Buenos dias!

Thank you for waking up early to be here.

My only spoken spanish is "lo siento" but I think I have gotten my slides into readable condition.

I'm from U of T, presenting on some work of the Computational Literature Project at the University of Arkansas.

PI Susan Gauch, Manisha Shukla, computer scientists

who have essentially humoured my intense curiosity about the social nature of plays, and thus the social nature of theatrical plots.

I'm going to tell you today about the first few steps we've taken to use social networks as proxies for plot at a large scale.

The first step was developing a way to turn a play into a network.

Our parser takes TEI encoded XML files and makes a number of assumptions to generate a social network from it.

- Tracks characters present on stage at the same time

- Uses stage directions to subdivide scenes into “blocks” based on character groupings

- A directional edge is created when a character is on stage to (probably) hear another character’s speech

- Edges are weighted by the total words spoken by Character A while Character B can hear

One of the reasons we began with shakespeare is that there already exist a number of handmade social networks of his plays which let us check our work, most famously Franco Moretti's.

These networks look a little different here, but represent the same basic structure, but with a higher level of detail in our network, which features directional and weighted edges.

The other thing we've done to expand from Moretti's work is that we have networks of all 36 First Folio plays, plus Pericles (for reasons I will mention later).

Beautiful, eh? Already you can see that there is a lot of variation, even within this small single-author corpus.

So the question is, does this variation tell us anything about genre?

Here is where the computer scientists stepped in again. We used a support vector machine to test out the 17 mathematical network features that Gephi provides, cross-validating with training sets of 31 plays and testing sets of 6, to see if any of them could be used to identify genre.

The most useful single feature was graph density (83% accuracy)

If multiple features are used in combination, the classification can achieve 100% accuracy

Combinations seem to need a metric of the size of the play (lines, edges, words, nodes), and another metric of its interconnectedness (eccentricity, degree, eigenvector, path length, density)

So, what does that mean in terms of the plays themselves?

The relevance of these features is most apparent when we look at the constrast between comedy and history

Histories and comedies have very different graph densities

Histories have lots of peripheral minor characters, and each character only knows a few others

In comedies, even with different subplots, everybody eventually knows everybody

Just as a few more examples, we can see two more histories

-- and two more comedies.

Although the classifier does identify tragedies successfully, it's less obvious to me as the human what sets them apart. They're usually in between comedies and histories in size and interconnectedness, but they vary.

My hypothesis right now is that they often feature a radiating structure with a "tragic hero" who monologues a lot.

It you look at Macbeth and Iago and Othello, althoughthe networks are otherwise fairly different, they have this feature roguhly in common.

We can also see it of course with Hamlet, the supreme monologuer, and King Lear.

This may be why our classifier was so certain that the Roman plays are all tragedies, not histories -- Timon of Athens and Julius Caesar have similar "big talkers" too.

The Roman plays are good reminder that these genre classifications are by no mean as settled and decided as our SVM would like them to be.

There has been scholarly debate for hundreds of years about Shakespeare's use of genre, particularly when it comes to the sets of plays referred to as the "problem plays" and the "romances". We were curious what claim our classifier would try to stake in this debate, if we trained on all of the other plays and then asked it to classify just the problem plays or just the romances.

At some points it does uphold the First Folio, and in others it claims that tragic plays are in the end comedies.

This is why Pericles is included -- as one of the Romances, we wanted to see how it would be classified.

To look at just some of these results in more detail,

This obviously doesn't solve the "real" genre of these plays, but it gives us another way to look at what's complex about them.

The Winter's Tale has a dense "comedic" sub-network introduced in the second half of the play.

I'm not a Shakespearean, so I don't have the expertise to take these findings and attempt to change the debate about Early Modern genre. But I do think this work can show the relevance of social network graphs to plot analysis, including with supervised machine learning, and give a taste of something that is at a fairly "boutique" scale now but which can easily scale up.

The parser can be used on any TEI plays (including other languages), and we're eager to hear from scholars who have their own plays they'd like to explore with these methods.

We're developing a website to allow scholars to upload their own TEI files and download the resulting networks

In my own further work, now that I've confirmed with some better-understood plays that the underlying principle is sound, I'm excited to look at the social networks of eighteenth century tragedies, to answer some questions I have about the role of women on stage.

As part of that, the next feature we intend to add is to introduce a node for the audience, who hears all soliloquies and asides.

We have the rudimentary Shakespeare graphs available online in case you want to see them, though they're not as pretty as the ones I made for this presentation. If you do have plays you want to explore with us, or just want to hear more, I hope you will get in touch with me.

I hope this taste of our work has persuaded you that social network graphs can be used to distant-read theatrical plots, and I thank you again for your time. Gracias!