NETWORK ANALYSIS OF LITERARY TEXTS

Frank Fischer · Daniil Skorinkin

(Higher School of Economics, Moscow)

ffischer@hse.ru · dskorinkin@hse.ru

7 February 2018

TOC

1. INTRODUCTION

WHO WE ARE

- Centre for Digital Humanities at HSE, Moscow (hum.hse.ru/digital/)
- NUG on Digital Literary Studies (hum.hse.ru/digital/rusdracor/)
- Digital Humanities minor ("Современные методы в гуманитарных науках")

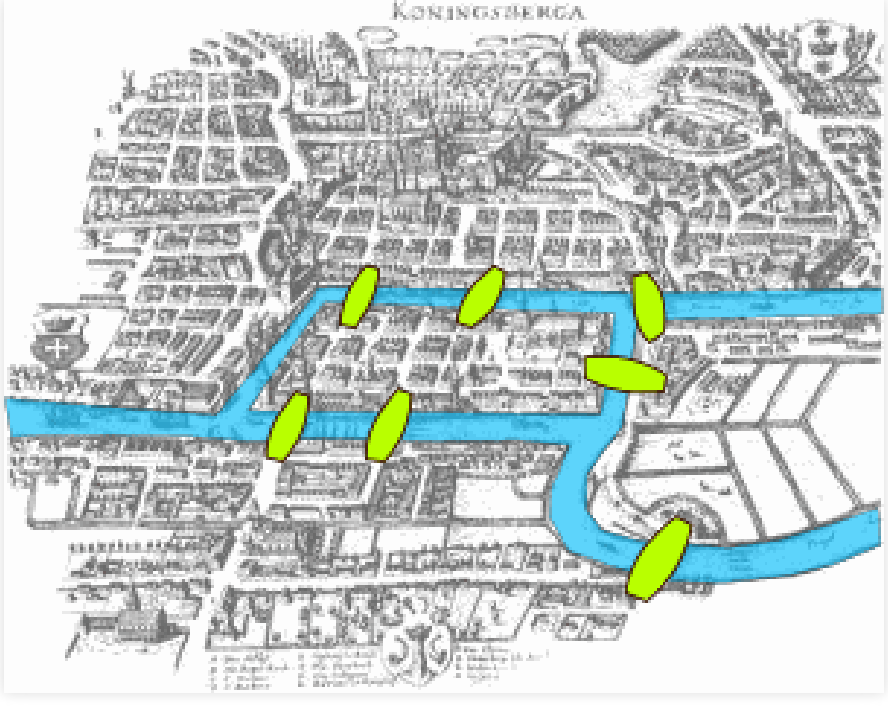
OBJECTIVE

- grow and maintain a corpus of Russian-language drama from around 1740 to around 1940
- XML-based markup standard TEI (Text Encoding Initiative)
- focus on extractable structures, especially social relations, but also linguistic properties
- main goal: a large-scale social network analysis of literary (dramatic) texts ("Distant Reading")

2. SOCIAL NETWORK ANALYSIS

THE EMERGENCE OF GRAPH THEORY

Euler's solution of the Königsberg bridge problem:

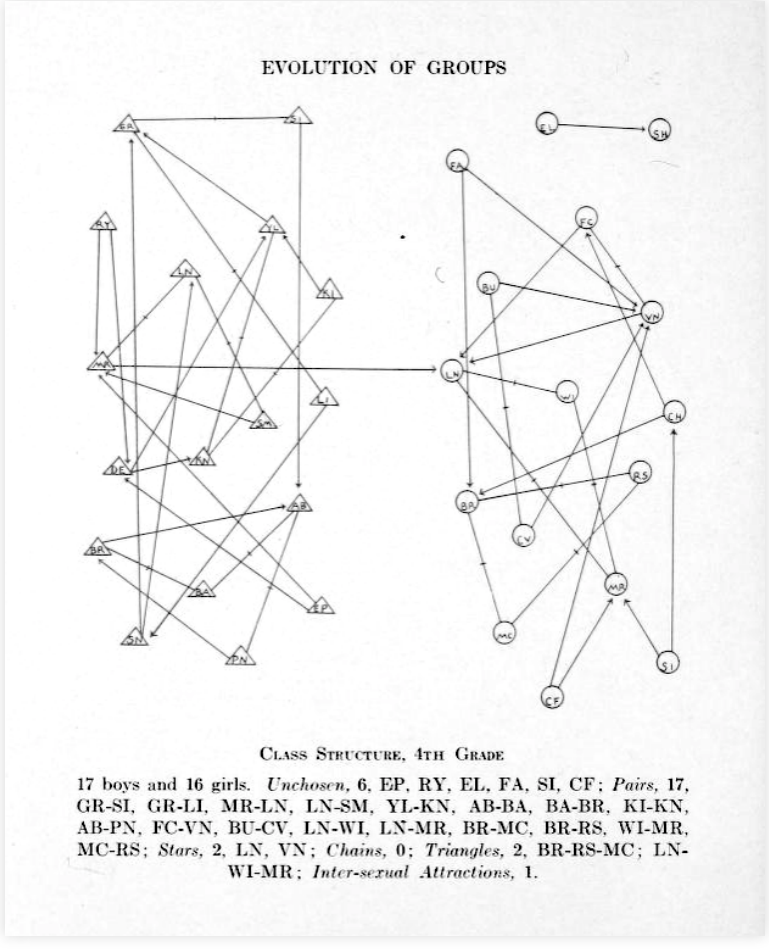


Historical map of Königsberg, highlighting river Pregel and the seven bridges. (Source: Wikimedia Commons)

SOCIAL NETWORK ANALYSIS (SNA)

- SNA started as a bundle of approaches developed in the Social Sciences since the 1930s
- 'Harvard Breakthrough' in the 1960s: sociologists, mathematicians and programmers flesh out a distinct structural-analytical perspective on social phenomena
- 1990s: new perspectives on network structures with the advent of the internet, applications in Physics and Bioinformatics; Literary Studies started to focus on SNA around half a decade ago

ONE OF THE FIRST SOCIOGRAMS



"Class structure, 4th grade". Work by Jacob L. Moreno (1889–1974).
(First published in NYT, 3 April 1933. Source for this PNG: martingrandjean.ch.)

3. DISTANT READING

ARNO SCHMIDT ON THE LIMITS OF READING

"Life is so short! Even if you are a bookworm and only need five days to read a book twice, you will not manage to read more than 70 per annum. And for the 45 years of receptiveness, from age 15 to age 60, they sum up to only 3,150 books: these have to be chosen wisely!"

Arno Schmidt: *Ich bin erst sechzig* (1955). In: Bargfelder Ausgabe, Werkgruppe I, Vol. 4. Zurich: Haffmans 1987, pp. 30. (My trans.)

THE INFAMOUS QUOTE

"[…] if you want to look beyond the canon […], close reading will not do it. It's not designed to do it, it's designed to do the opposite. […] we know how to read texts, now let's learn how *not* to read them. Distant reading: where distance […] *is a condition of knowledge* […]."

Franco Moretti: *Conjectures on World Literature*. In: *New Left Review* 1 (2000).

DISTANT READING – THE BOOK

Franco Moretti: Distant Reading (2013)

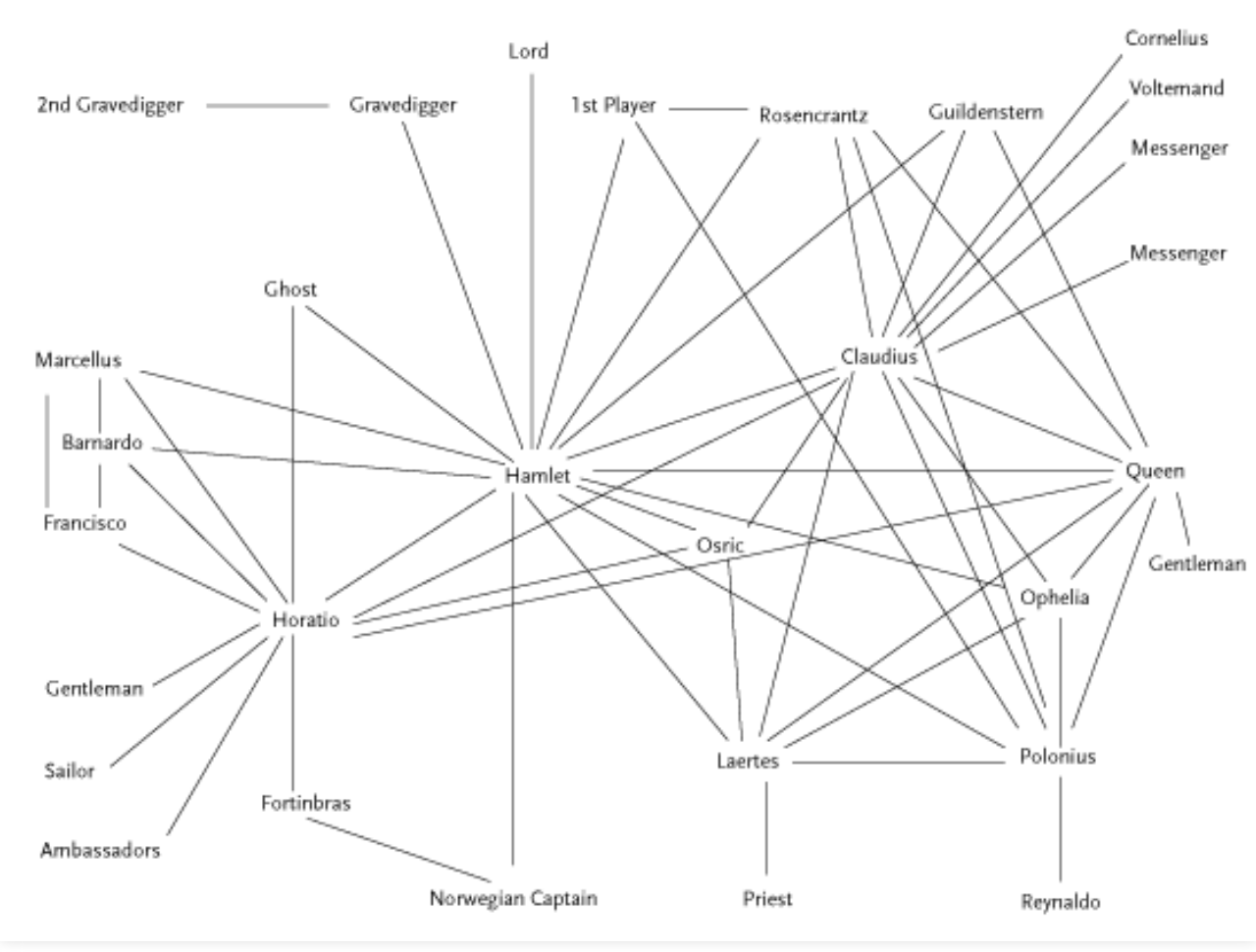
(Russian Translation:)

Франко Моретти: Дальнее чтение (2016)

EXAMPLE:
SCALING UP RESEARCH QUESTIONS

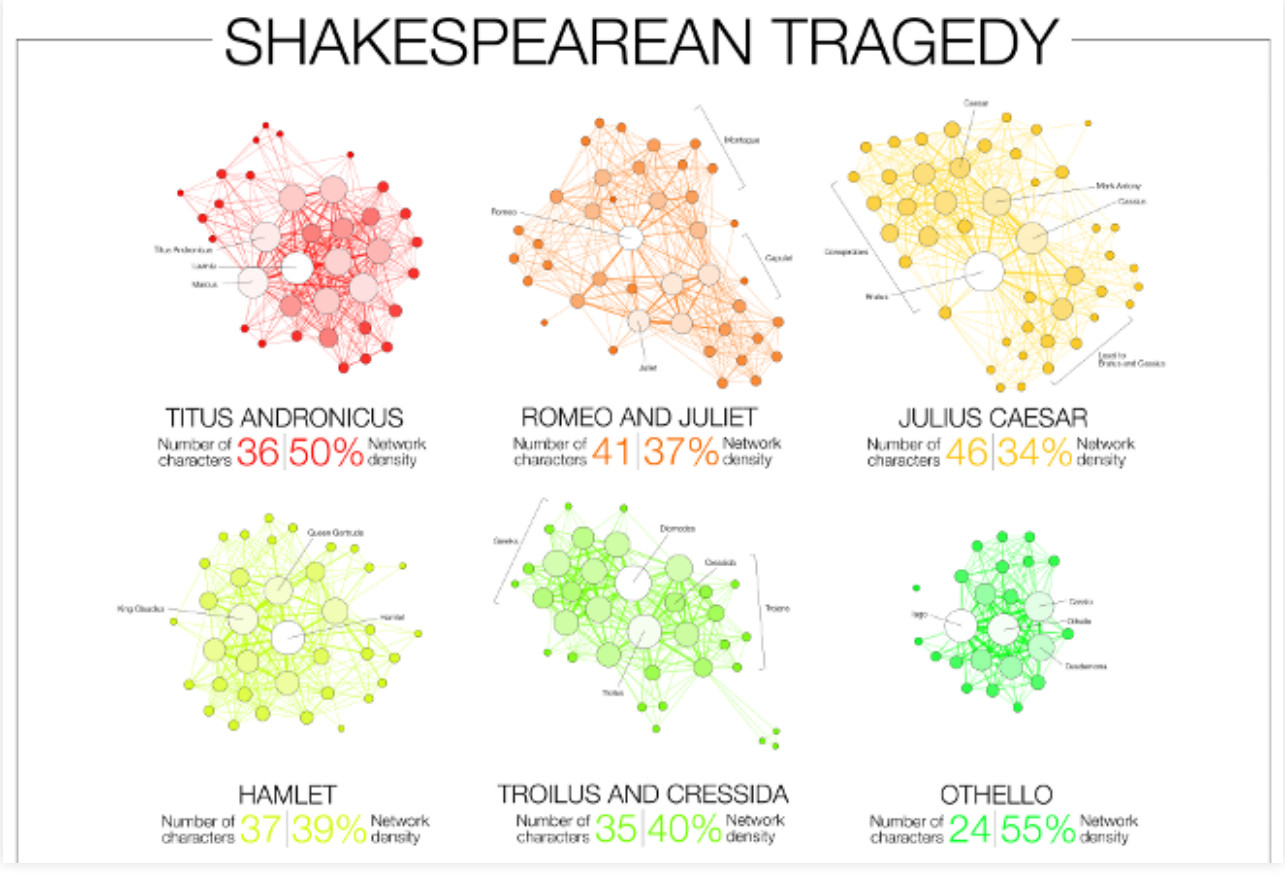
- social network analysis based on simple formalisation (inspired by Solomon Marcus, "Poetica matematică", 1970): two characters are linked to each other if both are performing a speech act in a given segment of a play (act, scene)

MORETTI'S ANALYSIS OF "HAMLET" (2011)



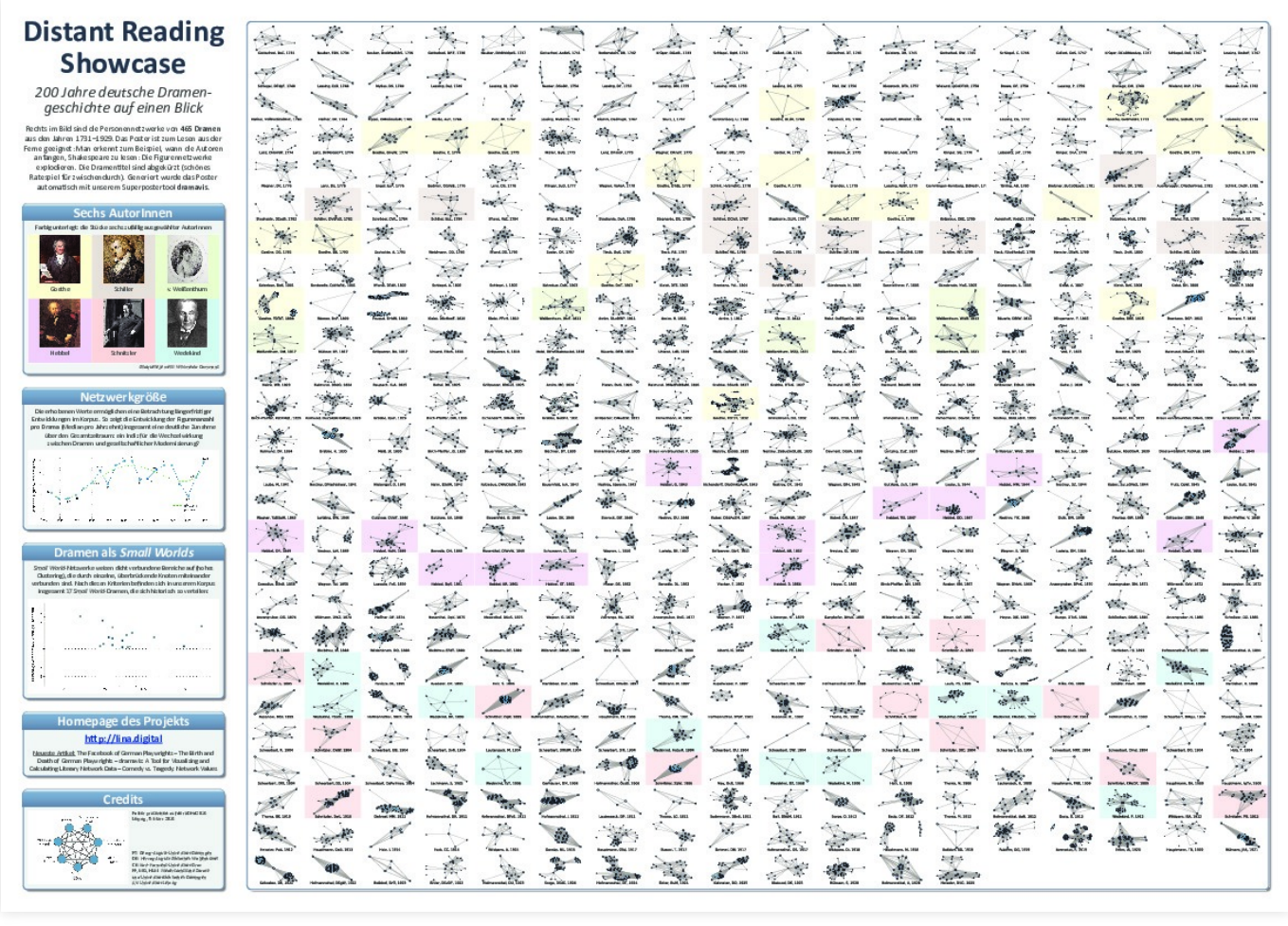
Source: newleftreview.org

"MAPPING SHAKESPEARE'S TRAGEDIES"



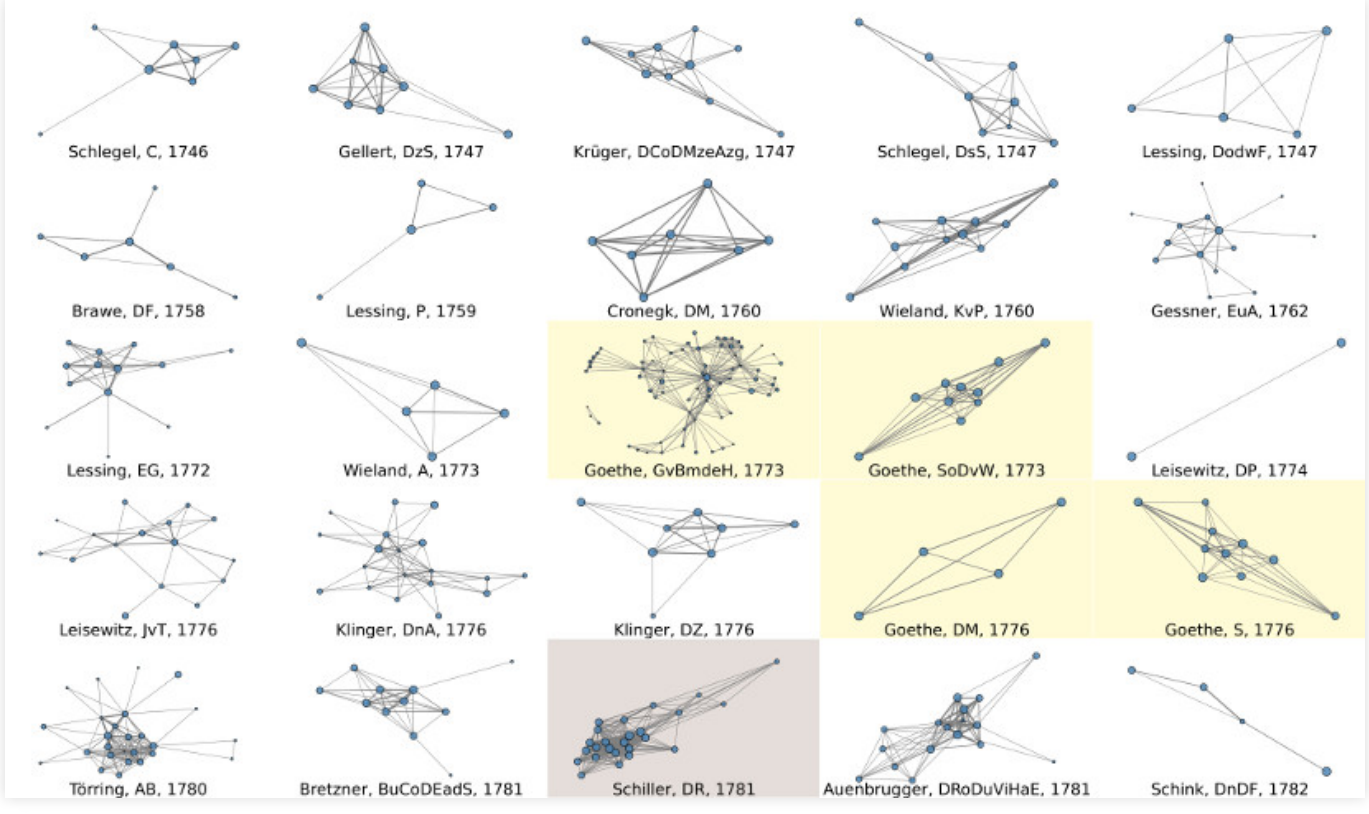
Martin Grandjean's network visualisation of 6 (out of 11) Shakespearean tragedies (Dec., 2015).
Full poster and explanations on Grandjean's website.

"DISTANT-READING SHOWCASE"



"Distant-Reading Showcase", poster presented at #DHd2016 in Leipzig (9 Mar 2016). Download in full-res (28.88 MB) via Figshare. DOI: 10.6084/m9.figshare.3101203.v1.

POSTER DETAIL



Zooming in to Goethe's pivotal play "Götz von Berlichingen" (1773).

4. RUSSIAN DRAMA CORPUS

DRAMA CORPORA IN TEI

- "Théâtre Classique": 1080 plays from the 17th and 18th century
- "Shakespeare His Contemporaries": 860 plays written between 1550 and 1700
- German Drama Corpus: 466 German-language plays from 1730 till 1930
- Letteratura teatrale nella Biblioteca italiana: 171 Italian plays
- Dramawebben: 62 Swedish plays

RUSDRACOR CORPUS METRICS

- **82** plays to date (January 2018) – our goal while growing the corpus is to increase representativeness
- <persName>: **1463** (male: 1023, female: 303)
- rest are uncertain or groups ("Народ", "Голоса")

GET THE CORPUS

```
svn export https://github.com/dracor-org/rusdracor/trunk/tei
```

- 82 files in TEI-XML (~ 15,4 MB)

ALPHA VERSION OF RUSDRACOR

- https://rus.dracor.org/

OUR APPROACH TO LITERARY NETWORK ANALYSIS (1)

- following older structuralist approaches in Literary Studies (Barthes 1972, Lotman 1977), but automatising data collection
- long-term objective: provide data to describe the structure and evolution of different compositional types of plays

OUR APPROACH TO LITERARY NETWORK ANALYSIS (2)

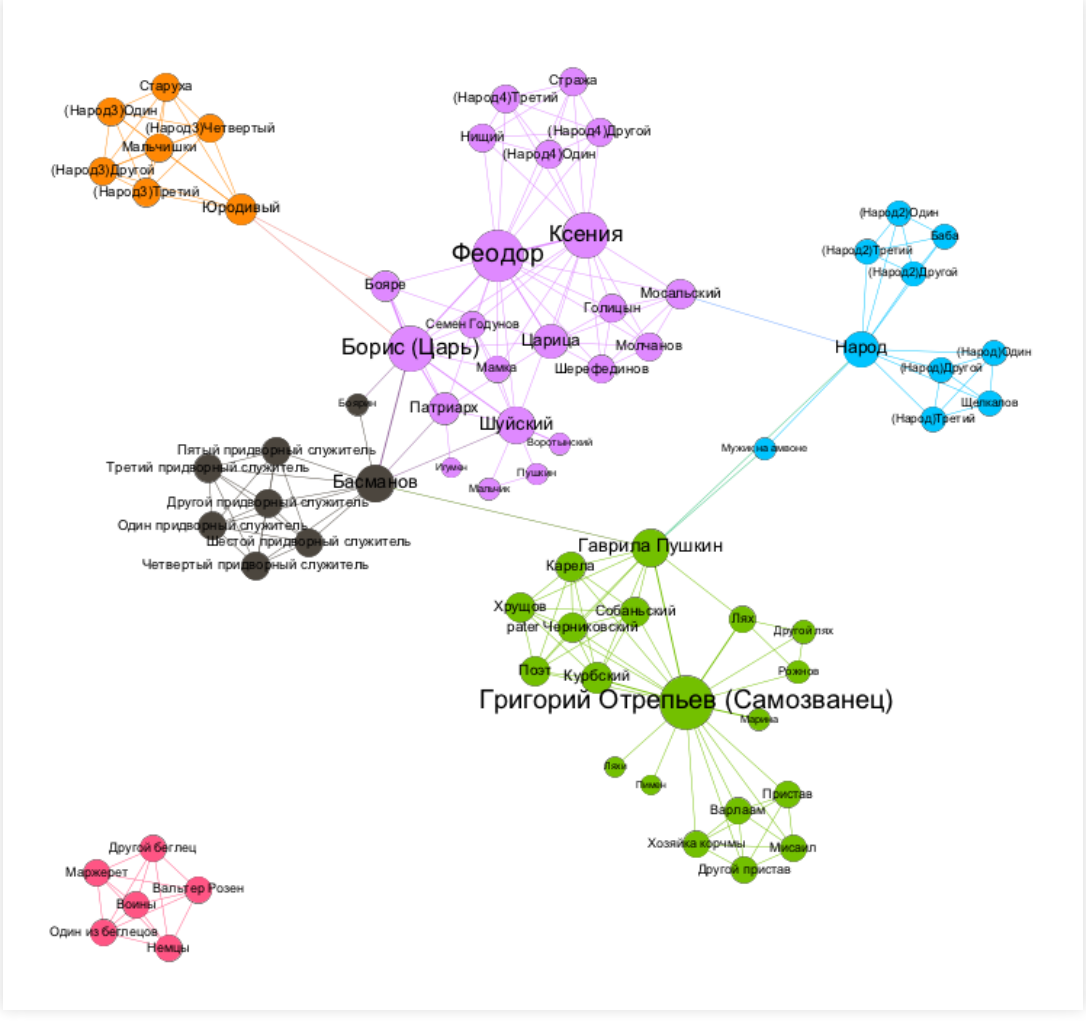
- network data = interactions between characters
- operationalisation of 'interaction': "Two characters interact with one another if they perform a speech act within the same segment of a drama (usually a 'scene')." (following Solomon Marcus, "Poetica matematică", 1970)

EXTRACT FROM PUSHKIN'S "BORIS GODUNOV"

```
(…)
<text>
 <front>
  <docTitle>
   <titlePart type="main">Борис Годунов</titlePart>
  </docTitle>
  <div type="dedication">
   <p>Драгоценной для россиян памяти Николая Михайловича
     Карамзина сей труд, гением его вдохновенный, с
     благоговением и благодарностию посвящает</p>
   <p>Александр Пушкин</p>
  </div>
 </front>
 <body>
  <div type="scene">
   <head>КРЕМЛЕВСКИЕ ПАЛАТЫ</head>
```

(encoded in TEI)

NETWORK GRAPH FOR "BORIS GODUNOV"



(extracted with our TEI2CSV converter, visualised in Gephi)

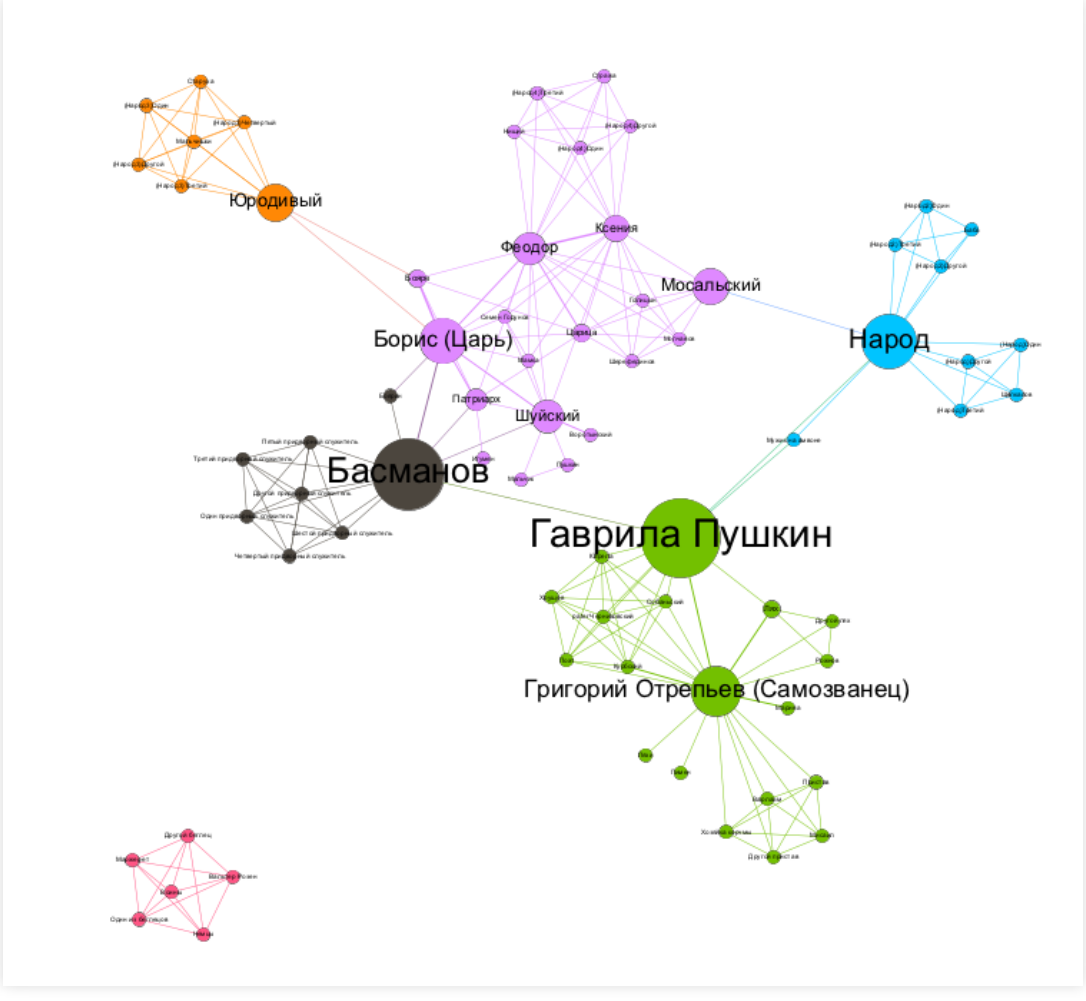
"BORIS GODUNOV", SOME METRICS PER CHARACTER

Character	Degree	Betweenness Centrality
Григорий Отрепьев (Самозванец)	**18**	537.0
Феодор	17	276.28
Ксения	15	187.81
Народ	11	606.93
Гаврила Пушкин	11	**976.12**
Басманов	11	862.69
Борис (Царь)	11	484.57
Шуйский	10	282.14

BETWEENNESS CENTRALITY

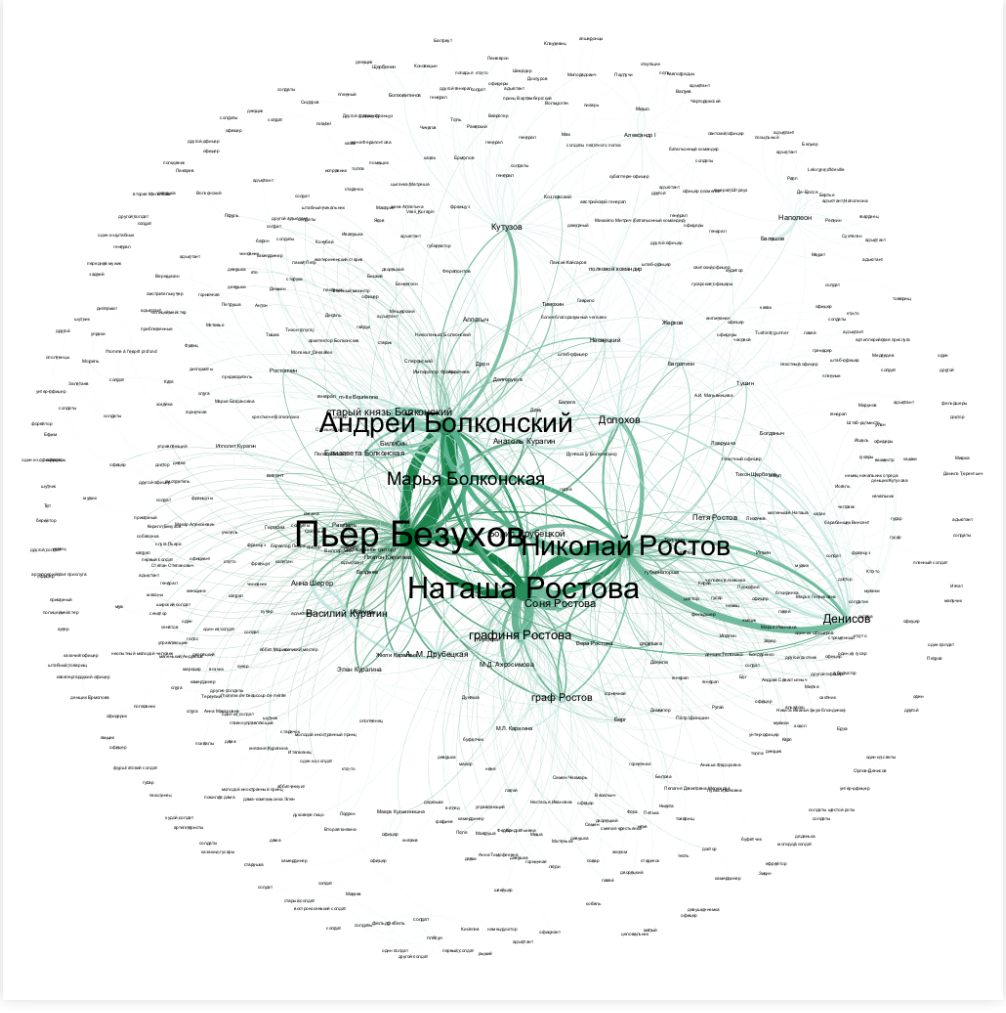
- "a measure of centrality in a graph based on shortest paths"
- "For every pair of vertices in a graph, there exists a shortest path between the vertices such that either the number of edges that the path passes through (for undirected graphs) or the sum of the weights of the edges (for directed graphs) is minimized. The betweenness centrality for each node is the number of these shortest paths that pass through the node." (Wikipedia)

NETWORK GRAPH FOR "BORIS GODUNOV"



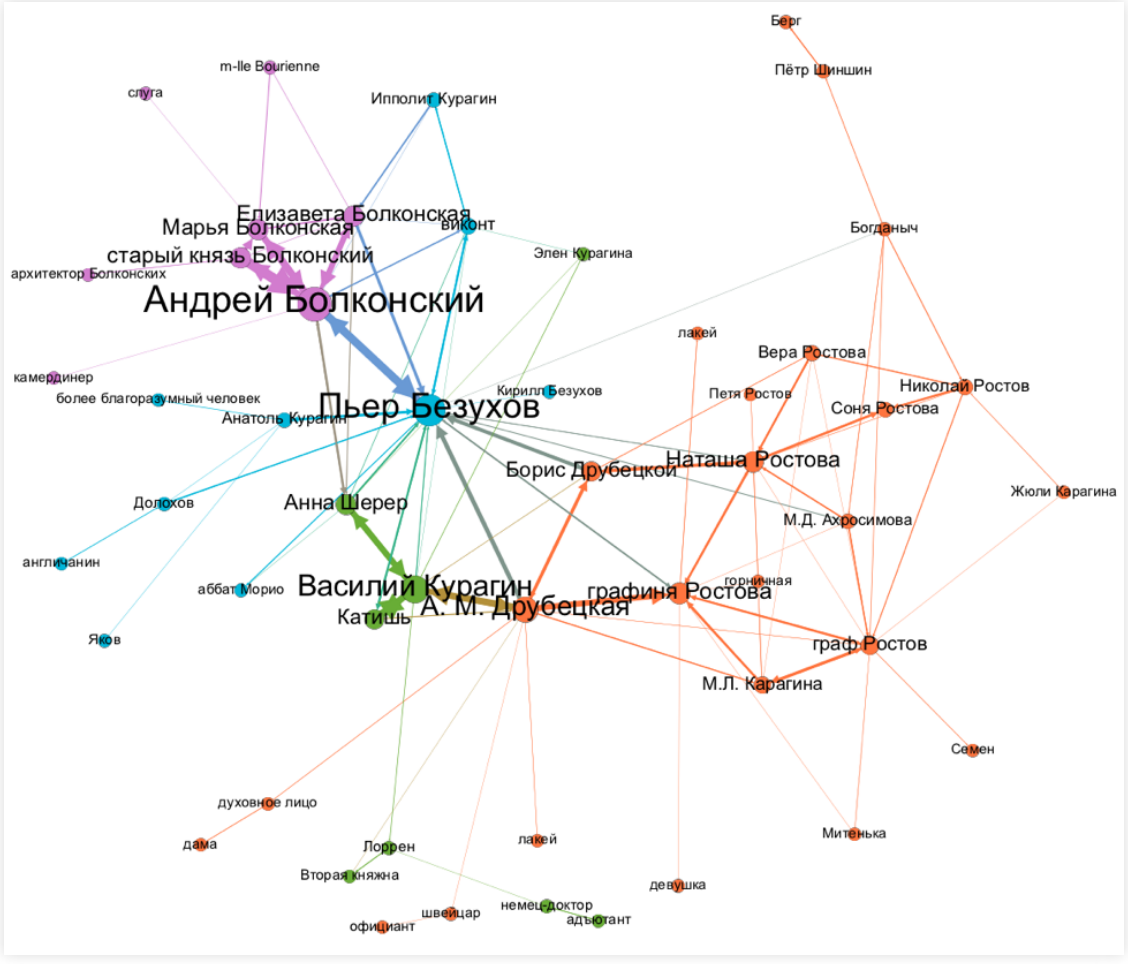
(visualisation based on values for Betweenness Centrality)

"WAR AND PEACE" (1/6)



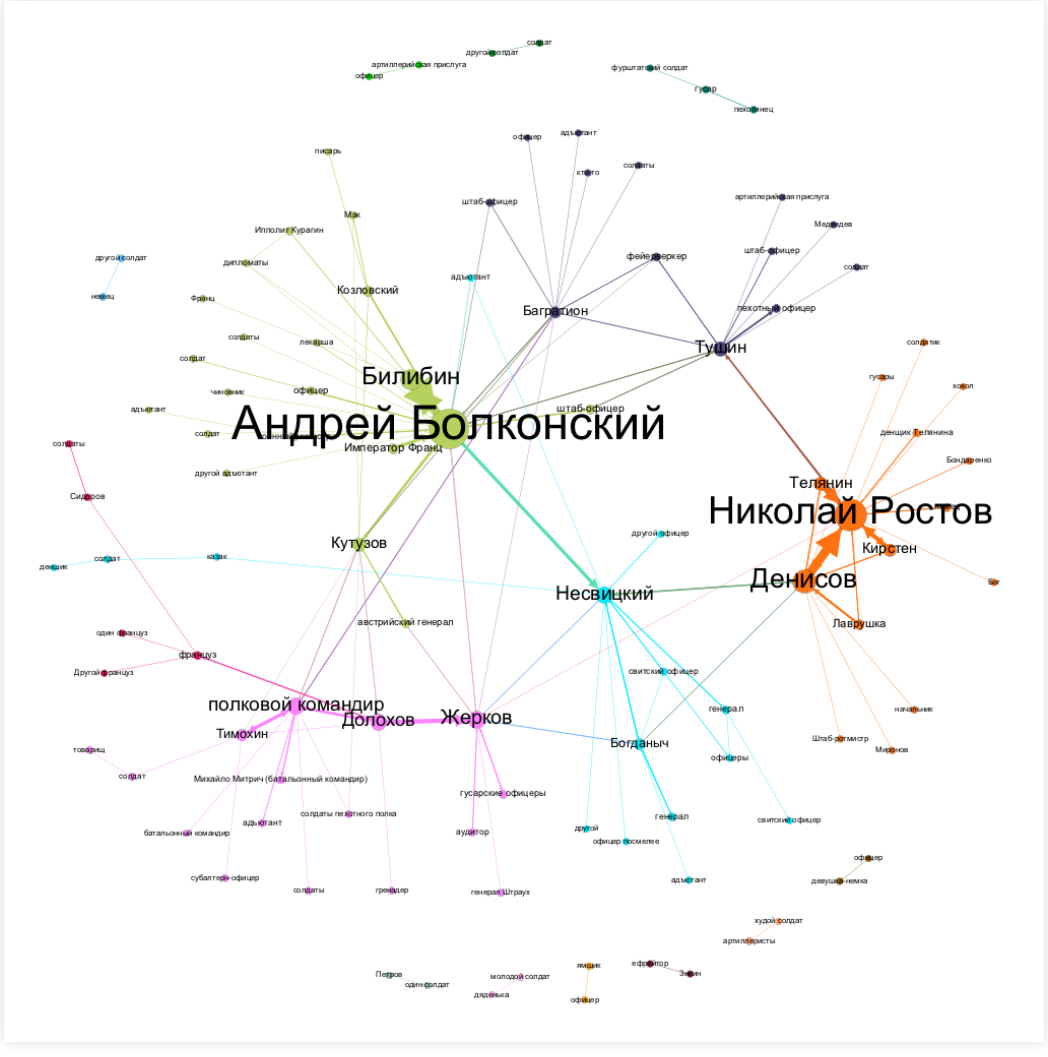
(conversational graph)

"WAR AND PEACE" (2/6)



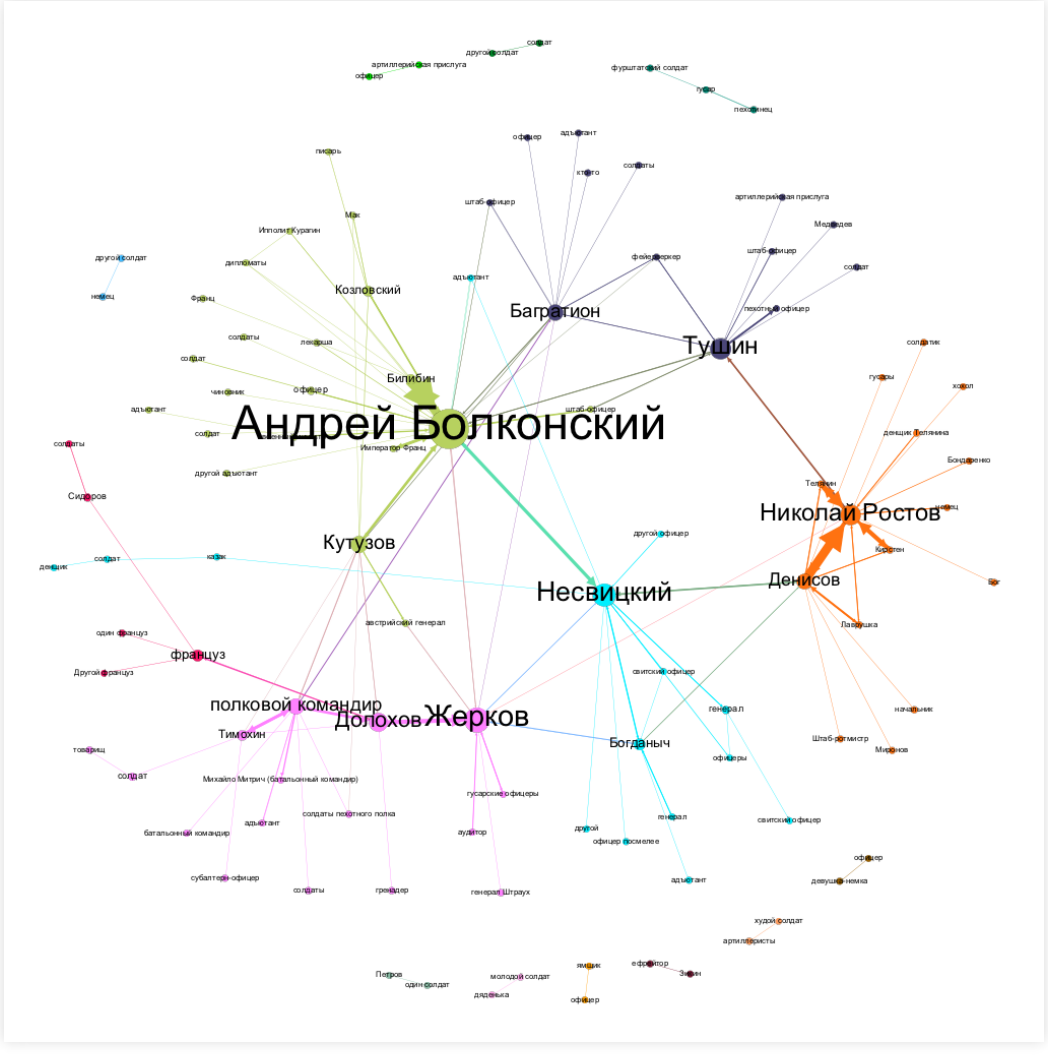
(часть 1 т. I)

"WAR AND PEACE" (3/6)



(часть 1 т. II)

"WAR AND PEACE" (4/6)



(часть 1 т. II: betweenness centrality)

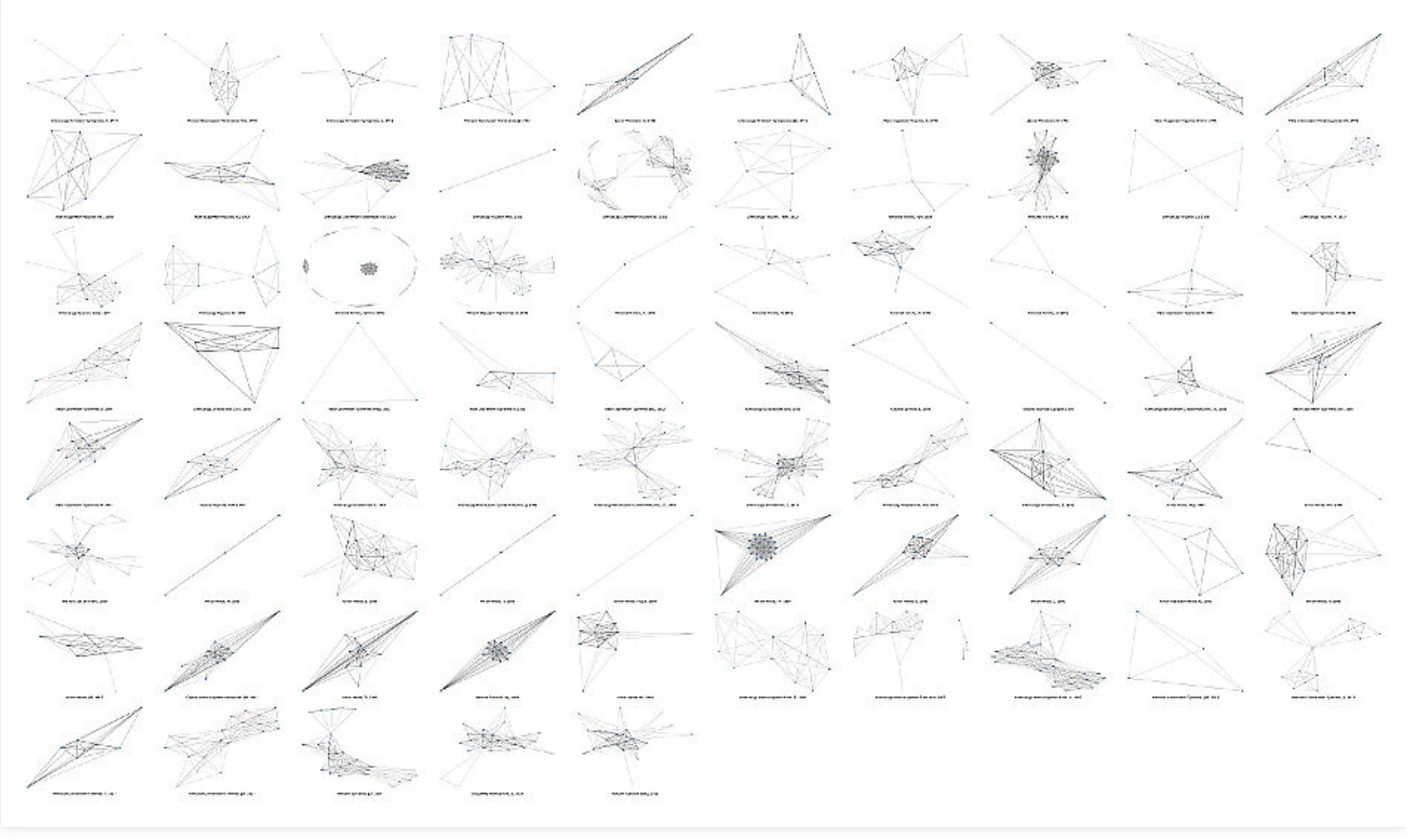
"WAR AND PEACE" (5/6)

(Война и мир целиком)

"WAR AND PEACE" (6/6)

(Война и мир целиком)

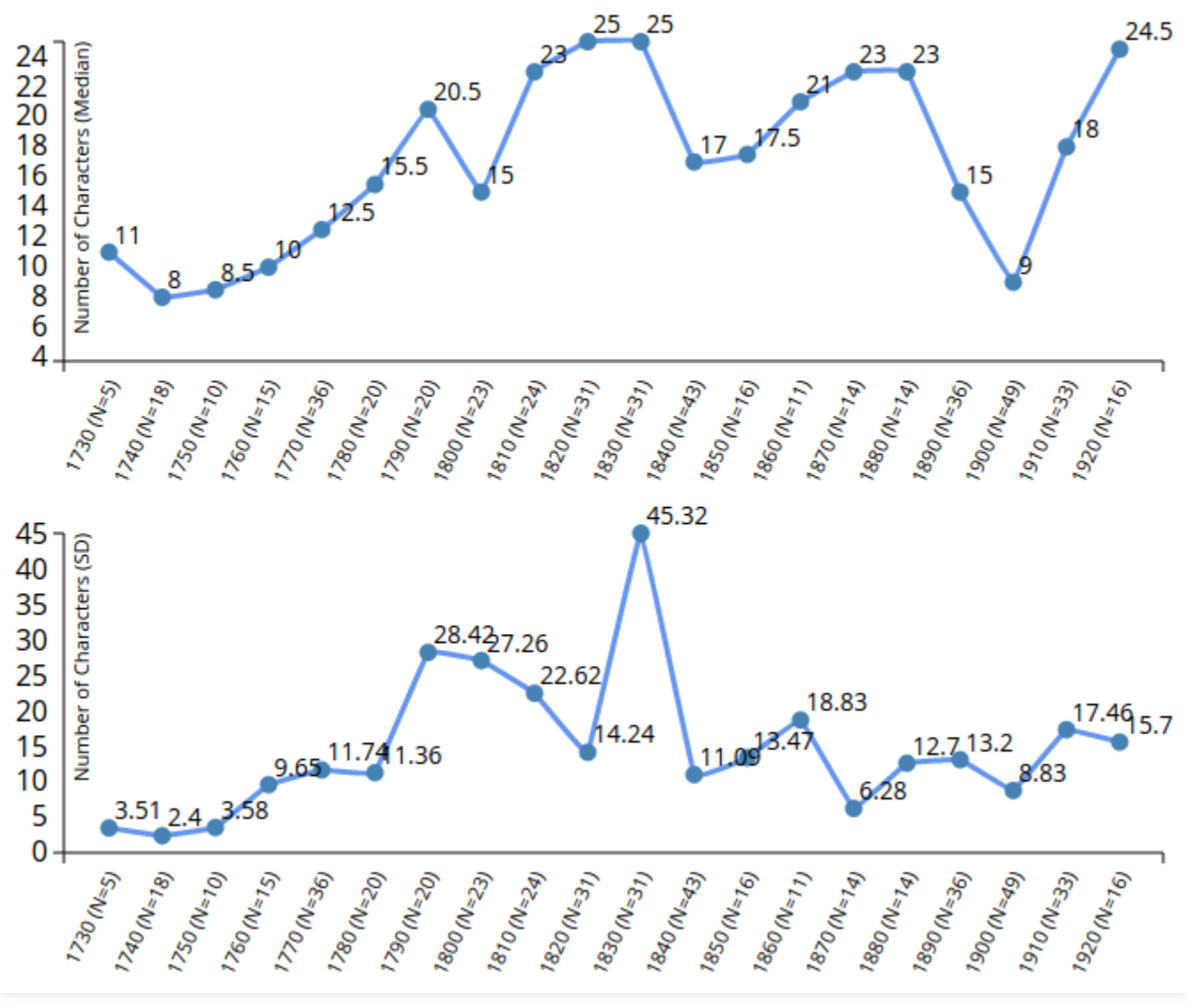
RUSSIAN SUPERPOSTER



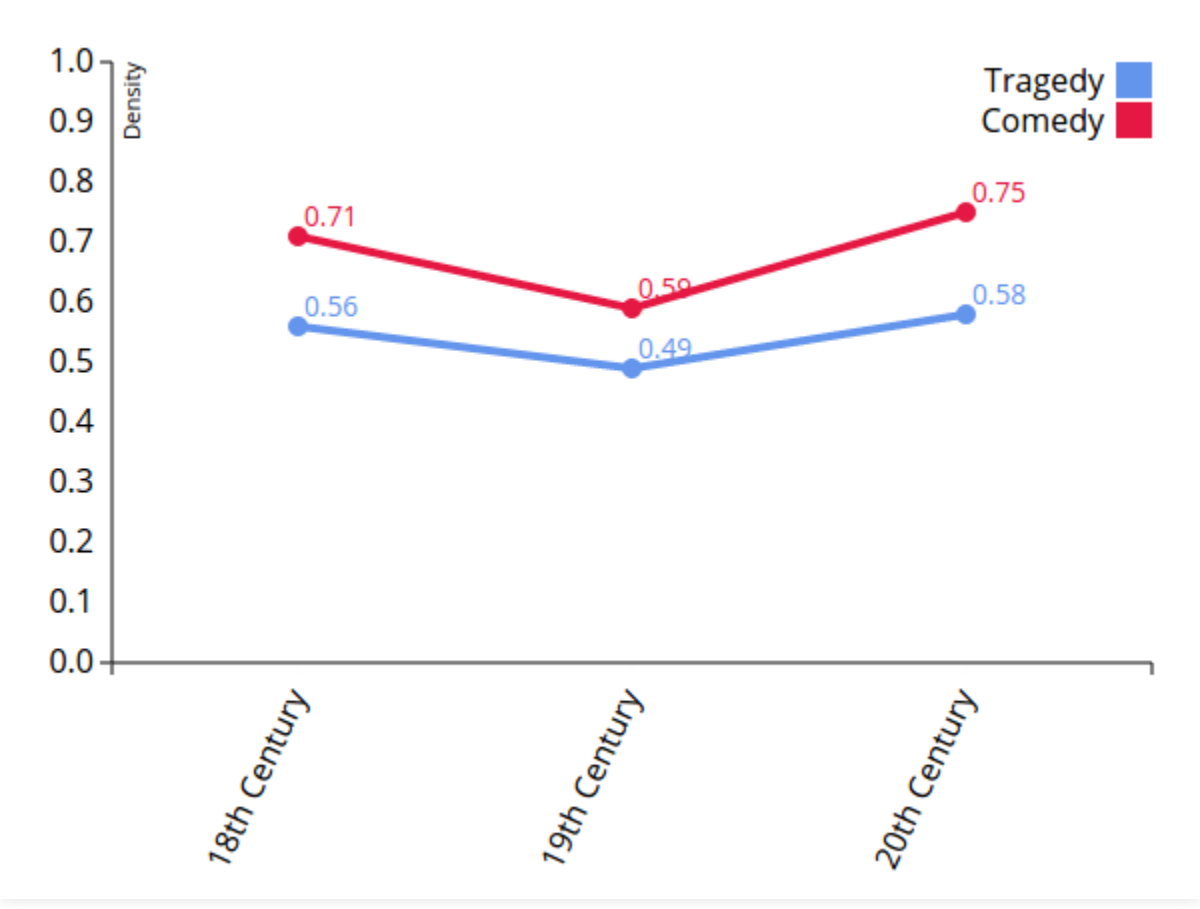
Social networks extracted from 75 Russian plays (1740–1930).

5. STRUCTURAL EVOLUTION

NETWORK SIZES AND SD (GERDRACOR)



NETWORK DENSITIES (GERDRACOR)



6. CASE STUDY: SMALL WORLDS

TYPES OF DRAMA NETWORKS

- Dramatic texts are context-sensitive aesthetic models of social formations, which means that …
  - … dramas represent social formations (e.g., nuclear family, royal court, 'society');
  - … these social formations only exist in their aesthetic representation, as models;
  - … these models are potentially context-sensitive and interact with real social formations.

'SMALL WORLD' NETWORKS

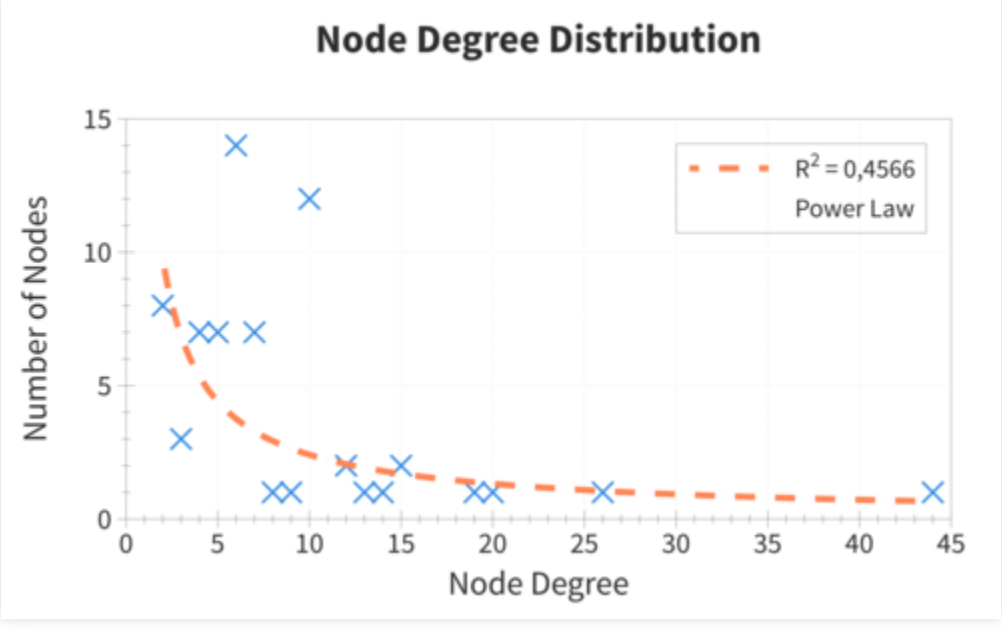
- "widespread in biological, social and man-made systems" (Watts & Strogatz 1998, 442)
- "highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs" (Watts & Strogatz 1998, 440)
- already applied on dramatic texts (Shakespeare): Stiller, Nettle & Dunbar 2003; Stiller & Hudson 2005

'SMALL WORLD' CRITERIA (1/2)

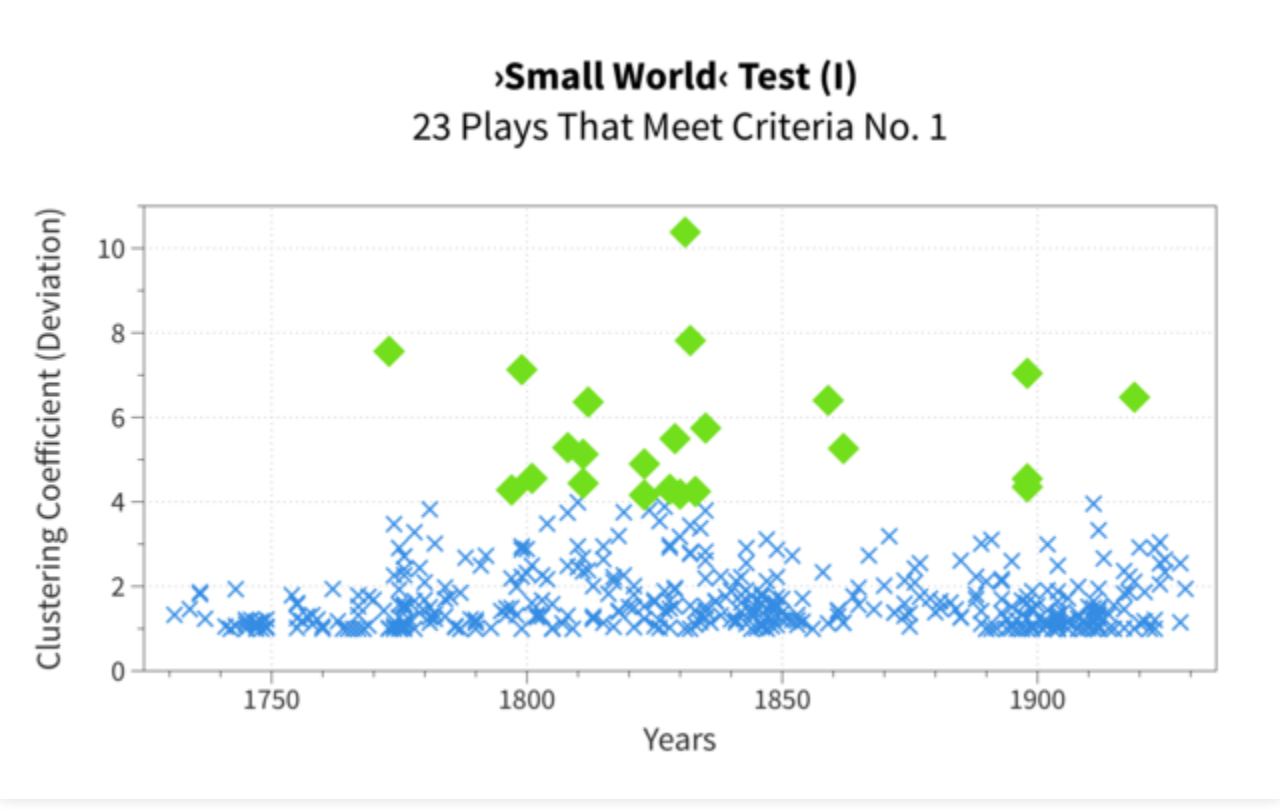
- Criterion 1: The clustering coefficient (C) of an observed network, in our case the character network of a dramatic text, is significantly higher than the C of a corresponding random network.
- Criterion 2: The average path length (APL) of an observed network does not differ significantly from the APL of a corresponding random network.

'SMALL WORLD' CRITERIA (2/2)

- Criterion 3: 'scale free' (variant of 'small world' networks described by Albert & Barabási 2002).
- 'Scale free' networks feature a node-degree distribution following a power law:

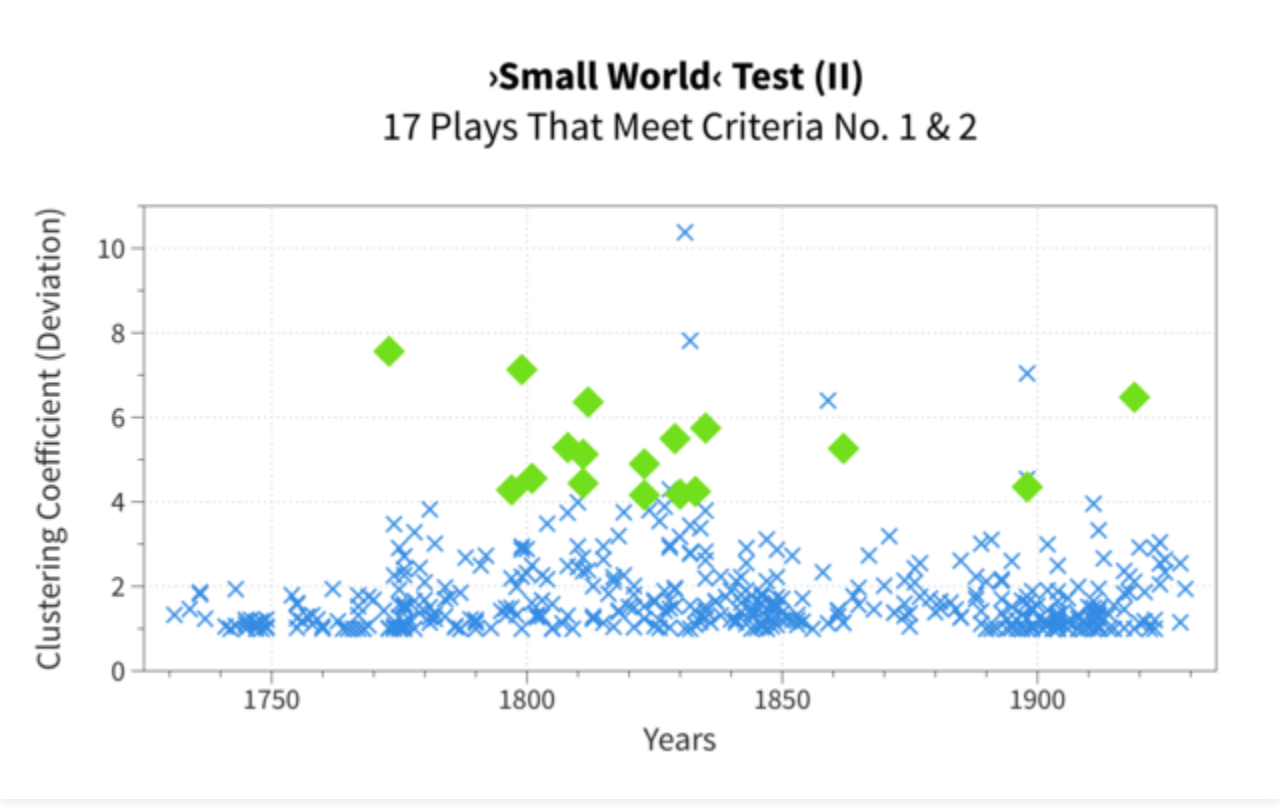


CRITERION 1



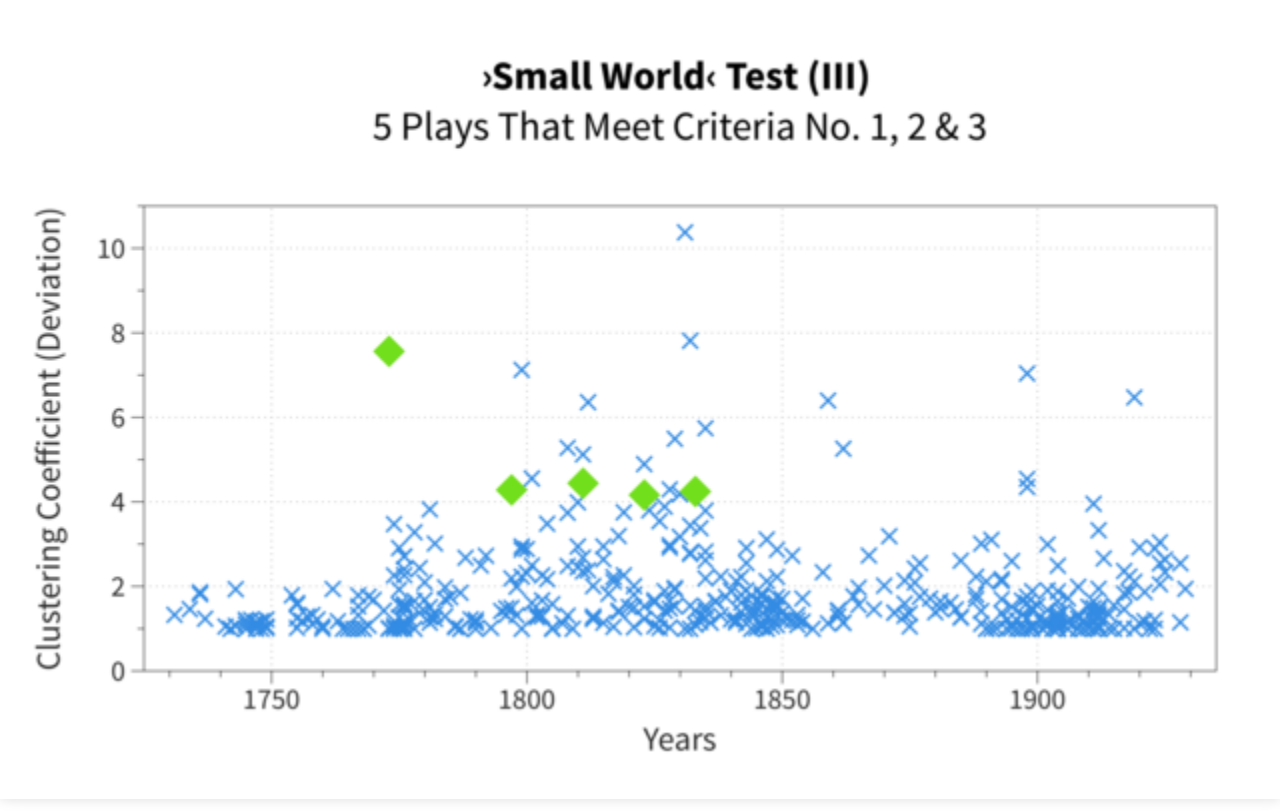
- Formed quotient of clustering coefficients of each individual play and the mean clustering coefficient of 1,000 random networks → identified all dramas where quotient was significantly higher (i.e., bigger than Mean + 2 × SD).

CRITERION 2



- Further exclusion of all plays with average path length significantly different from that of the random network (i.e., smaller than Mean – 2 × SD or bigger than Mean + 2 × SD, respectively).

CRITERION 3



- Further exclusion of all plays not showing power-law regression in their node-degree distribution.

7. TOOLCHAIN

TOOLCHAIN

- HTML to TEI converter (Beautiful Soup, etc.)
- Oxygen XML Editor for correction and maintenance
- ezlinavis (Easy Linavis) for simple formalisations in class: https://ezlinavis.dracor.org/
- dramavis (Python script collection): https://github.com/lehkost/dramavis
- dracor.org (drama corpora maintained by ourselves, includes an API): https://dracor.org/
- Shiny App: https://shiny.dracor.org/

ACKNOWLEDGEMENTS

This talk was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017–2018 (grant No 17-05-0054) and by the Russian Academic Excellence Project "5-100".