**Analyzing Social Networks Of XML Plays:**

**Exploring Shakespeare’s Genres**

*Lawrence Evalyn, Susan Gauch, and Manisha Shukla*

# 1. Introduction

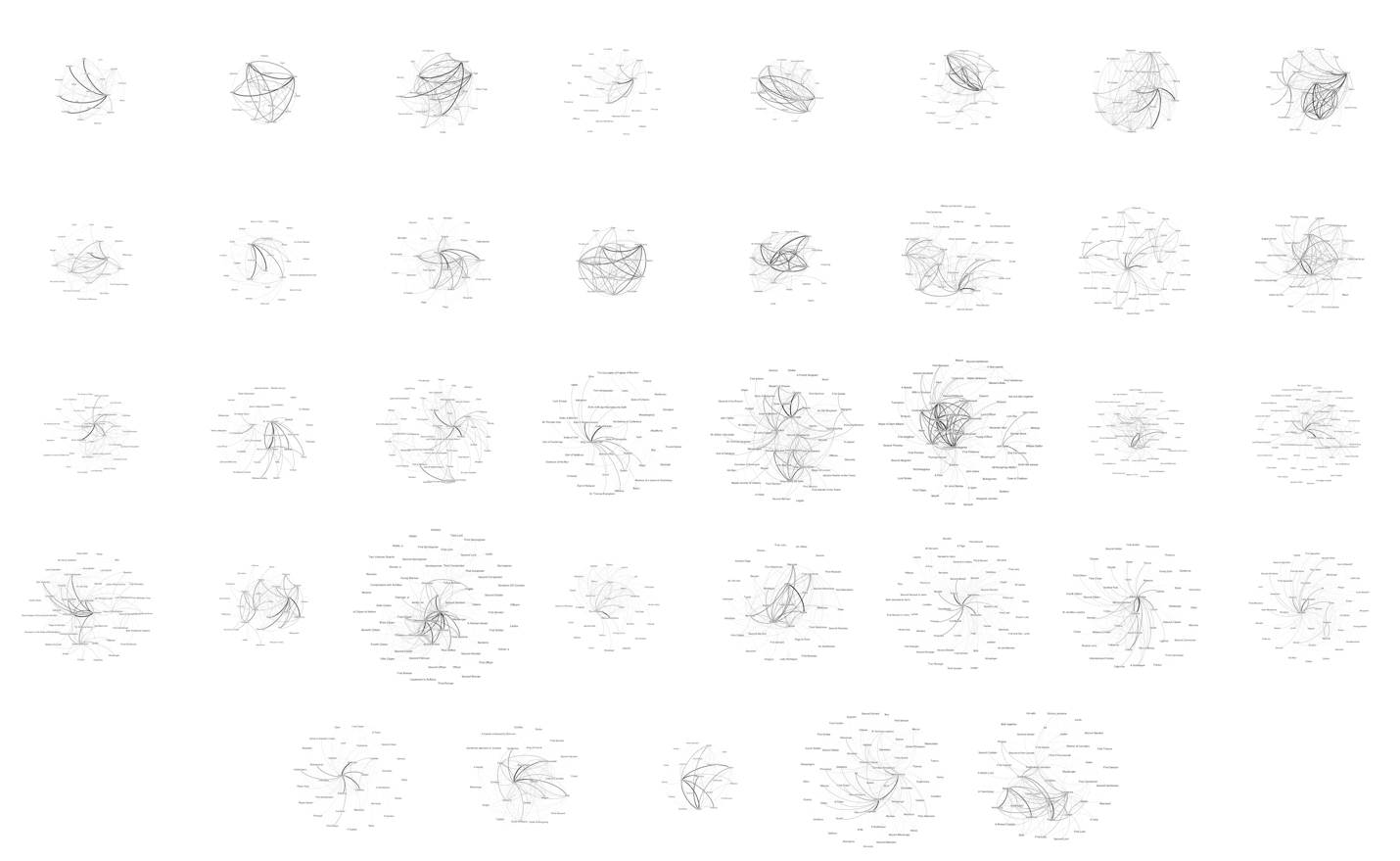
Digital scholarship in literature has a long history of distant reading “macroanalysis” (Jockers), which seeks to answer large questions of cultural history. These techniques are also often fruitful for something more akin to ‘close reading at a distance’ or what Hoover terms “microanalysis,” which uses “style variation within a single work or author as a way of better understanding authorship and style” (2016). At both scales, a great deal of work exists using stylometry to examine long prose works, with methods that are often thoughtfully adapted to address poetry as well. Drama, however, poses a unique challenge to distant reading: its meaning is only partly captured by the page. In the long arc of theatrical criticism, the process of reading drama, as opposed to watching it, has required justification: a play and its script are not identical, because a play always gestures toward an embodied, social experience.[[1]](#footnote-1)

Rather than adapting distant reading methodologies which were designed for prose and retrofitted for drama, we set out to define a method which was designed for drama. We have begun with social network graphs. Ours is not the first project to conceive of plays as social networks. 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). However, as fictions, these networks are shaped not only by sociological principles, but also by narrative structures. Moretti uses social networks to examine the plots of three Shakespearean tragedies, and to contrast the structure of chapters in English and Chinese novels (2011). Alberich et al. (2002) and Sparavigna (2013) also discuss the interplay between social and narrative constraints on networks. We emphasize this distinction to look for specifically literary features of our networks.

Modelling a play as a social network of character interaction renders the speech act as the fundamental “unit” of a play’s script. People talking to each other is what makes a play happen.[[2]](#footnote-2) Length, direction, number, and reciprocity of speech are shaped by underlying social relations. The action of the play, too — its narrative content — is a result of varying social relations changing over the course of the play and attempting to work themselves out. What these patterns of speech acts ultimately mean will vary based on theatrical context.[[3]](#footnote-3) Matching these speech-act patterns to specific narrative content enables theorization of literary groupings such as genre, period, and author. Our paper uses a case study of Shakespeare’s works and the Early Modern generic categories of “tragedy,” “comedy,” and “history” to demonstrate that the social networks of plays do encode narrative content. Moreover, we shows that this encoding of narrative content is meaningful for both microanalysis and macroanalysis.

# 2. Methods

Examining network graphs requires, first, a substantial body of network graphs. Many scholars have handmade theatrical network graphs, both as wholly artistic interpretations (Moretti) and as mathematically-expressed networks (Trilcke et al., Fischer et al.). Like the Digital Literary Network Analysis (dlina) working group, we acquire our networks by parsing plays which have been encoded according to the Text Encoding Initiative (TEI) standards. Our parser takes in a TEI-encoded play, extracts characters for the nodes of a network, and connects nodes via a directional edge each time a character speaks while another character is on stage as a potential auditor.[[4]](#footnote-4) By using directional edges, weighting character connections based on speech, and by sub-dividing the play into blocks smaller than the unit of the scene based on the configurations of characters present on stage, we create networks at a higher level of detail than prior work which creates edges simply for co-occurrence within a scene. We began by replicating Moretti’s results in order to verify our parser’s accuracy,[[5]](#footnote-5) then generated networks for a full corpus of Shakespeare’s works.[[6]](#footnote-6)



*Figure #: Network graphs of all 37 plays*

We then used 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 asked whether social networks can be used to distinguish between the dramatic genres of tragedy, comedy, and history.

We decided that the answer would be “yes” if we could train a classifier to replicate the First Folio’s classifications for Shakespeare’s 36 plays. We tested 21 different mathematical features of the networks, with three different classifiers. These features are a combination of extracted play features (such as the number of lines), node features averaged across all nodes (such as average eigenvector centrality), and graph features (such as graph density). The most accurate classifier was a support vector machine with one vs one classification (SVM with OVO); Naïve Bayes was typically a few percentage points below SVM with OVO, and k-nearest neighbours (KNN) performed so poorly in comparison that we eliminated it from our experimentation.[[7]](#footnote-7) We found that even one feature could achieve classification accuracy substantially above random chance. Since there are three possibilities, a random guess would be correct 33% of the time, but as Figure 1 shows, SVM achieved 66.43% accuracy examining only path length, and Naïve Bayes achieved 63.57% examining only the number of lines.

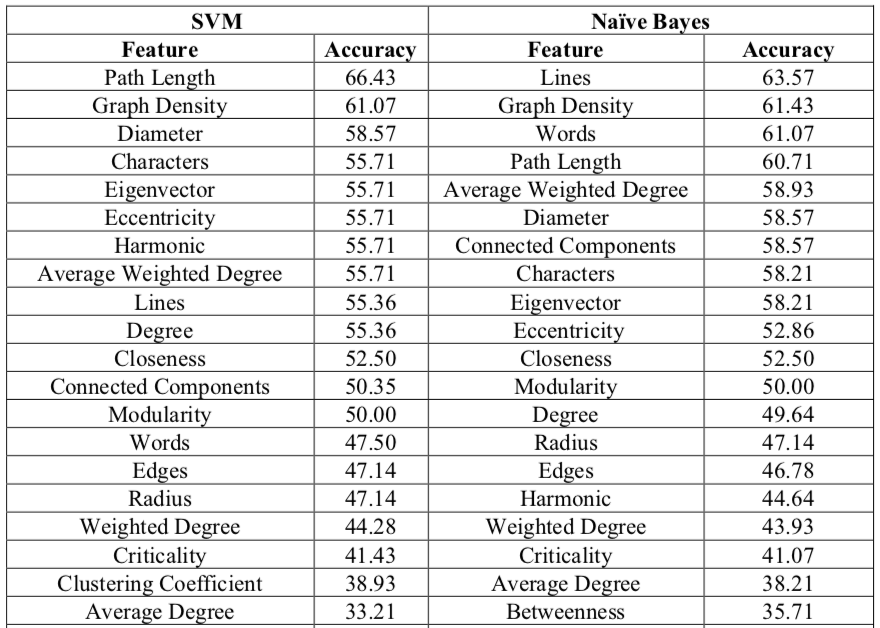


Figure 1: features which perform better than random in classification.

Accuracy increases as more features are considered in combination: with two features, SVM achieves 72.5% accuracy and Naïve Bayes achieves 77.86%.[[8]](#footnote-8) With three features, SVM achieves 83.57% accuracy and Naïve Bayes achieves 80.71%.[[9]](#footnote-9) We evaluated combinations of features up to 7 maximum. Our highest accuracy (88.93%) used SVM to classify the plays with five features, a combination of play features (words, lines), node features (closeness), and network features (density and degree).

# 3. Findings

Comedies are, it turns out, the most straightforward networks to identify. They have fewer nodes, with high graph density. Although comedies often have multiple subplots, we find that these separate stories do not result in highly-separated networks. We theorize that comedic networks are strongly shaped by the final “resolution” scenes, which bring the cast together for a restoration of social order, and therefore create edges knitting the network together. (See, for example, A Midsummer Night’s Dream.) The average eigenvector centrality of the characters in comedies is also much higher than in tragedies or histories; this suggests that many more of the characters in a comedy are “important,” reflecting a focus on ensemble stories.

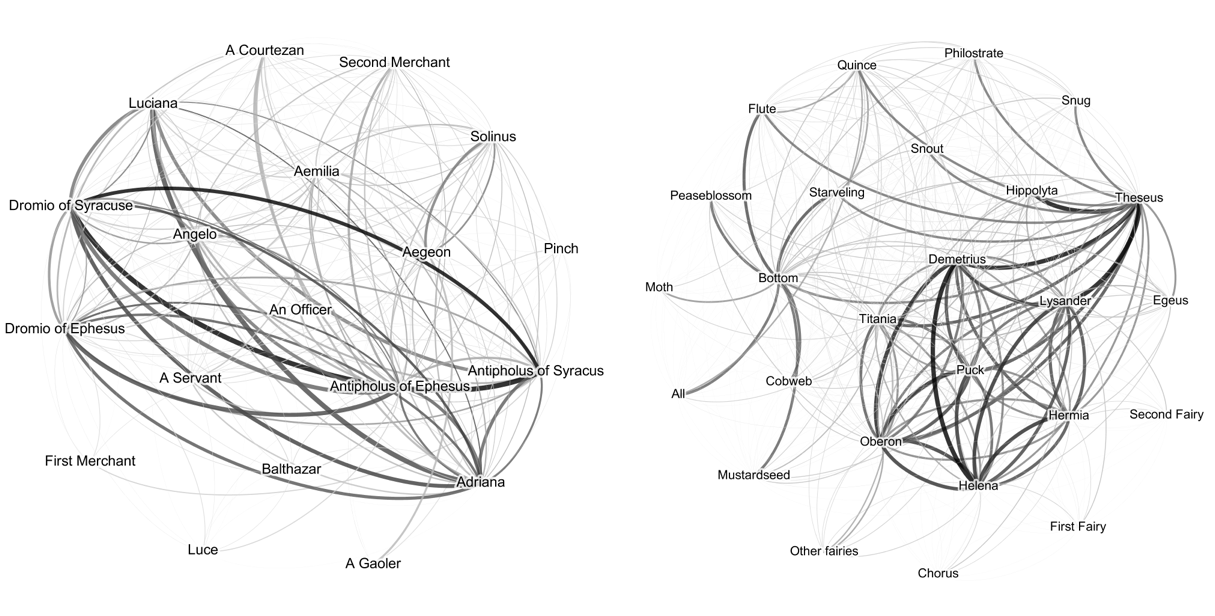


Figure #: A Comedy of Errors, A Midsummer Night’s Dream

In contrast, histories feature highly dispersed networks, with large numbers of very minor characters,[[10]](#footnote-10) such as “First,” “Second,” and “Third” members of groups like soldiers and ambassadors, who each interject briefly in a single scene. Connections form chains of acquaintance with little overlap, so even the monarchs have low eigenvector centrality. Whereas the directional edges in comedies tended to be reciprocal and balanced, in histories they are often unidirectional, as characters give orders or statements to each other.

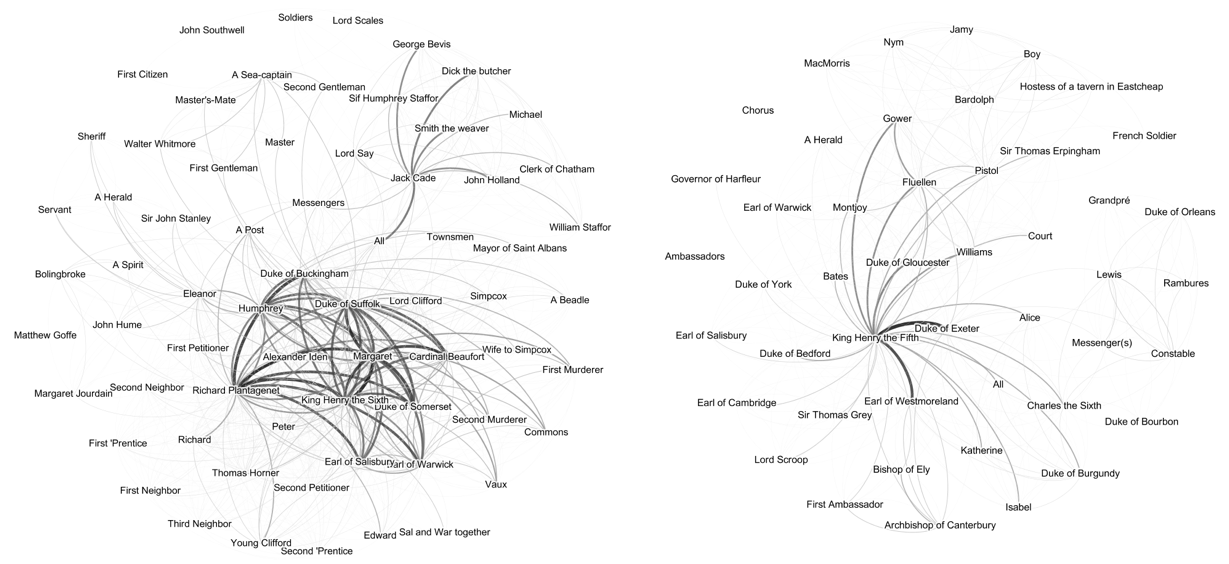


Figure #: Henry VI Part 2, Henry V

Tragedies are more formally unpredictable, and sit somewhere ambiguously between comedies and histories. Many have a dense core with a secondary ring of more peripheral characters. Many have a node who forms a distinct, singular centre of the network, connected to a large number of nodes that are only sparsely connected to each other. Directional edges are, as in histories, often unreciprocated, and the “tragic hero” nodes are particularly likely to have very heavily weighted edges connecting outward (indicating that they have given long speeches) with light to no reciprocation.

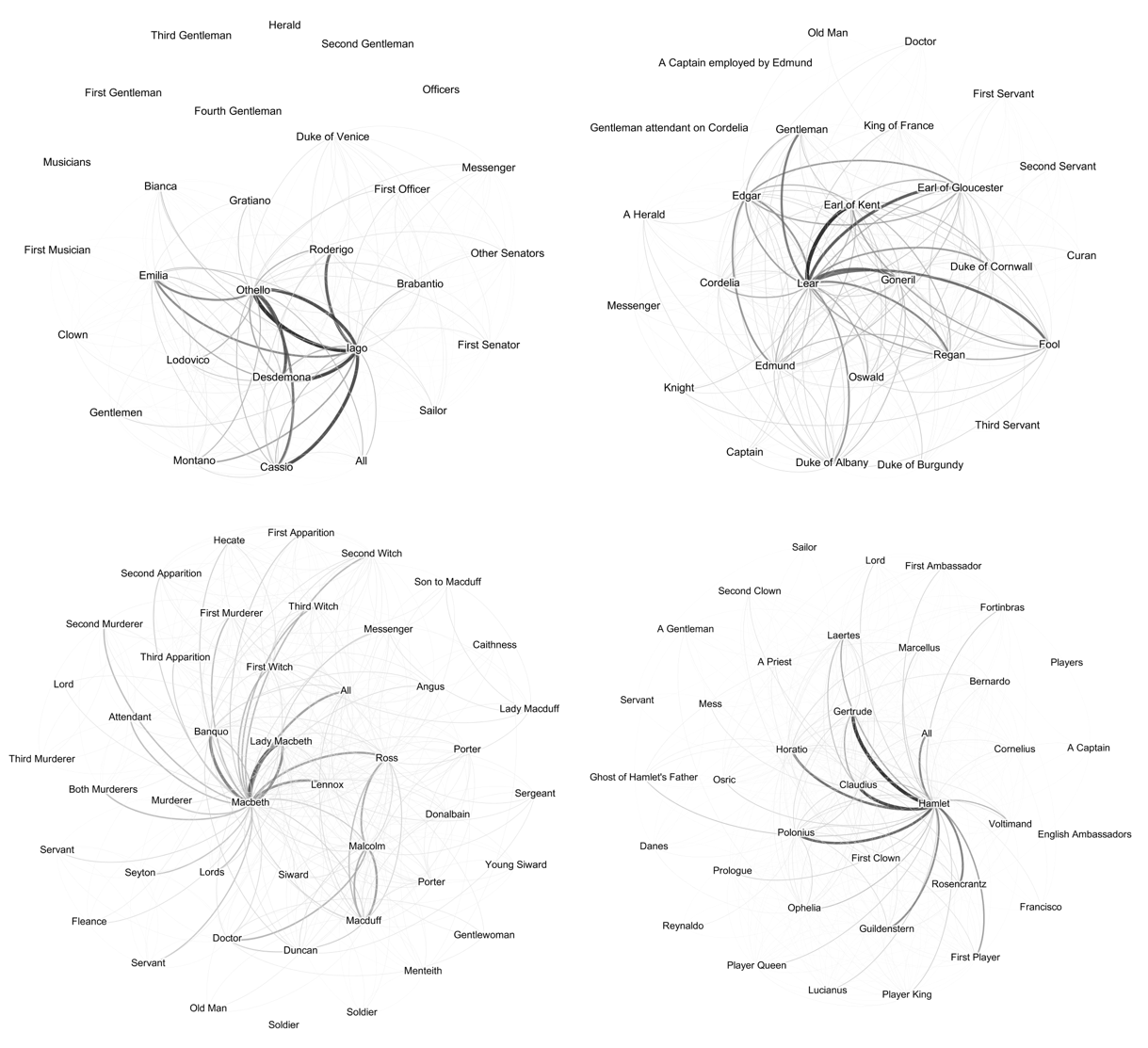


Figure #: Othello, King Lear, MacBeth, Hamlet

Having verified that the network graphs do encode literary information, future work can apply these methods for macroanalysis. Our SVM trained on Shakespeare will only give meaningful results in an Early Modern context, but could nonetheless be straightforwardly applied to a larger Early Modern corpus. Algee-Hewitt has found, for example, [a change in eigenvector over time]. Our work suggests that he may be seeing, in part, a change in the popular genres.

[The rest of this paper is going to be microanalysis.] 36 plays is a small enough corpus that they can be scrutinized individually, and their network graphs visually examined. Much of the difficulty in identifying genre in the first place stems from genre’s slippery status. Even in evaluating accuracy, we actually determine the classifier’s ability to predict how the First Folio would have labelled a work, not the work’s “true” and objective genre. Shakespearean scholars by no means universally accept the First Folio’s assignments. Accordingly, once we trained our SVM classifier, we queried two sets of ambiguous texts, the “problem plays” and the “romances,” to see what social network analysis might offer these debates.

As Figure 2 shows, both classifiers generally concur with the First Folio’s classification of the “problem plays,” with the exception that the SVM classifier identifies Troilus and Cressida as a comedy rather than a tragedy. The romances (Figure 3) are slightly more ambiguous.

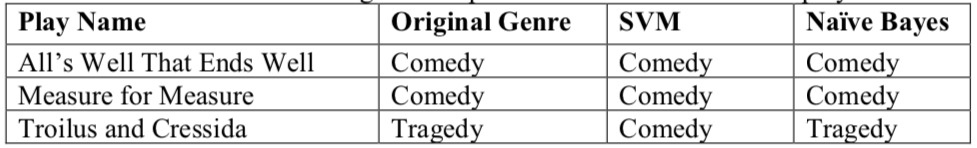


Figure 2: classifications for the “problem plays.”

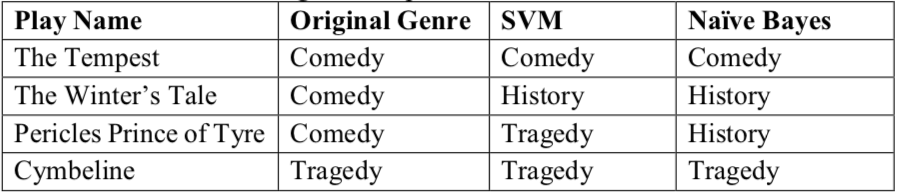


Figure 3: classifications for the romances.

The classifiers concur with the First Folio regarding The Tempest, but neither classifies The Winter’s Tale or Pericles[[11]](#footnote-11) as a comedy. Moreover, the classification of Cymbeline (Figure 4) as a tragedy, even though this classification concurs with the First Folio, constitutes a surprise. Audiences since Samuel Johnson in 1765[[12]](#footnote-12)  have complained of Cymbeline’s failures as a tragedy (among other things, it has a happy ending), and Barbara Mowat refers to the Folio’s identification of the play as “The Tragedie of Cymbeline” in the index, title page, and running heads, as “misleading paratextual matter” (130). Our classification, in contrast, supports Valerie Wayne and Richard Proudfoot’s assessment that although “[i]t is not a tragedy as that form is most widely understood today or as Johnson understood it in the eighteenth century, […] it did sufficiently meet Renaissance definitions” of that genre (Wayne 403).

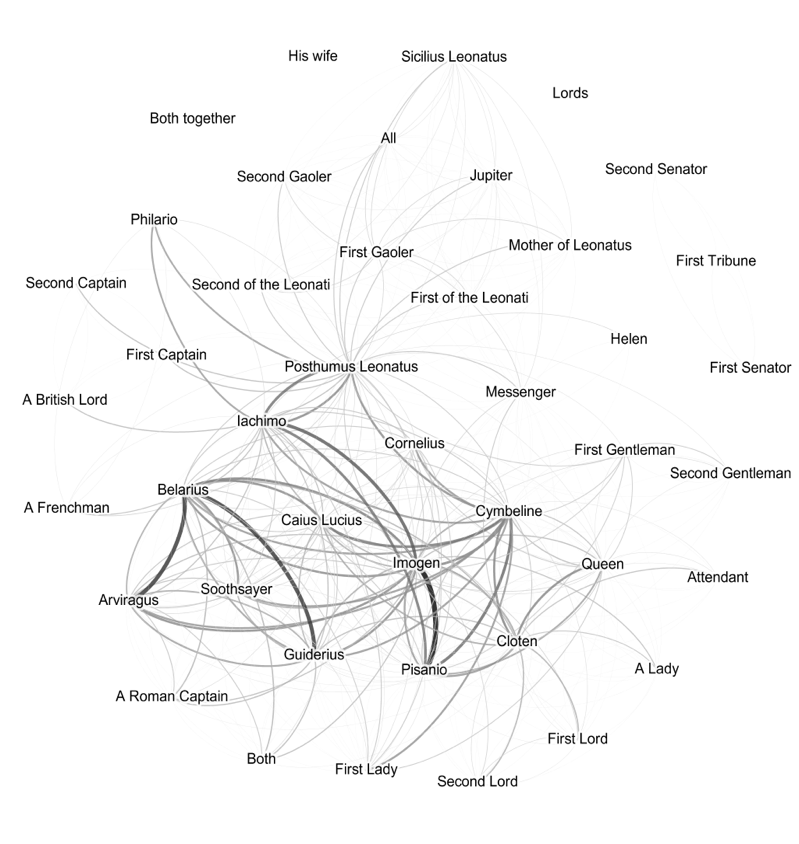


Figure 4: network graph of character interaction in Cymbeline.

The Winter’s Tale provides a particularly useful case study for close-reading a play through the lens of its social network. The play earns its designation as a “romance” due to its two-part structure:[[13]](#footnote-13) the first X acts play like a tragedy, in which Leontes gradually ruins his own life. At the beginning of Act X, the character of “Time” comes on stage to declare that XX years have passed, and the play begins an extended pastoral comedy which ultimately resolves all the looming tragedies of the beginning of the play. These two theatrical dynamics are visible in the social network graph, as seen in Figure 5: Leontes radiates unreciprocated and unbalanced edges, looking like a Hamlet of his world. But the characters who enter in the latter half of the play — NAMES HERE — connect to each other in a dense, comedic structure. As a result, the play’s network as a whole is neither a comedy nor a tragedy, which the classifiers evidently deem a history.

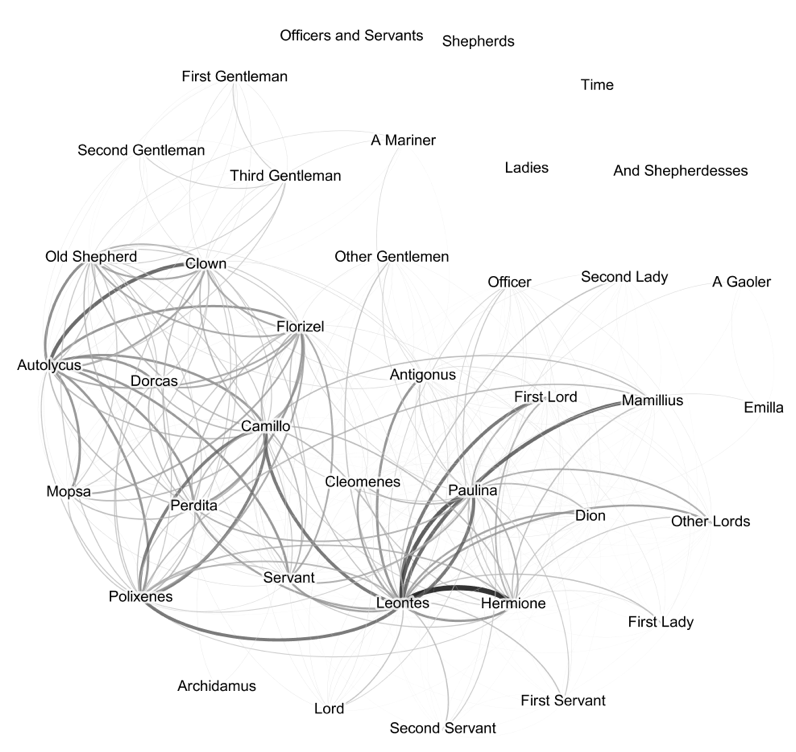


Figure 5: network graph of character interaction in The Winter’s Tale.

# 4. Conclusions

We treat none of these identifications as definitive declaration of the plays’ “real” genres, but use them to distinguish between plays whose generic ambiguity lies in their subject matter, and plays whose ambiguity lies in their structure.

Our parser successfully and rapidly produces sophisticated social network graphs of TEI plays, which can be used to computationally identify theatrical genre in Shakespeare’s plays. Thirty-seven plays is a small scale for this approach: since the parser is highly extensible and can be used with any plays encoded in TEI, future work need not be restricted to the Early Modern period. It need not even be restricted to works written in English. These future projects [will need to take their local content into account], but [thanks to us they now know to look at density, eigenvector centrality, path length, and reciprocation]. Our networks of the well-studied works of Shakespeare can provide a baseline against which to contextualize analysis of these elements in works for which there is far less critical consensus.

# References

1. Alberich, R.,Miro-Julia, J., and Rosselló., F. (2002). Marvel Universe Looks Almost Like a Real Social Network. arXiv:cond-mat/0202174v1
2. Algee-Hewitt, M. (2017). Distributed Character: Quantitative Models of the English Stage, 1500-1920. *Digital Humanities 2017: Book of Abstracts*. Montreal: McGill University and Université de Montréal, pp. 119–21.
3. Hoover DSH 2016
4. Bastian M., Heymann S., Jacomy M. (2009). *Gephi: an open source software for exploring and manipulating networks.* International AAAI Conference on Weblogs and Social Media.
5. Frank Fischer, Mathias Göbel, Dario Kampkaspar, Christopher Kittel, Hanna-Lena Meiners, Danil Skorinkin, and Peer Trilcke. Digital Literary Network Analysis (dlina). https://dlina.github.io/
6. Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., and Trilcke, P. (2017). Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts. Digital Humanities 2017: Book of Abstracts. Montreal: McGill University and Université de Montréal, pp. 437–41.
7. Jockers - macroanalysis
8. Johnson, S. *Selections from Johnson on Shakespeare*, ed. Bertrand H. Bronson with Jean M. O’Meara (New Haven, CT: Yale UP, 1986), 307.
9. Moretti, F. (2011). Network Theory, Plot Analysis. New Left Review, 68: 80–102.
10. Mowat, B. “‘What’s in a Name?’: Tragicomedy, Romance, or Late Comedy,” in *A Companion to Shakespeare’s Works*, ed. Richard Dutton and Jean E. Howard, vol. 4, *The Poems, Problem Comedies, Late Plays* (Oxford: Blackwell, 2003)
11. Shakespeare, William. A Midsummer Night’s Dream.
12. Shakespeare, William. Cymbeline.
13. Shakespeare, William. Hamlet.
14. Shakespeare, William. The Winter’s Tale.
15. Stiller, J., Nettle, D., and Dunbar, R. I. M. (2003). The Small World of Shakespeare's Plays. Human Nature, 14(4): 397–408.
16. Shukla ?
17. Sparavigna, A. C. (2013). On Social Networks in Plays and Novels. International Journal of Sciences, 2: 20–25.
18. Trilcke, P., Fischer, F., Göbel, M., Kampkaspar, D. (2016). Theatre Plays as ‘Small Worlds’? Network Data on the History and Typology of German Drama, 1730–1930. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 385-387.
19. Wayne, V. "The First Folio’s Arrangement and Its Finale." Shakespeare Quarterly, vol. 66 no. 4, 2015, pp. 389-408. Project MUSE, doi:10.1353/shq.2015.0064
20. WordHoard Shakespeare

1. Even in periods when the reading of drama was valorized, as in the eighteenth century, this was due to the perceived perfections of the “theatre of the mind” as superior to the physical theatre in giving free play for that which constitutes the play but is not present on the page. [↑](#footnote-ref-1)
2. Or at the very least, the words that people say to each other are the subset of what happens in the play which the writer of that play is able to direct; the meaning created by gesture, intonation, and action is the purview of the director and actor. [↑](#footnote-ref-2)
3. For example, in John Dryden’s heroic drama *The Conquest of Granada*, it is a mark of Boabdelin’s social importance in the first act that he can demand and recieve long reports of others’ deeds, whereas in William Shakespeare’s *Hamlet*, it is a mark of Hamlet’s importance in the play that he *gives* long speeches. [↑](#footnote-ref-3)
4. Although we have not yet solved the problem of guessing the specific addressees of a character’s speech, we do use stage directions to track character entrances and exits, thereby reducing some of the false positives created when edges are created at the scale of shared scenes. [↑](#footnote-ref-4)
5. The existence of well-crafted graphs for comparison partly motivated our selection of Shakespeare as our first subject. In general, our graphs differ from Moretti’s primarily by including characters whom he excludes: we have nodes for minor characters like “Servant,” for example, and we include all the players of the play-within-the-play in Hamlet. Because our edges are directional, we are also able to represent details of character interaction which Moretti regretted leaving out, such as the fact that Horatio addresses the Ghost, but the Ghost speaks only to Hamlet. Despite these differences, our network graphs generally support Moretti’s literary readings. [↑](#footnote-ref-5)
6. Specifically, we generated networks for the 36 plays included in the First Folio, and (to contribute to scholarly conversations about Shakespeare’s “romances”) *Pericles, Prince of Tyre*, for 37 total plays. All play texts came from Northwestern University’s “WordHoard Shakespeare.” [↑](#footnote-ref-6)
7. KNN is particularly vulnerable to the over-representation of comedies among Shakespeare’s works. Naïve Bayes treats each feature as independent, an assumption unlikely to be correct in our case, since, e.g., the number of words in a play are likely to be closely related to the number of lines in the play. [↑](#footnote-ref-7)
8. SVM achieves 72.5% with four different pairs: harmonic & diameter; harmonic & path length; graph density & diameter; and graph density & path length. Naïve Bayes achieves 77.86% with lines & eigenvector. [↑](#footnote-ref-8)
9. SVM achieves 83.57% accuracy with words, characters, & lines. Naïve Bayes achieves 80.71% with lines, graph density, & degree. [↑](#footnote-ref-9)
10. Early Modern theatre companies remained generally constant in size, so these additional characters likely indicate increased doubling (or tripling, etc) of roles [↑](#footnote-ref-10)
11. *Pericles* is not, of course, a First Folio play, and has not formed part of our corpus for the other stages of this project, but was graphed and classified in service of this question. Its “original” designation as a comedy comes form general critical consensus rather than the First Folio. [↑](#footnote-ref-11)
12. Every discussion of *Cymbeline*, and particularly of its genre, must duly quote Johnson’s judgment that “This play has many just sentiments, some natural dialogues, and some pleasing scenes, but they are obtained at the expense of much incongruity. To remark the folly of the fiction, the absurdity of the conduct, the confusion of the names and manners of different times, and the impossibility of the events in any system of life, were to waste criticism upon unresisting imbecility, upon faults too evident for detection, and too gross for aggravation” (307). I hereby discharge this solemn duty. [↑](#footnote-ref-12)
13. probably cite somebody about this [↑](#footnote-ref-13)