Shakespearean Social Network Analysis using Topological Methods

Bastian Rieck

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

HGS MathComp

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Who was Shakespeare?

- Baptized on April 26th 1564 in Stratford-upon-Avon
- Died on April 23rd 1616 in Stratford-upon-Avon
- 38 plays
- 154 Sonnets
- Broad classification into *tragedies*, *comedies*, and *histories*.

Shakespeare's plays

Comedies	Tragedies	Histories
A Midsummer Night's Dream	Antony and Cleopatra	The Life and Death of King John
All's Well That Ends Well	Coriolanus	Henry IV, Part 1
As You Like It	Cymbeline	Henry IV, Part 2
Cymbeline	Hamlet	Henry V
The Comedy of Errors	Julius Caesar	Henry VI, Part 1
Love's Labour's Lost	King Lear	Henry VI, Part 2
Measure for Measure	Macbeth	Henry VI, Part 3
The Merchant of Venice	Othello	Henry VIII
The Merry Wives of Windsor	Romeo and Juliet	Richard II
Much Ado About Nothing	Timon of Athens	Richard III
Pericles, Prince of Tyre	Titus Andronicus	
The Taming of the Shrew		
The Tempest		
Twelfth Night		
The Two Gentlemen of Verona		
The Winter's Tale		

Why Shakespeare?

Idioms

2016 marks the 400th anniversary of Shakespeare's death. He continues to have a lasting influence on the English language:

- 'A dish fit for the gods' (Julius Caesar)
- 'A foregone conclusion' (Othello)
- 'A horse, a horse, my kingdom for a horse' (Richard III)
- 'Brevity is the soul of wit' (Hamlet)
- 'Give the Devil his due' (Henry IV)
- 'Heart of gold' (Henry V)
- 'Star-crossed lovers' (Romeo & Juliet)

Why Shakespeare?

Humour

SECOND APPARITION: Macbeth! Macbeth! Macbeth!
MACBETH: Had I three ears, I'd hear thee.
SECOND APPARITION: Be bloody, bold, and resolute.
Laugh to scorn the power of Man, for none of woman born
shall harm Macbeth.

— Macbeth, Act IV, Scene I

Disclaimer

I'm a linguistic barbarian. Please correct me if what I am telling you makes absolutely no sense or is in direct opposition to linguistic research.

Motivation

Stories appear to follow some basic patterns. We all know certain tropes that appear and re-appear.

- J. Campbell, *The Hero with a Thousand Faces*: Myths from around the world share the same narrative structure.
- K. Vonnegut, *The Shapes of Stories*: By graphing the 'ups' and 'downs' of a character, the story reveals its shape.
- A. J. Reagan, *The emotional arcs of stories are dominated by six basic shapes*: Many stories share the same 'emotional arcs'.

Motivation

How can we measure *structural similarities* between Shakespeare's plays in a mathematically sound way?

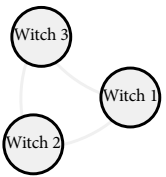
How to represent a story?

- Create a *social network*—a graph—from the play
- Every character becomes a *vertex* in the graph
- If two characters talk in the same scene, connect them by an *edge*
- Data source: Tagged corpus[1]
- Using conversion scripts by Ingo Kleiber[2,3]

[1] http://lexically.net/wordsmith/support/shakespeare.html
[2] https://kleiber.me
[3] https://github.com/IngoKl/shakespearesna1406

Example

```
<SCENE 1>
<A desert Heath.>
<STAGE DIR>
<Thunder and lightning. Enter three Witches.>
</STAGE DIR>
<WITCH 1><0%>
        When shall we three meet again
        In thunder, lightning, or in rain?
</WITCH 1>

<WITCH 2><1%>
        When the hurlyburly's done,
        When the battle's lost and won.
</WITCH 2>

<WITCH 3><1%>
        That will be ere the set of sun.
</WITCH 3>

<WITCH 1><1%>
        Where the place?
</WITCH 1>

<WITCH 2><1%>
        Upon the heath.
</WITCH 2>

<WITCH 3><1%>
        There to meet with Macbeth.
</WITCH 3>

<WITCH 1><1%>
        I come, Graymalkin!
</WITCH 1>

<WITCH 2><1%>
        Paddock calls.
</WITCH 2>

<WITCH 3><1%>
        Anon.
</WITCH 3>

<ALL><1%>
        Fair is foul, and foul is fair:
        Hover through the fog and filthy air.
</ALL>
<STAGE DIR>
<Exeunt.>
</STAGE DIR>

</SCENE 1>
```

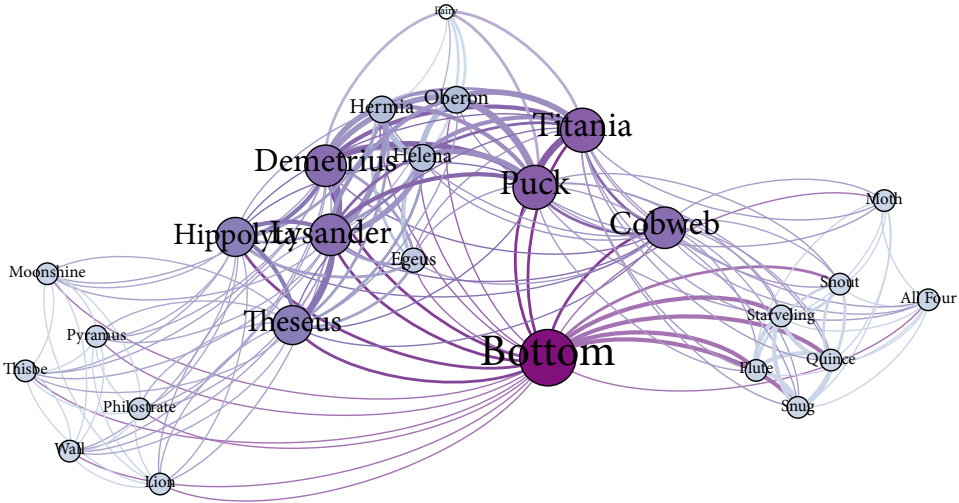


For larger graphs

- Use force-directed graph layout algorithms
- Scale node by its *degree*
- Assign edge weights based on the number of common scenes
- Colour & scale edges by their *weight*

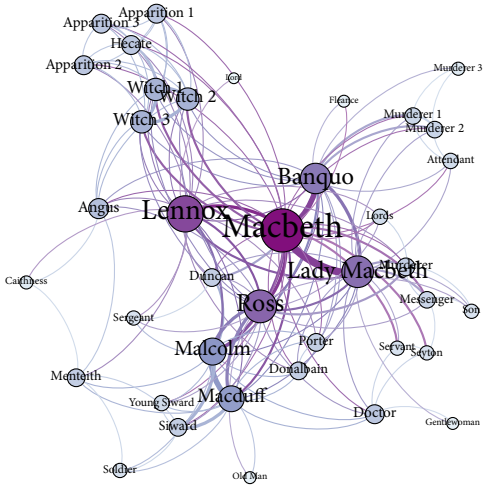
A Midsummer Night's Dream

Comedy



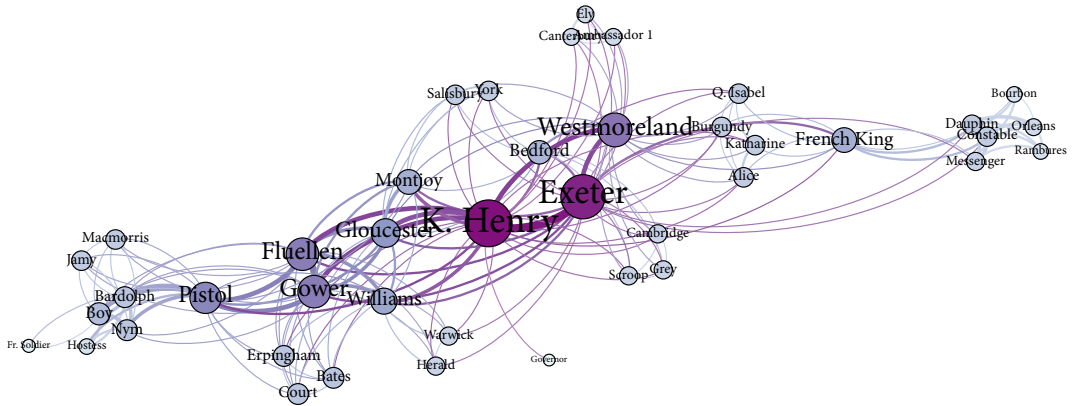
Macbeth

Tragedy

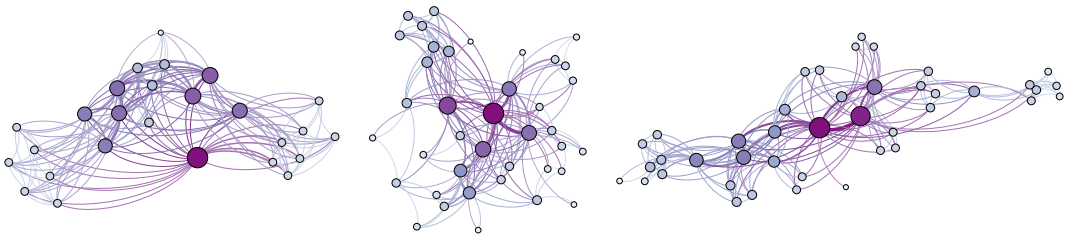


Henry V

History



So far

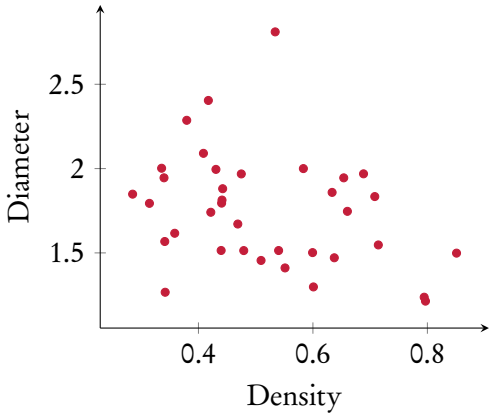


Some *apparent* structural differences—but how to quantify them correctly? Are they an artefact of the layout algorithm?

Measuring structural properties, I

Simple properties

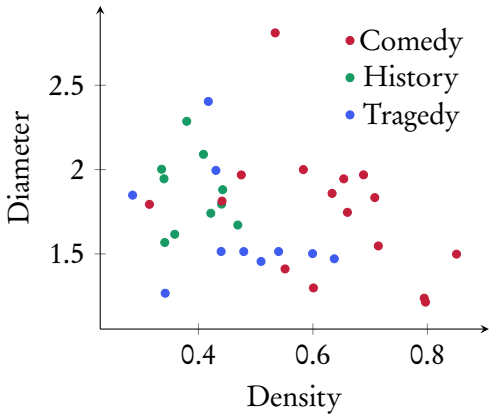
- *Density* : How different is the graph from a complete graph on n vertices?
- *Diameter*: How long is the longest shortest path?



Measuring structural properties, I

Simple properties

- *Density* : How different is the graph from a complete graph on n vertices?
- *Diameter*: How long is the longest shortest path?



What can we make of this?

- Histories have a *low density* and a *medium diameter*: Characters talk in smaller groups; groups are somewhat removed from each other
- Comedies have a *high density* and a *low–medium diameter*: Characters talk in larger groups; 'Much Ado about Nothing' and 'The Merry Wives of Windsor' have a very high cohesion, i.e. a small diameter, while 'Measure for Measure' has a very loose cohesion
- 'Measure for Measure' is one of Shakespeare's *problem plays* because it has a rather complex and ambiguous tone

Very 'coarse' measures, but still interesting.

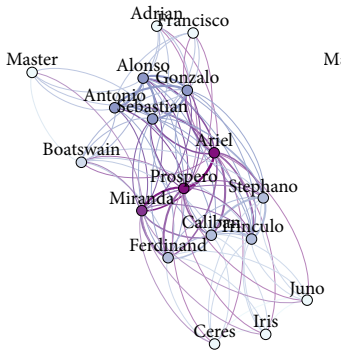
Measuring structural properties, II

Centrality measures

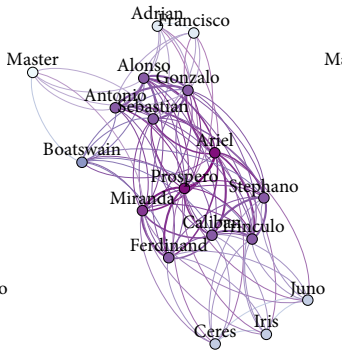
- *Betweenness centrality*: What fraction of all shortest paths within the graph use the current vertex?
- *Closeness centrality*: How far removed is the current vertex from the remaining vertices?
- *Eigenvector centrality*: Perform an eigenanalysis of the weighted adjacency matrix and use the components of the eigenvector corresponding to the largest eigenvector
- *Weighted degree centrality*: Use the sum of all weights of incident edges

Slightly different notions of ‘centrality’

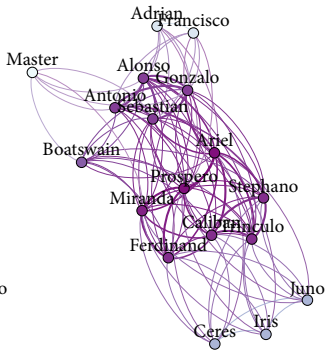
The Tempest



Betweenness centrality



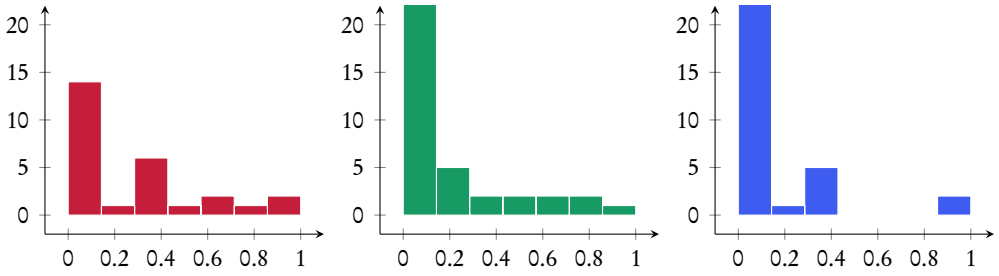
Closeness centrality



Eigenvector centrality

How to compare centrality measures mathematically?

Betweenness centrality distribution



Histogram distance measures (χ^2, Kullback–Leibler, …), Euclidean distance, Earth Mover's distance, …

However: Low discriminative power in the context of networks!

How to increase the discriminative power?

Persistent homology

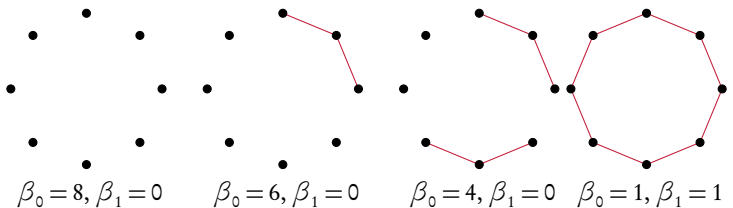
- Take *structural features* of the graph into account
- In particular, let's focus on the *connectivity* of the graph
- Natural problem for topological data analysis

How does this work?

Given a graph G, decompose it via a *graph filtration*,

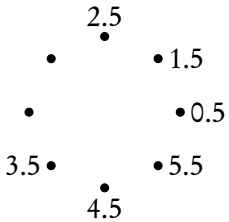
$$\emptyset = G_0 \subseteq G_1 \subseteq \cdots \subseteq G_n = G, \tag{1}$$

and study how the connectivity of the graph changes. In particular, we are interested in *connected components* and *loops*.

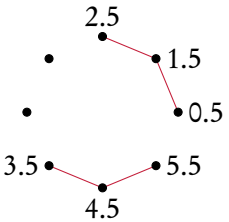


Key idea: If every graph in the filtration has a *weight function* assigned, we may measure how long structural features *persist* over the range of the function!

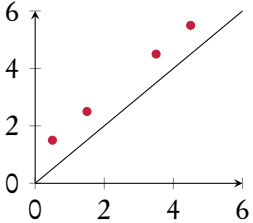
Collecting scale information



$\beta_0 = 8, \beta_1 = 0$



$\beta_0 = 6, \beta_1 = 0$



Persistence diagram

Why persistence diagrams?

- *Salient* shape descriptor for high-dimensional data sets
- Known stability & robustness results[4]
- Well-defined distance measures: *Bottleneck distance*, *Wasserstein distance*[5]
- Vector space formulation is possible—*averages* can be calculated!

p-norm summary statistic:

$$||\mathscr{D}||_2 := \left(\sum_{(c,d)\in\mathscr{D}} \text{pers}(c,d)^p \right)^{\frac{1}{p}} \tag{2}$$

[4] Cohen-Steiner et al.: *Stability of Persistence Diagrams*, Discrete & Computational Geometry 37:1, 2007

[5] Essentially, an Earth Mover's Distances between diagrams

Filtrations

Graph distances

$$f(v) := 0 \tag{3}$$

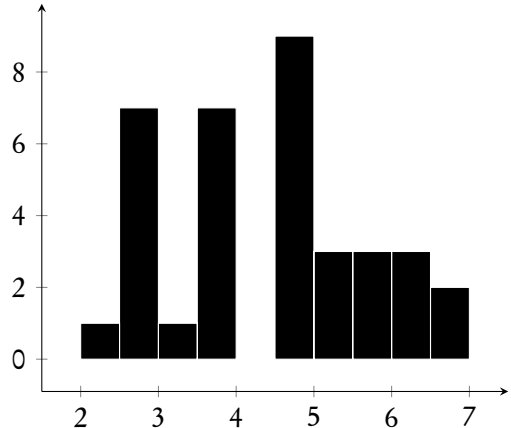
$$f(u,v) := \frac{1}{w(u,v)} \tag{4}$$

Properties:

- Naturally models a *metric* on the graph
- The distance is inversely proportional to the edge weight—characters that appear together in many scenes are considered to be close

Results

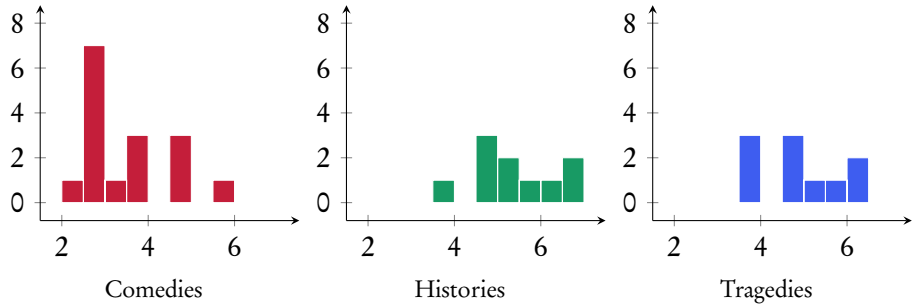
Graph distances



2-norm distribution

Results

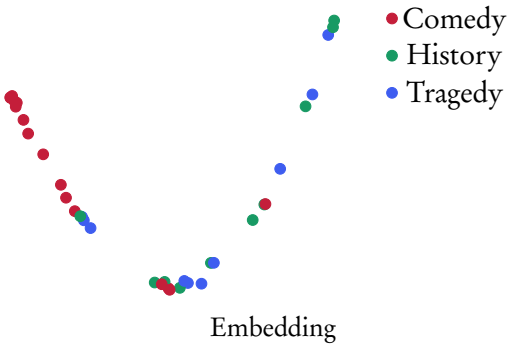
Graph distances



2-norm distributions, split by category

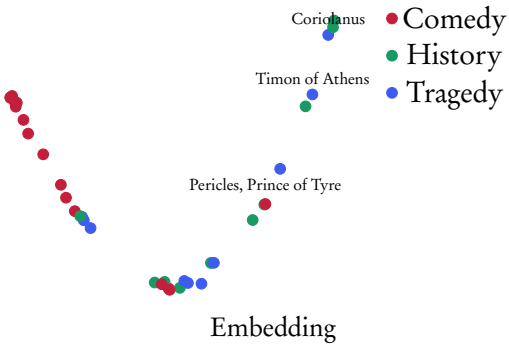
Results

Graph distances



Results

Graph distances



Filtrations

Centrality measures, merged

Let $c(v)$ denote a vertex-based centrality measure. Set weights to:

$$f(v) := c(v) \tag{5}$$

$$f(u,v) := \max(c(u), c(v)) \tag{6}$$

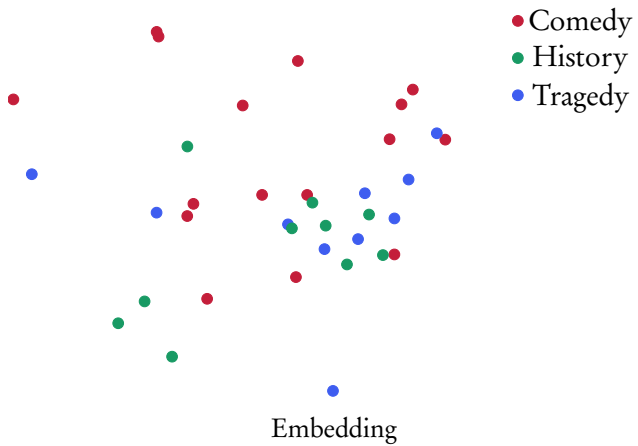
Properties:

- *Function-based* filtration; is able to capture the shape of networks slightly better[6]
- Affords calculation of *extended persistence*
- However, this does not model a metric!
- Merge corresponding persistence diagrams; simple *bag-of-features* approach

[6] Carlsson: *Topological Pattern Recognition for Point Cloud Data*, Acta Numerica 23, 2014

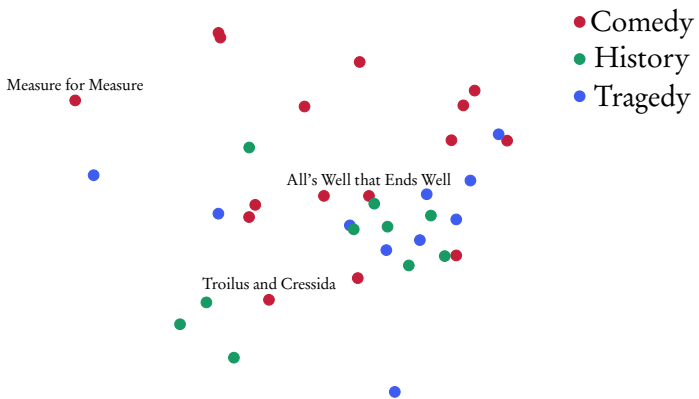
Results

Centrality measures, merged



Results

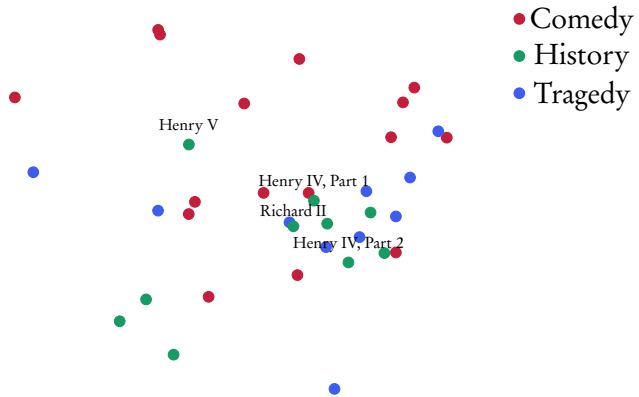
Centrality measures, merged



The 'problem plays'

Results

Centrality measures, merged



Structural changes in the *Henriad*

Filtrations

Centrality measures, mixed

Let $c(v)$ denote a vertex-based centrality measure. Set weights to:

$$f(v) := 0 \tag{7}$$

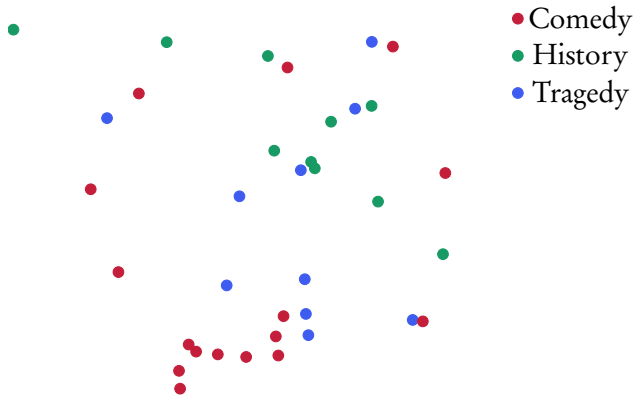
$$f(u, v) := \max(c(u), c(v)) \tag{8}$$

Properties:

- Pretend that $c(v)$ describes a metric
- By setting vertex weights to 0, more information about merges is retained
- Somewhat unjustified…

Results

Centrality measures, mixed (eigenvector centrality)



Conclusion & outlook

- Lots of structural, discriminative information available
- Robust topological analysis yields some (simple) insights
- Does a similar topology imply a similar story?
- Everything hinges on the definition of the graph…
  - Inclusion of *sentiment analysis*
  - Graph filtration based on temporal evolution of the play

Applications:

- *Recommending* a play; comparing plays of different authors to Shakespeare's plays
- Quantifying dissimilarity between different editions

'Bard Data' instead of 'Big Data'?