STYLOMETRY

THE IDEA

«Measuring» authorial style

(Mendenhall, 1887)

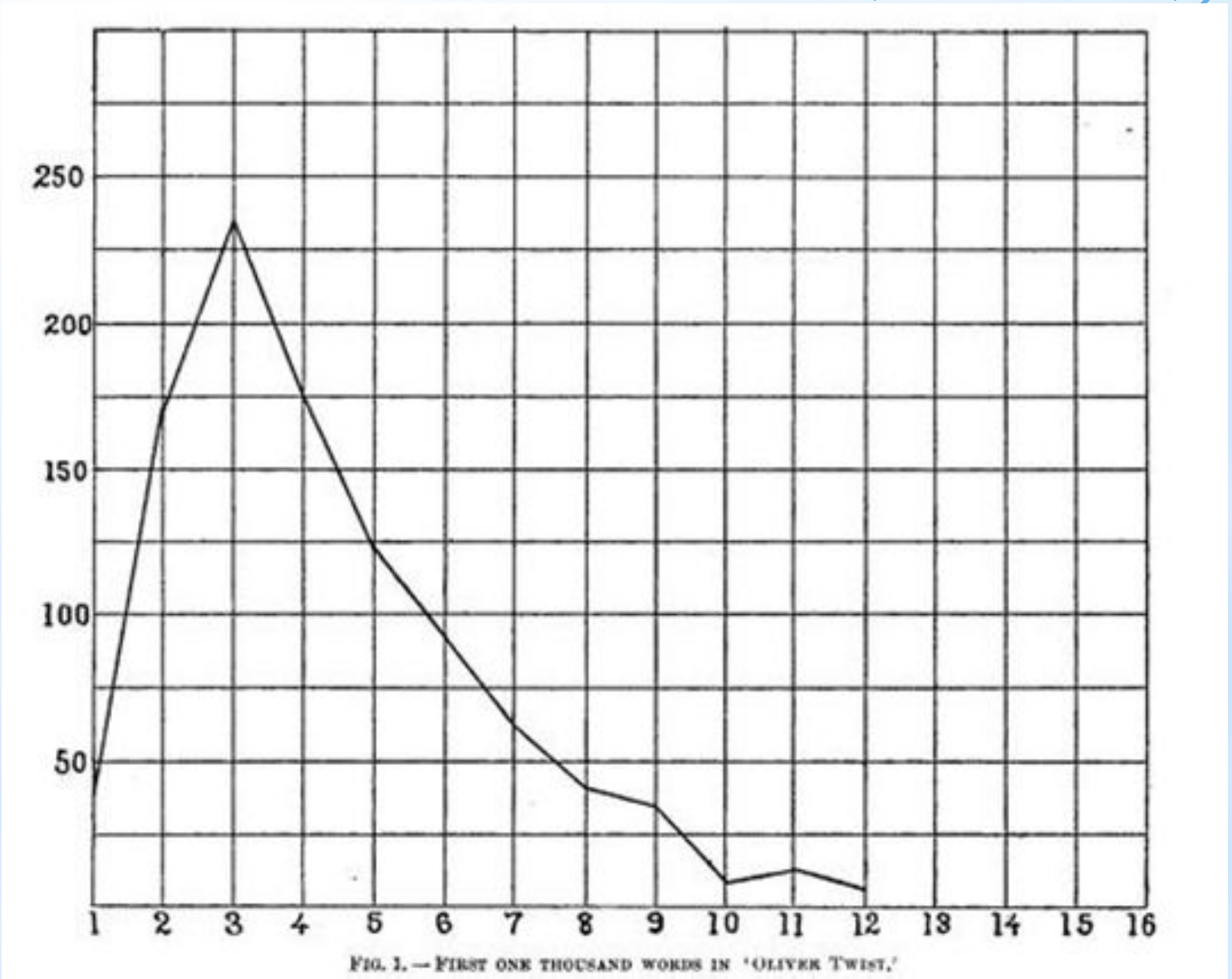


FIG. 1. — FIRST ONE THOUSAND WORDS IN 'OLIVER TWIST.'

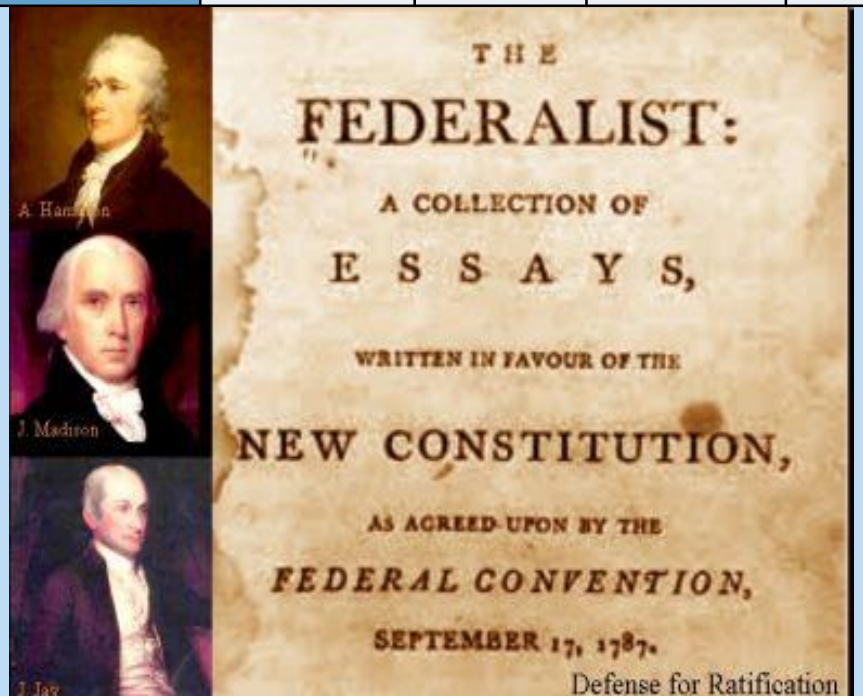
THE IDEA

«Measuring» authorial style

Successes

(Mosteller and Wallace 1964)

	enough	while	whilst	upon
Hamilton	0.59	0.26	0	2.93
Madison	0	0	0.47	0.16
Disputed texts	0	0	0.34	0.08
Co-authored texts	0.18	0	0.36	0.36



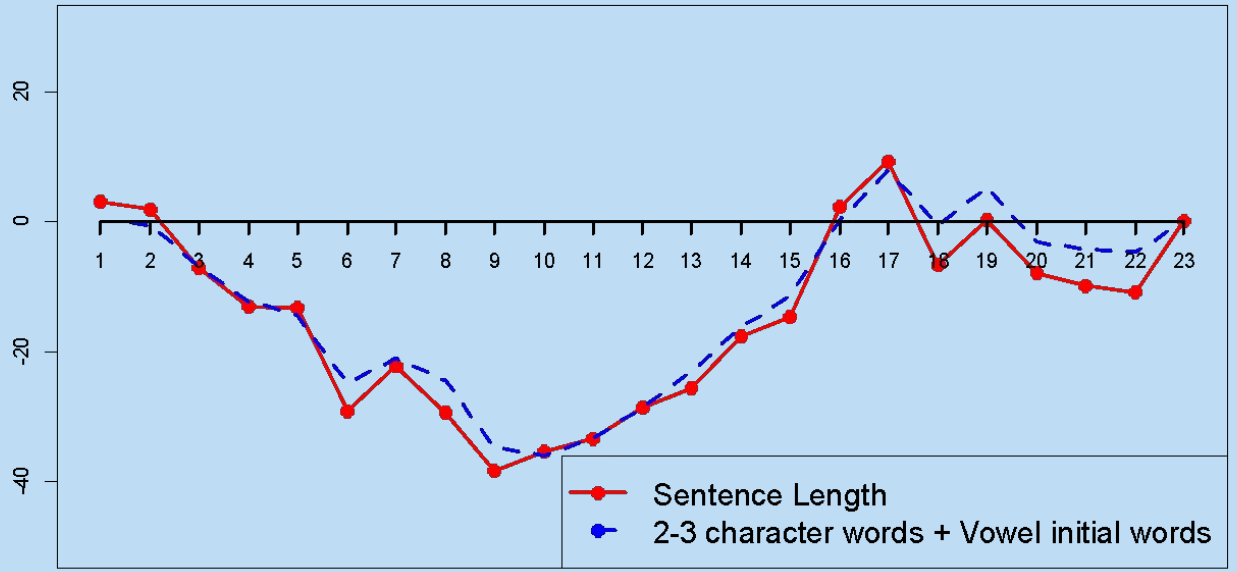
THE IDEA

«Measuring» authorial style

Failures

Andrew Morton in the early ‘60 adapted **Cu**mulative **Sum** – CUSUM or QSUM

During a BBC live show (1993):
Documents of convicted criminals were attributed to … the Secretary of State for Justice!!!



THE (PLETHORA OF) METHODS FOR STYLOMETRY AND AUTHORSHIP ATTRIBUTION

- Character-level analysis
- Syntax-level analysis
- Multi-method analysis (e.g. JGAAP, PAN competition software…)
- …and many others
- In this lesson, just two methods:
  - Delta method (for authorship attribution)
  - Zeta method (for the quantitative analysis of style)

WORD-FREQUENCY BASED STYLOMETRY

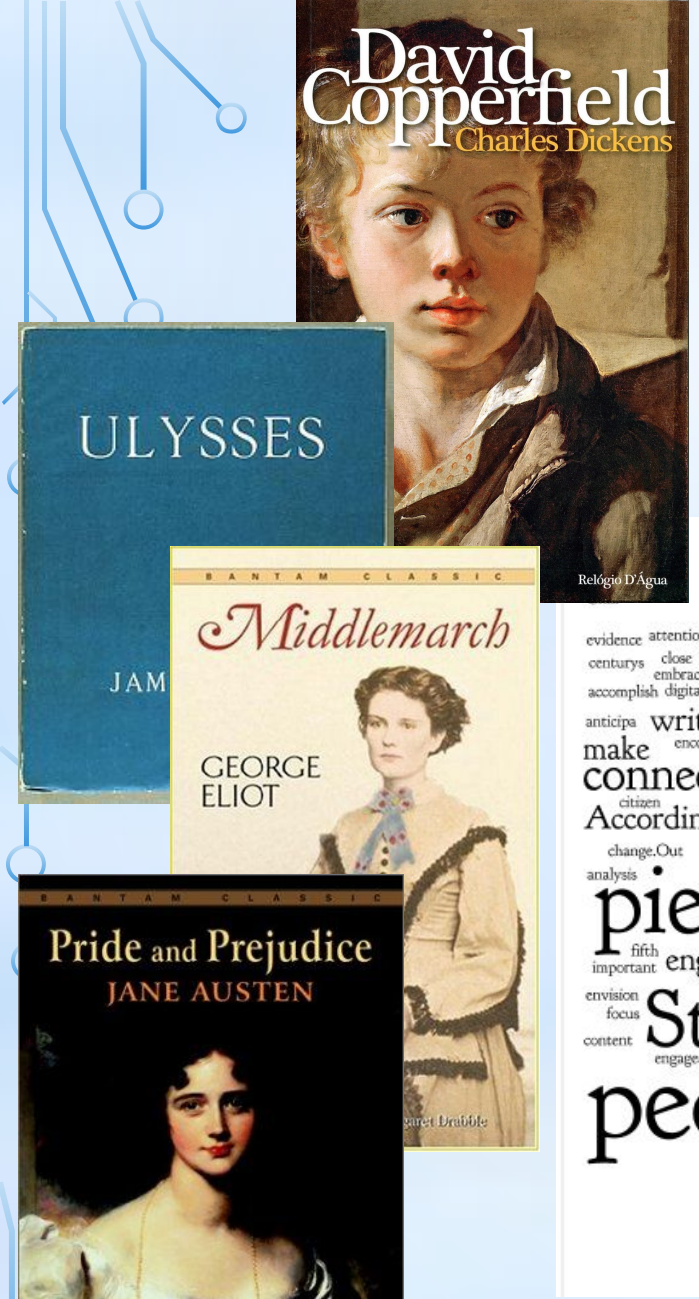
'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship[1]

John Burrows
University of Newcastle, Australia

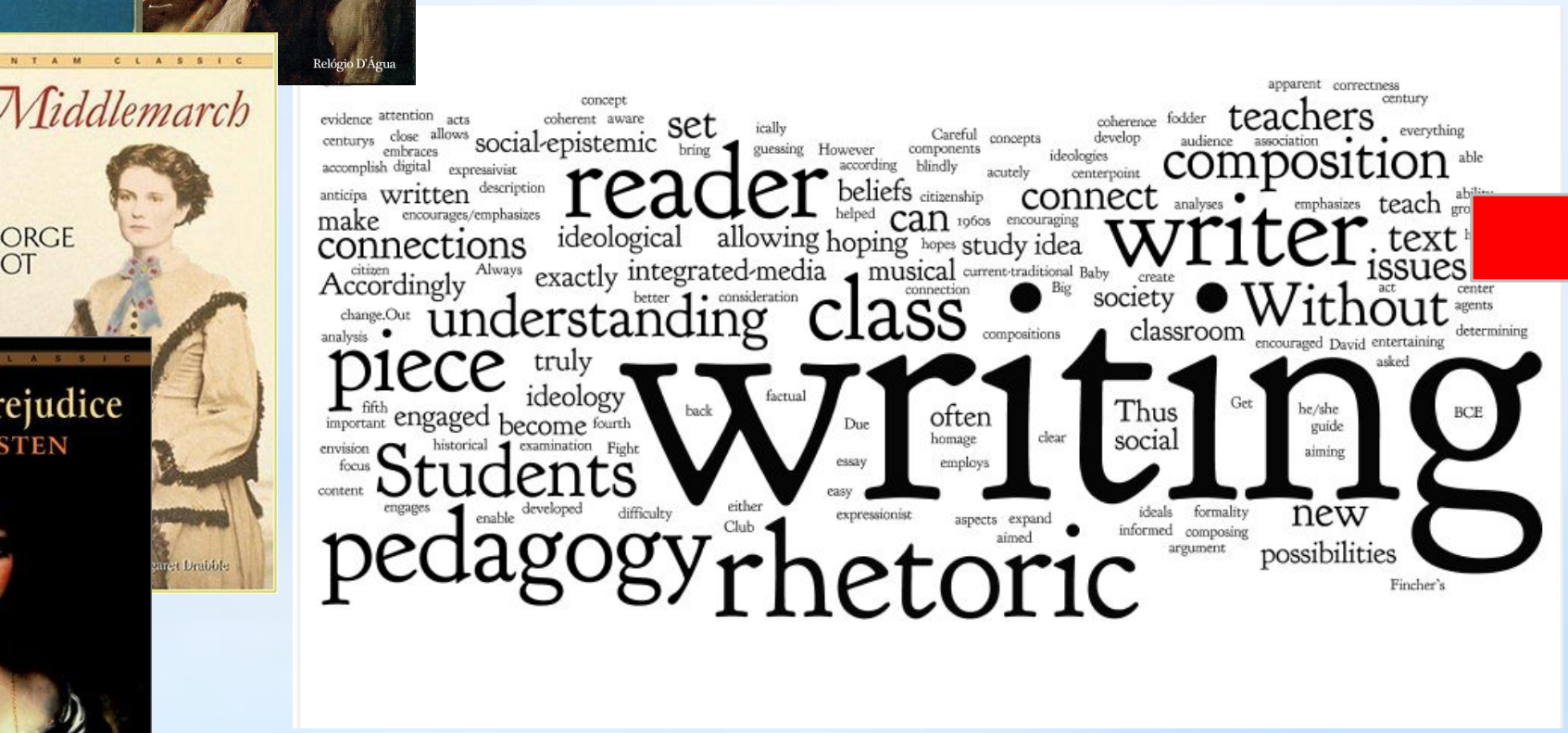
"Literary and Linguistic Computing" 17, no. 3 (2002): 267–87

Abstract

This paper is a companion to my 'Questions of authorship: attribution and beyond', in which I sketched a new way of using the relative frequencies of the very common words for comparing written texts and testing their likely authorship. The main emphasis of that paper was not on the new procedure but on the broader consequences of our increasing sophistication in making such comparisons and the increasing (although never absolute) reliability of our inferences about authorship. My present objects, accordingly, are to give a more complete account of the procedure itself; to report the outcome of an extensive set of trials; and to consider the strengths and limitations of the new procedure. The procedure offers a simple but comparatively accurate addition to our current methods of distinguishing the most likely author of texts exceeding about 1,500 words in length. It is of even greater value as a method of reducing the field of likely candidates for texts of as little as 100 words in length. Not unexpectedly, it



DELTA DISTANCE



1. the
2. and
3. of
4. to
5. a
6. i
7. in
8. he
9. was
10. it
11. that
12. you
13. his
14. her
15. with
16. as
17. had
18. she
19. for

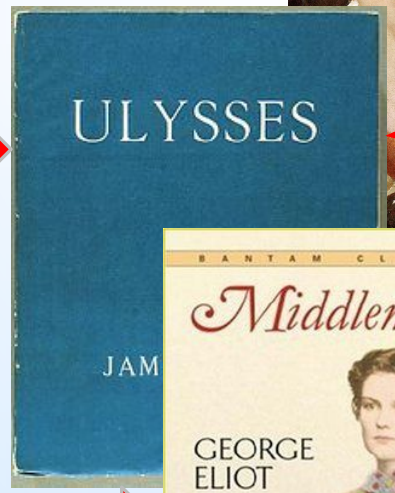
1. the
2. and
3. of
4. to
5. a
6. i
7. in
8. he
9. was
10. it
11. that
12. you
13. his
14. her
15. with
16. as
17. had
18. she
19. for

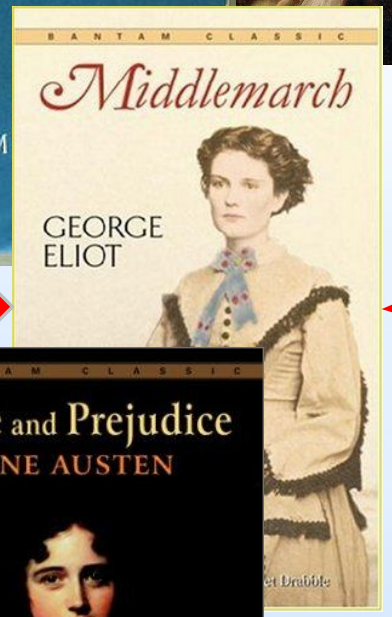
5.1%
3.2%
2.4%
2.5%

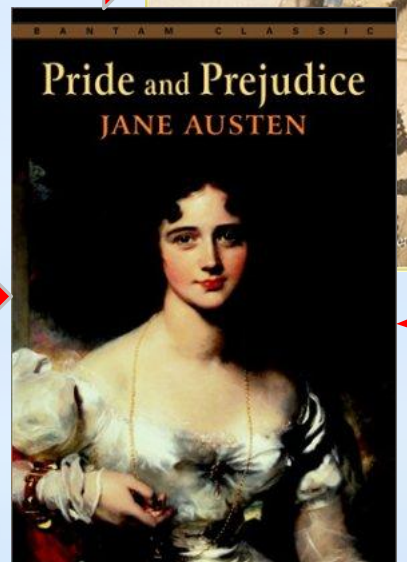
4.1%
3.3%
2.2%
2.7%

3.1%
4.2%
1.4%
1.2%

5.2%
3.2%
2.4%
2.5%







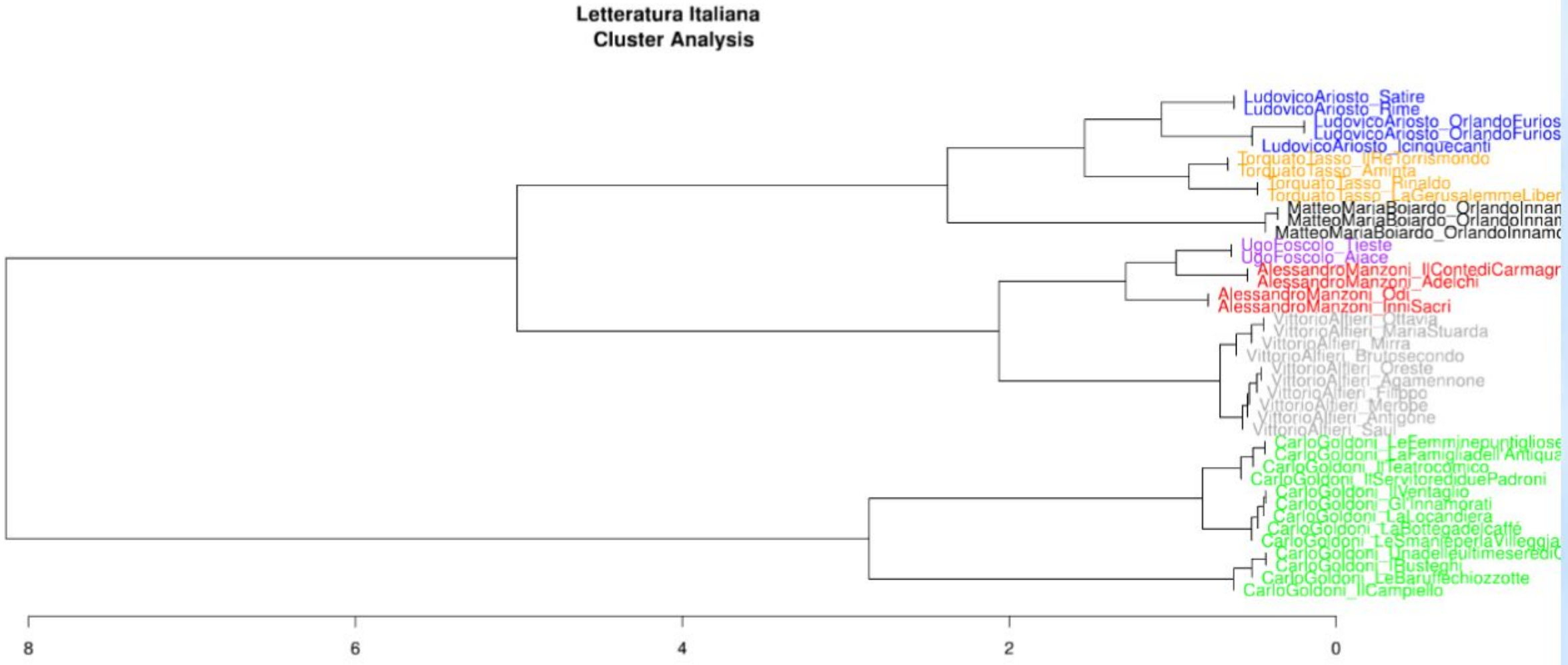
	A	B	C	D	E	F
1		AlessandroManzoni_Adelchi	AlessandroManzoni_IlContediCarmagnola	AlessandroManzoni_InniSacri	AlessandroManzoni_Odi	AlessandroManzoni_Poesiegio
2	AlessandroManzoni_Adelchi	0	0,481290655	0,666926925	0,738545533	0,5688
3	AlessandroManzoni_IlContediCarmagnola	0,481290655	0	0,746348745	0,814261157	0,654
4	AlessandroManzoni_InniSacri	0,666926925	0,746348745	0	0,633663965	0,634
5	AlessandroManzoni_Odi	0,738545533	0,814261157	0,633663965	0	0,733
6	AlessandroManzoni_Poesiegiovanili	0,568820863	0,654375023	0,634854567	0,733827682	
7	CarloGoldoni_Gl'Innamorati	0,980786338	0,936018177	1,013723738	1,101305203	0,950
8	CarloGoldoni_IlCampiello	1,016924762	1,031300757	1,018625104	1,092680684	0,929
9	CarloGoldoni_IlServitorediduePadroni	0,94860233	0,926662976	0,976288639	1,080804722	0,918
10	CarloGoldoni_IlTeatrocomico	0,915941412	0,896367382	0,971870697	1,085346366	0,898
11	CarloGoldoni_IlVentaglio	1,011953514	1,00041649	1,074888328	1,131792245	0,997
12	CarloGoldoni_IRusteghi	1,089096895	1,124315967	1,047451935	1,1240649	0,977
13	CarloGoldoni_LaBottegadelcaffé	0,997940632	0,980781404	1,069965126	1,139058754	0,993
14	CarloGoldoni_LaFamigliadell'Antiquario	0,97647637	0,968110166	1,038499373	1,080510085	0,953
15	CarloGoldoni_LaLocandiera	0,97946604	0,952399004	1,052505983	1,110322738	0,956
16	CarloGoldoni_LeBaruffechiozzotte	1,051753673	1,103993387	1,018834132	1,082447143	0,942
17	CarloGoldoni_LeFemminepuntigliose	0,940334542	0,938723973	1,008461186	1,076438004	0,917
18	CarloGoldoni_LeSmanieperlaVilleggiatura	1,023938091	0,964832878	1,056736183	1,148650567	1,007
19	CarloGoldoni_UnadelleultimerediCarnovale	1,045847956	1,085480986	1,047945641	1,10681856	0,948
20	VittorioAlfieri_Agamennone	0,684514153	0,743793265	0,829452563	0,905939302	0,7
21	VittorioAlfieri_Antigone	0,73781244	0,801189414	0,824156384	0,91495815	0,721
22	VittorioAlfieri_Brutosecondo	0,675393312	0,675937144	0,830722082	0,910174086	0,668
23	VittorioAlfieri_Filippo	0,69672213	0,73856813	0,806194725	0,93419818	0,669
24	VittorioAlfieri_MariaStuarda	0,693145931	0,715015202	0,806081448	0,948928306	0,673
25	VittorioAlfieri_Merope	0,735463235	0,783055974	0,855979157	0,971583955	0,709
26	VittorioAlfieri_Mirra	0,76329317	0,819104452	0,864045202	0,9659327	0,760
27	VittorioAlfieri_Oreste	0,70530237	0,777981376	0,829335057	0,930970217	0,715
28	VittorioAlfieri_Ottavia	0,762895099	0,791949819	0,874379901	0,96265065	0,722
29	VittorioAlfieri_Saul	0,645417404	0,735038238	0,760393582	0,871007648	0,666
30						

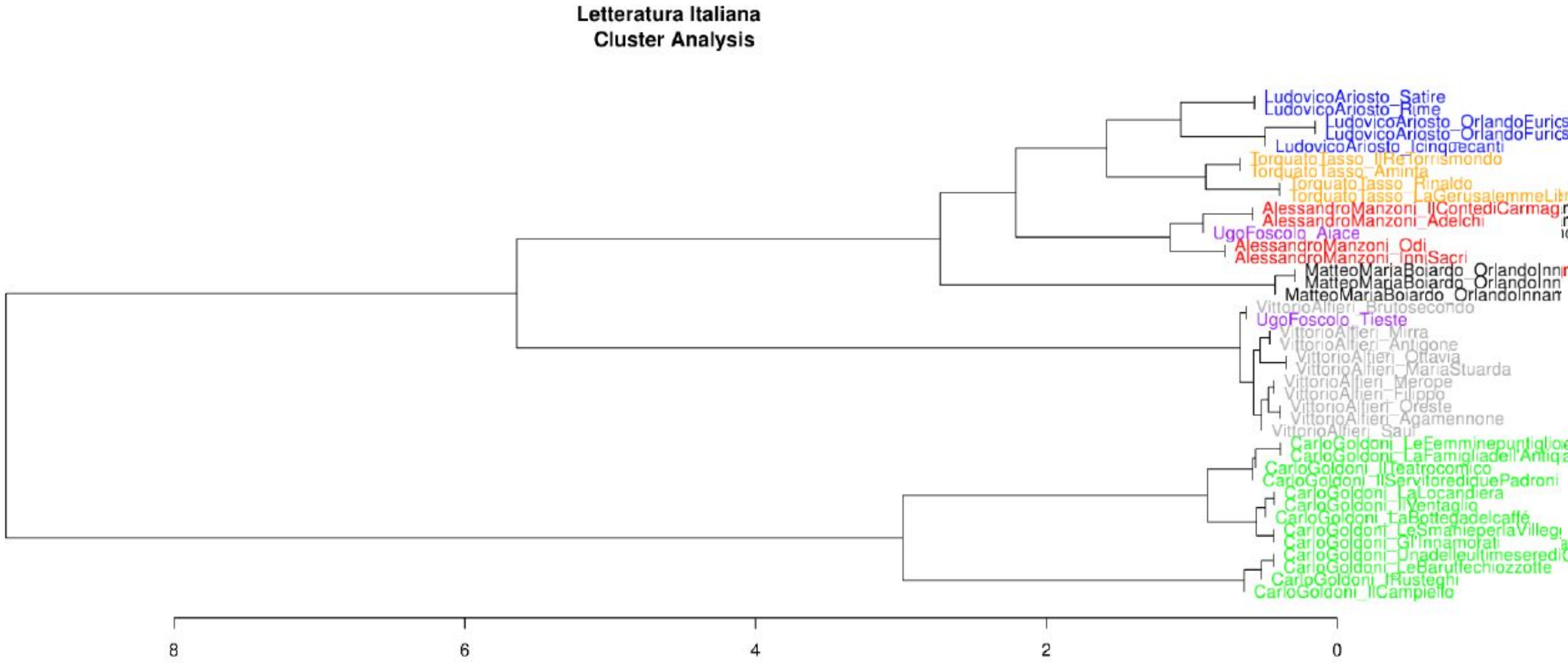
	Berlin	Brussels	Dublin	London	Madrid	Munich	Paris	Rome
Berlin	0	652	1315	930	1868	502	877	1182
Brussels	652	0	773	319	1314	602	261	1171
Dublin	1315	773	0	463	1450	1375	777	1882
London	930	319	463	0	1263	916	341	1431
Madrid	1868	1314	1450	1263	0	1485	1053	1361
Munich	502	602	1375	916	1485	0	685	698
Paris	877	261	777	341	1053	685	0	1106
Rome	1182	1171	1882	1431	1361	698	1106	0

VISUALIZATIONS

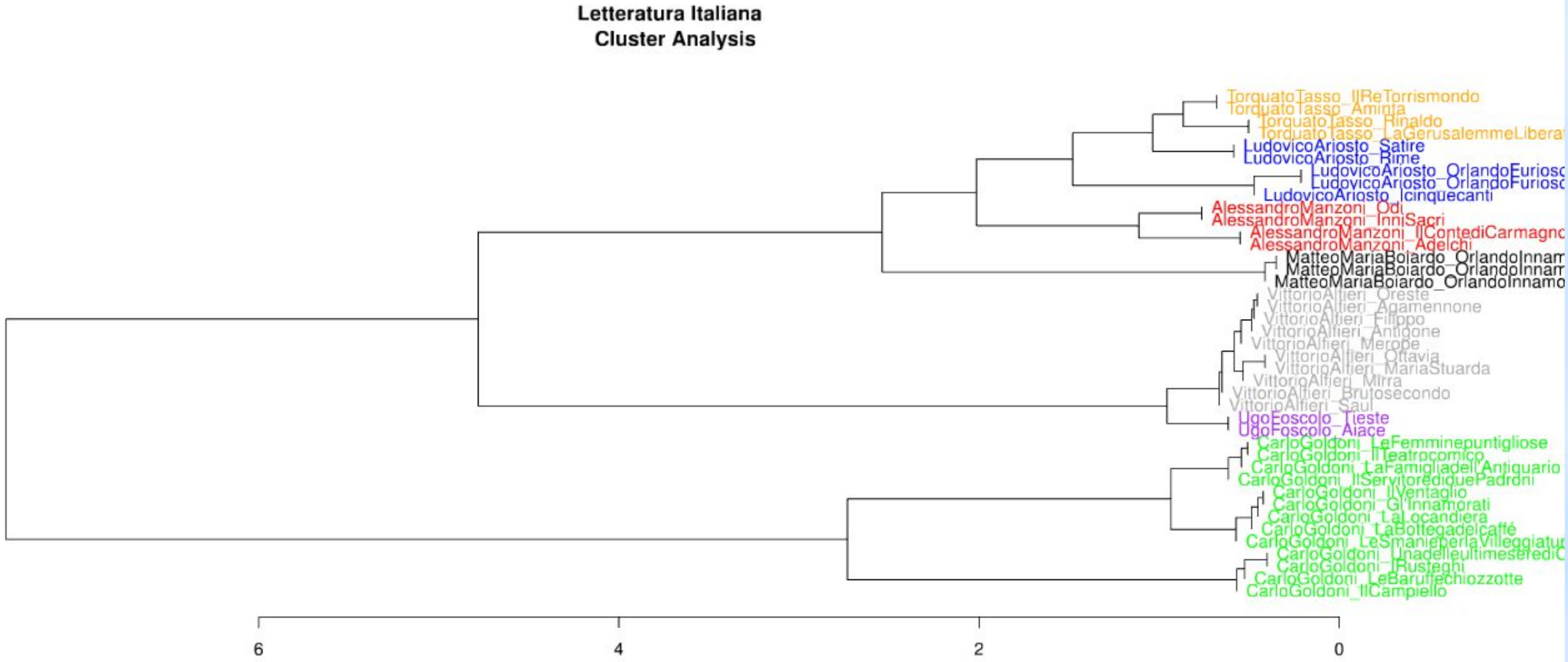
1. Dendrograms

Ward's clustering algorithm (Ward, 1963)

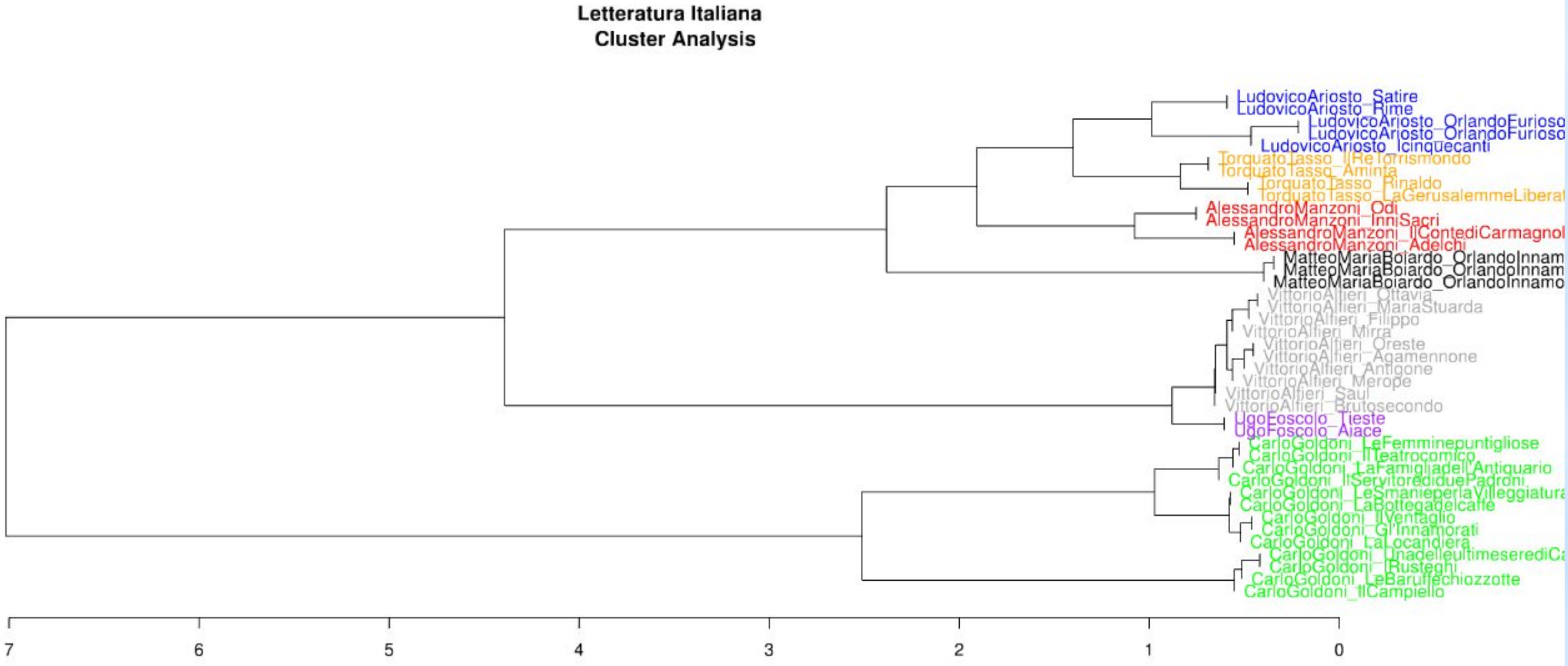

Letteratura Italiana
Cluster Analysis
LudovicoAriosto_Satire
LudovicoAriosto_Rime
LudovicoAriosto_OrlandoFurios
LudovicoAriosto_OrlandoFurios
LudovicoAriosto_Icinquecanti
TorquatoTasso_IlReTorrismondo
TorquatoTasso_Aminta
TorquatoTasso_Rinaldo
TorquatoTasso_LaGerusalemmeLiber
MatteoMariaBoiardo_OrlandoInnan
MatteoMariaBoiardo_OrlandoInnan
MatteoMariaBoiardo_OrlandoInnamo
UgoFoscolo_Tieste
UgoFoscolo_Ajace
AlessandroManzoni_IlContediCarmagr
AlessandroManzoni_Adelchi
AlessandroManzoni_Odi
AlessandroManzoni_InniSacri
VittorioAlfieri_Ottavia
VittorioAlfieri_MariaStuarda
VittorioAlfieri_Mirra
VittorioAlfieri_Brutosecondo
VittorioAlfieri_Oreste
VittorioAlfieri_Agamennone
VittorioAlfieri_Filippo
VittorioAlfieri_Merope
VittorioAlfieri_Antigone
VittorioAlfieri_Saul
CarloGoldoni_LeFemminepuntigliose
CarloGoldoni_LaFamigliadell'Antiqua
CarloGoldoni_IlTeatrocomico
CarloGoldoni_IlServitorediduePadroni
CarloGoldoni_IlVentaglio
CarloGoldoni_GliInnamorati
CarloGoldoni_LaLocandiera
CarloGoldoni_LaBottegadelcaffè
CarloGoldoni_LeSmanieperlaVilleggia
CarloGoldoni_Unadelleultimeseredi
CarloGoldoni_IBusiegni
CarloGoldoni_LeBarufechiozzotte
CarloGoldoni_IlCampiello
8
6
4
2
0
Burrows Delta
with 100 most frequent words (MFW)



Burrows Delta
with 200 most frequent words (MFW)



Cosine Delta
with 100 most frequent words (MFW)



My Weird Distance Measure
with 1,000,000 most frequent words (MFW)

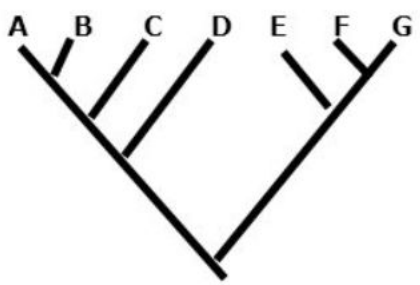
VISUALIZATIONS

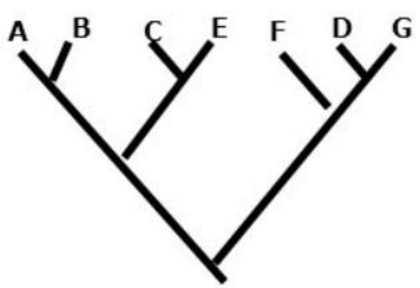
2. Consensus Trees

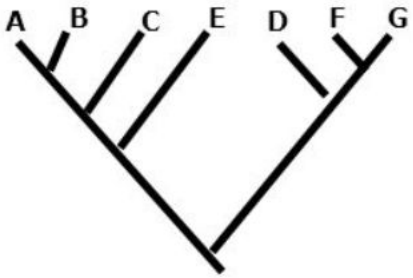
Method developed in philogenetics
(see Paradis et al. 2004)

Consensus Trees

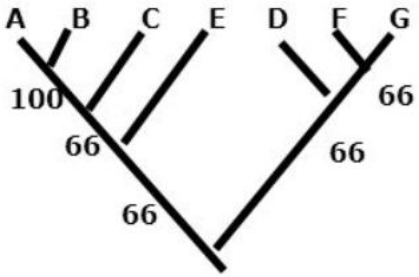
Majority rule consensus



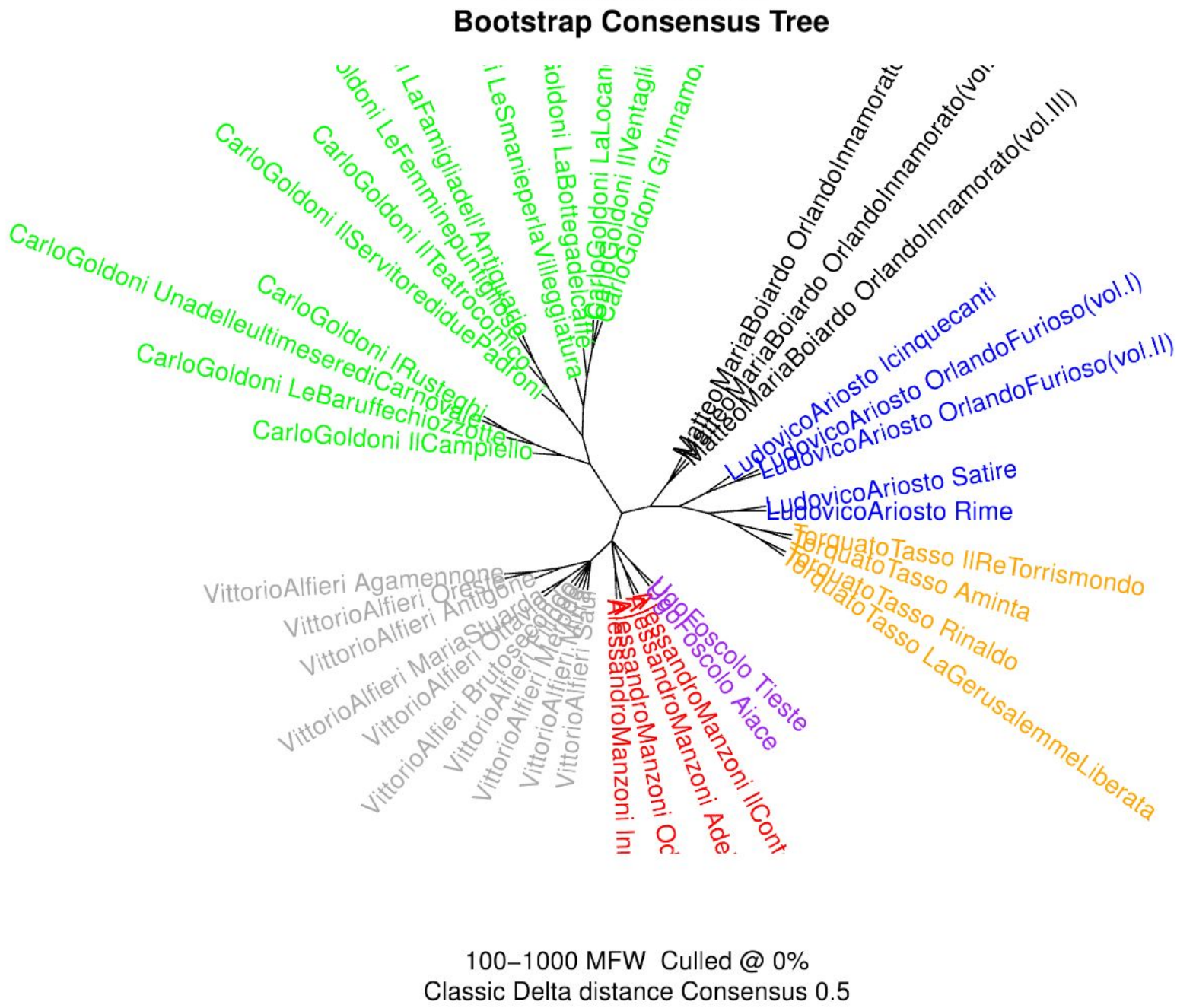




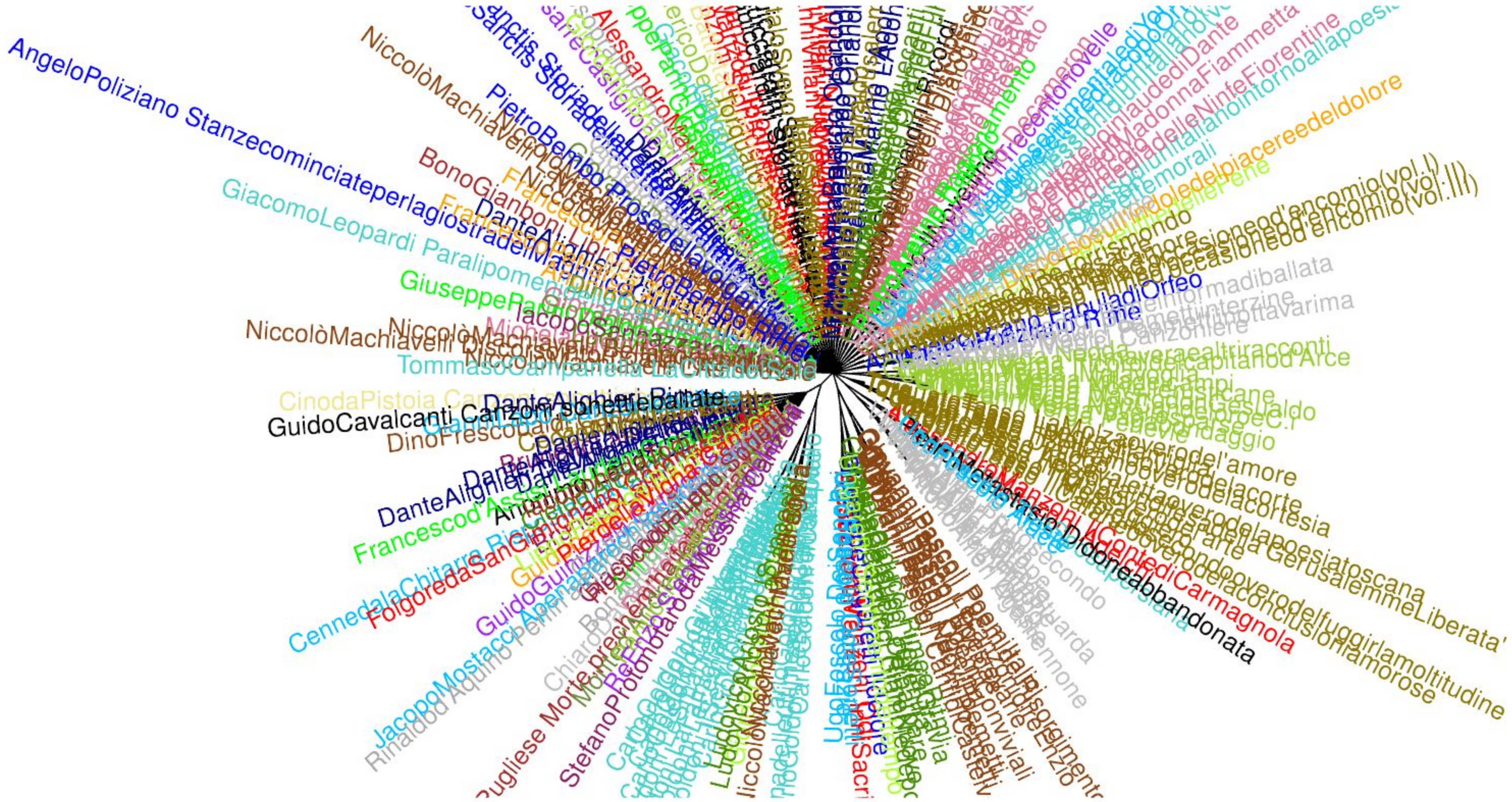
Numbers indicate frequency of clades in the fundamental trees



MAJORITY-RULE CONSENSUS TREE


Letteratura Italiana
Bootstrap Consensus Tree
CarloGoldoni LaFamigliadell'Antiquario
CarloGoldoni LeFemminepuntigliose
CarloGoldoni IlTeatrocomico
CarloGoldoni LeSmanieperlaVilleggiatura
CarloGoldoni LaBottegadelCaffe
CarloGoldoni IlServitoreDiDuePadroni
CarloGoldoni IRusteghi
CarloGoldoni UnadelleultimeserediCarnovale
CarloGoldoni LeBaruffechiozzotte
CarloGoldoni IlCampiello
MatteoMariaBoiardo OrlandoInnamorato
MatteoMariaBoiardo OrlandoInnamorato(vol.II)
MatteoMariaBoiardo OrlandoInnamorato(vol.III)
LudovicoAriosto Icinquecanti
LudovicoAriosto OrlandoFurioso(vol.I)
LudovicoAriosto OrlandoFurioso(vol.II)
LudovicoAriosto Satire
LudovicoAriosto Rime
TorquatoTasso IlReTorrismondo
TorquatoTasso Aminta
TorquatoTasso Rinaldo
TorquatoTasso LaGerusalemmeLiberata
VittorioAlfieri Agamennone
VittorioAlfieri Oreste
VittorioAlfieri Antigone
VittorioAlfieri MariaStuarda
VittorioAlfieri Filippo
VittorioAlfieri Saul
UgoFoscolo Tieste
UgoFoscolo Aiace
AlessandroManzoni IlConte
AlessandroManzoni Adelchi
AlessandroManzoni Oc
AlessandroManzoni In
100−1000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

Letteratura Italiana
Bootstrap Consensus Tree



100–1000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

VISUALIZATIONS

3. Network Graphs

See Eder, 2017

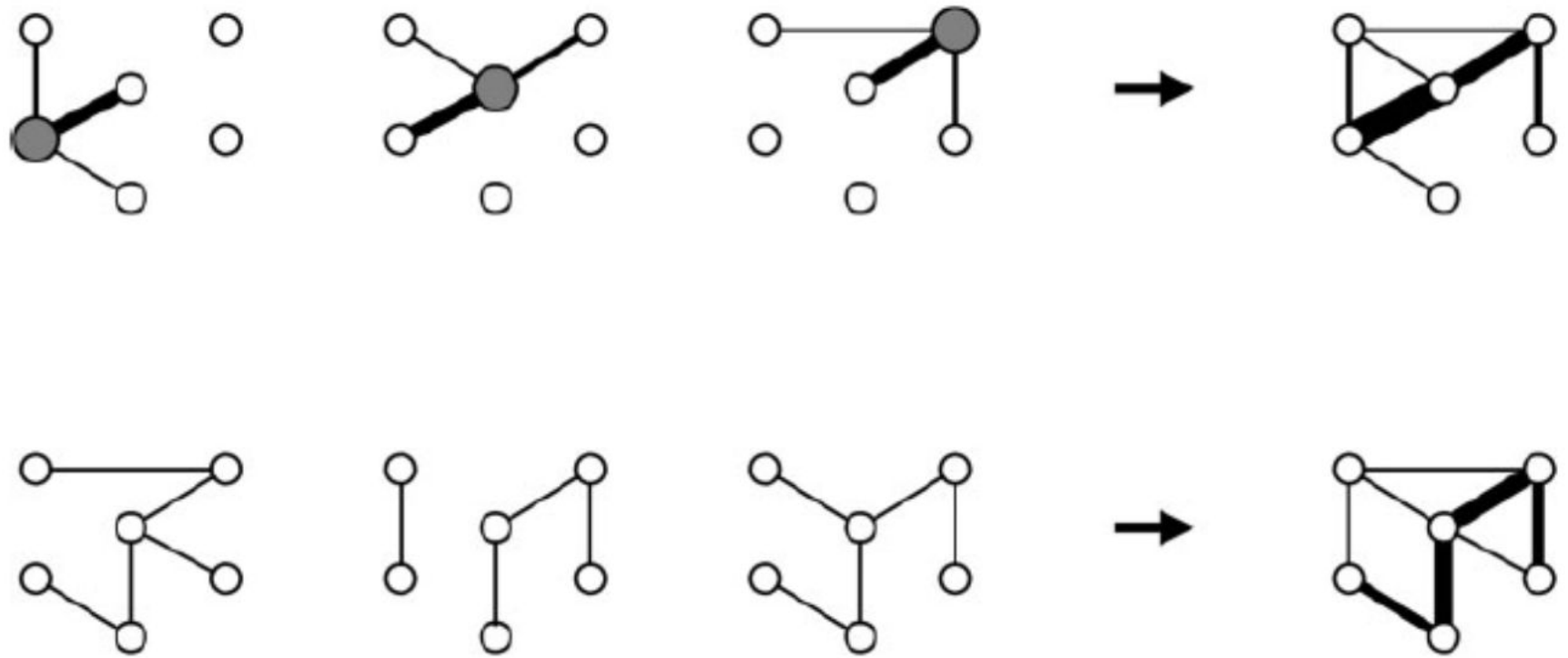
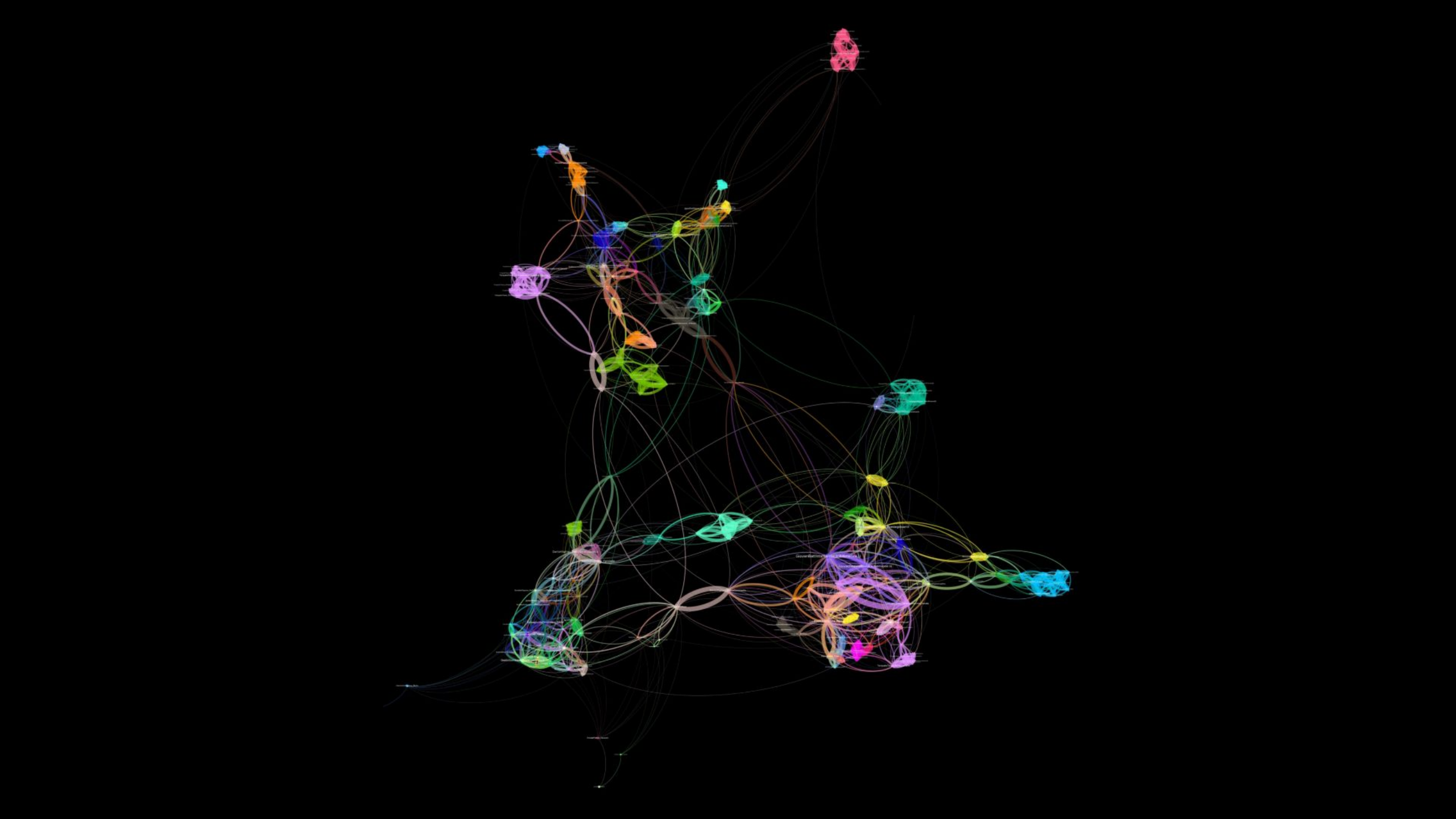
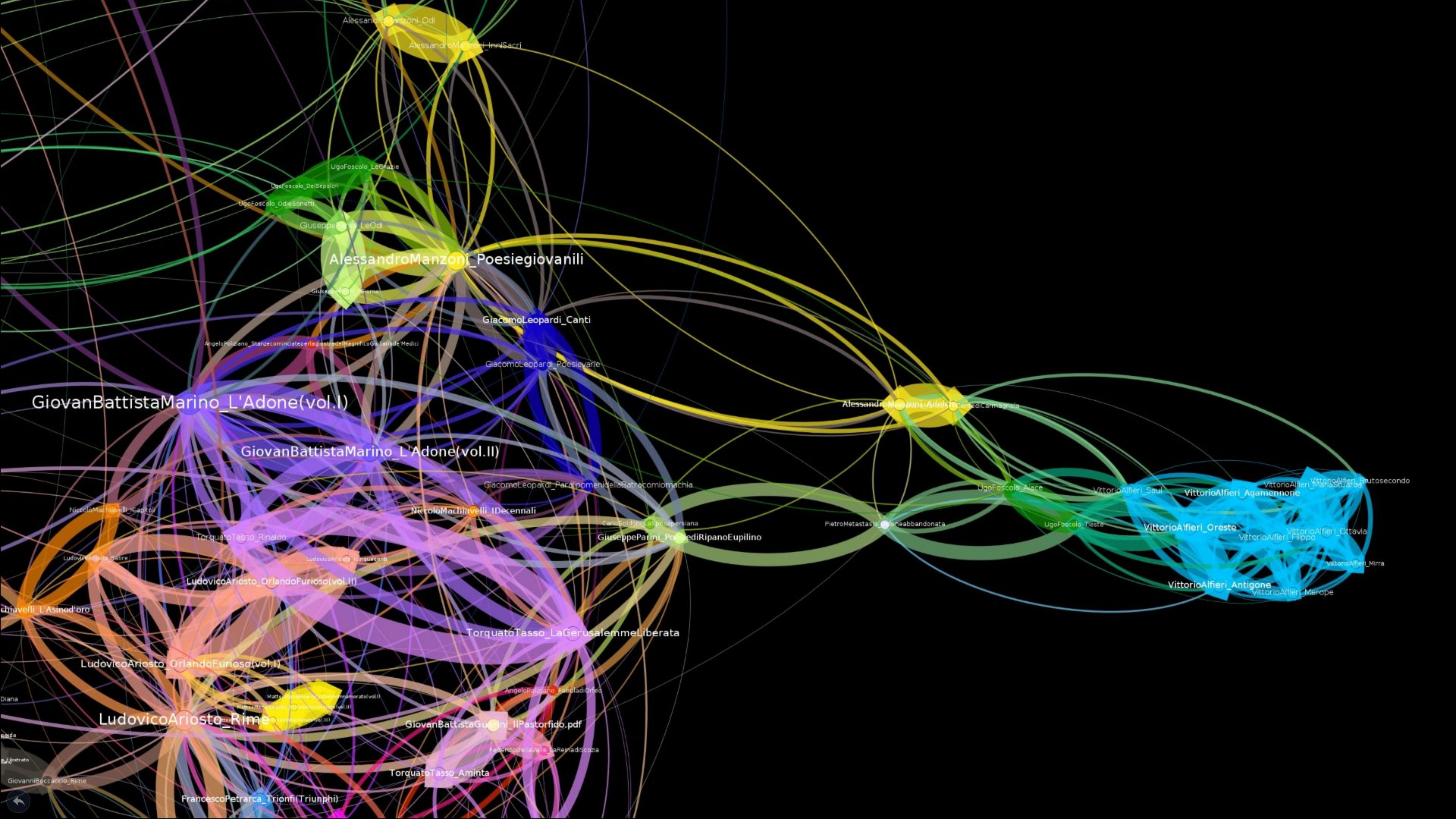
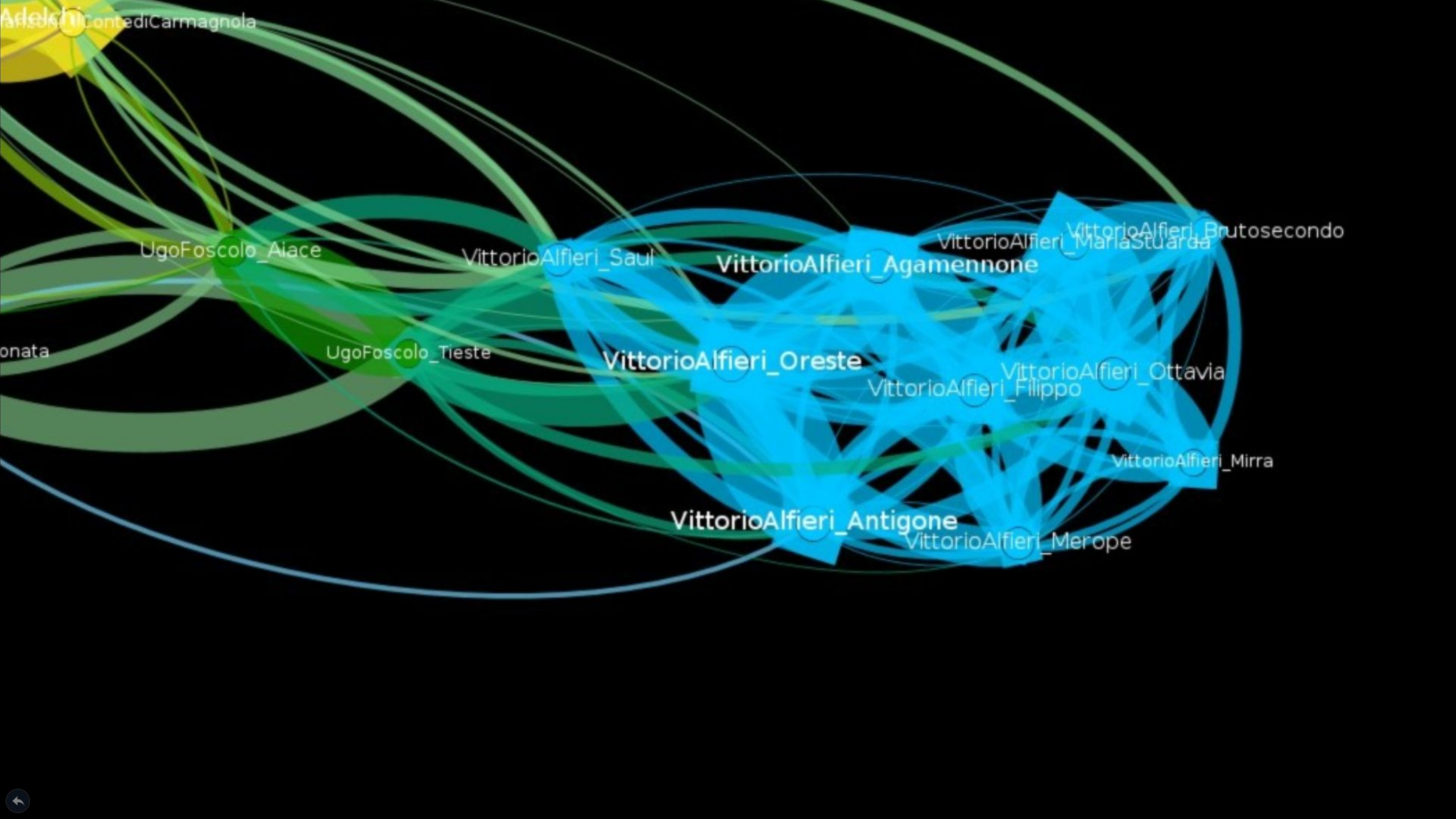


Fig. 6. Two algorithms of mapping textual relations: establishing weighted links to a nearest neighbor and two runners-up (top); producing a consensus network (bottom).

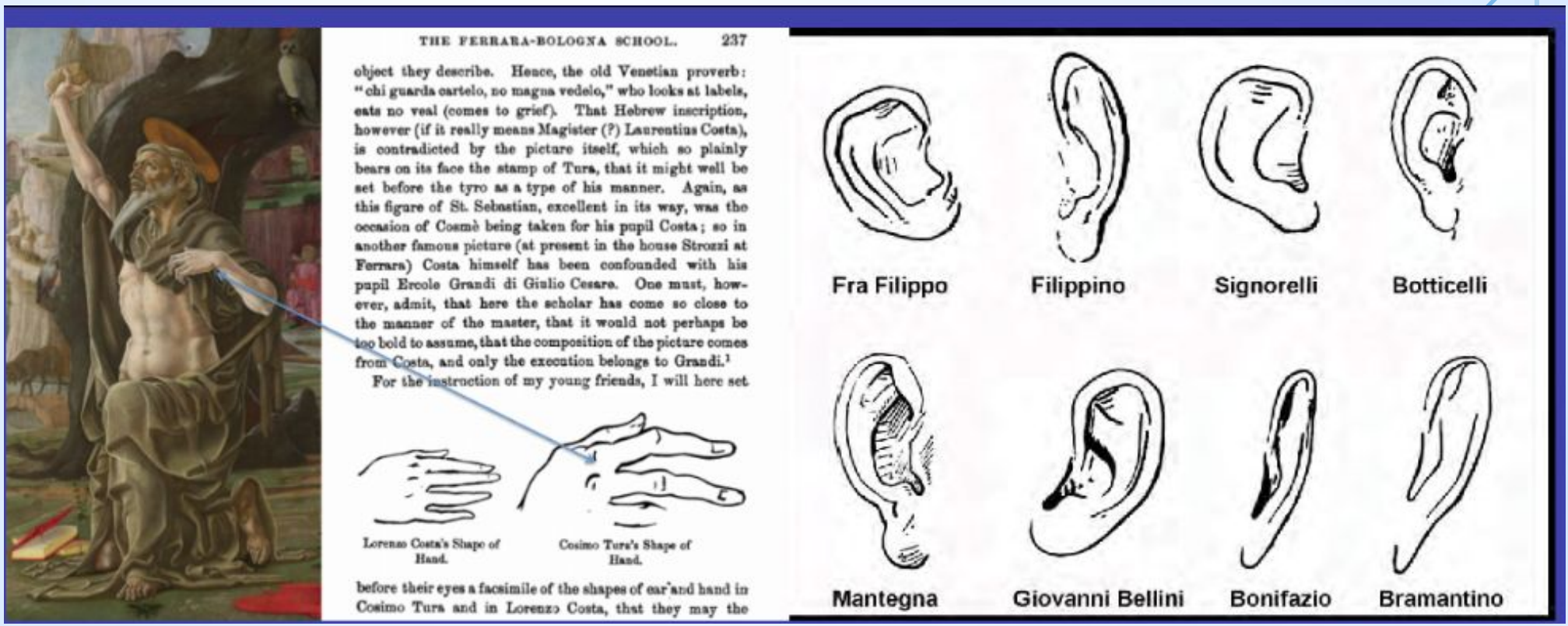




UgoFoscolo_LeGrazie
UgoFoscolo_DeiSepolcri
UgoFoscolo_OdieSonetti
AlessandroManzoni_Poesiegiovanili
GiacomoLeopardi_Canti
GiacomoLeopardi_Poesievarie
GiovanBattistaMarino_L'Adone(vol.I)
GiovanBattistaMarino_L'Adone(vol.II)
GiacomoLeopardi_ParalipomenidellaBatracomiomachia
NiccoloMachiavelli_ICapitoli
NiccoloMachiavelli_IDecennali
TorquatoTasso_Rinaldo
LudovicoAriosto_Satire
LudovicoAriosto_OrlandoFurioso(vol.II)
TorquatoTasso_LaGerusalemmeLiberata
LudovicoAriosto_OrlandoFurioso(vol.I)
LudovicoAriosto_Rime
GiovanBattistaGuarini_IlPastorfido.pdf
FedericoDellaValle_LaReinadiScozia
TorquatoTasso_Aminta
GiovanniBoccaccio_Rime
FrancescoPetrarca_Trionfi(Triunphi)
UgoFoscolo_Aiace
UgoFoscolo_Tieste
VittorioAlfieri_Saul
VittorioAlfieri_Agamennone
VittorioAlfieri_Brutosecondo
VittorioAlfieri_Oreste
VittorioAlfieri_Filippo
VittorioAlfieri_Ottavia
VittorioAlfieri_Mirra
VittorioAlfieri_Antigone
VittorioAlfieri_Merope

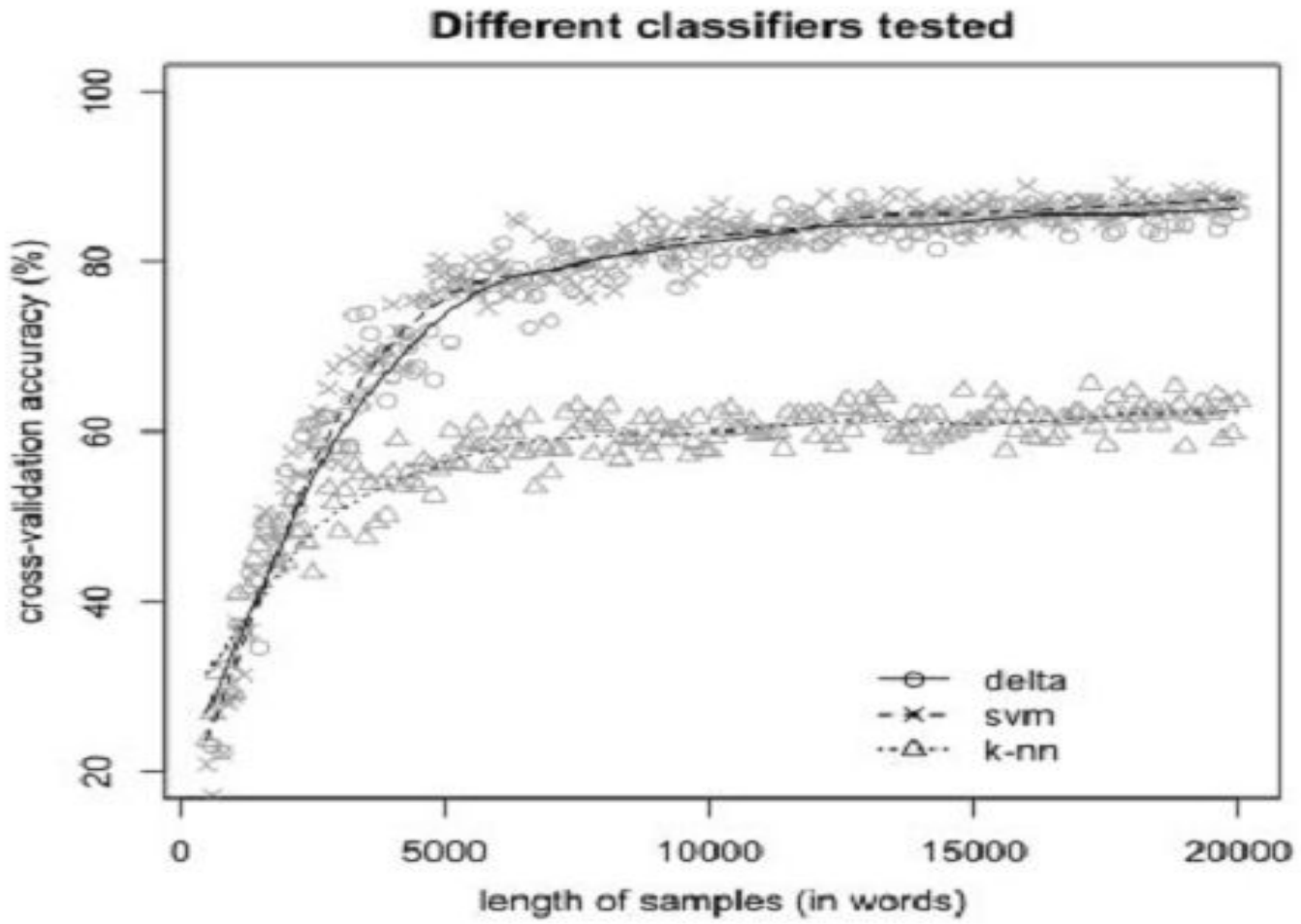

ContediCarmagnola
UgoFoscolo_Aiace
VittorioAlfieri_Saul
VittorioAlfieri_Agamennone
VittorioAlfieri_MariaStuarda
VittorioAlfieri_Brutosecondo
UgoFoscolo_Tieste
VittorioAlfieri_Oreste
VittorioAlfieri_Ottavia
VittorioAlfieri_Filippo
VittorioAlfieri_Mirra
VittorioAlfieri_Antigone
VittorioAlfieri_Merope

WHY DOES IT WORK?



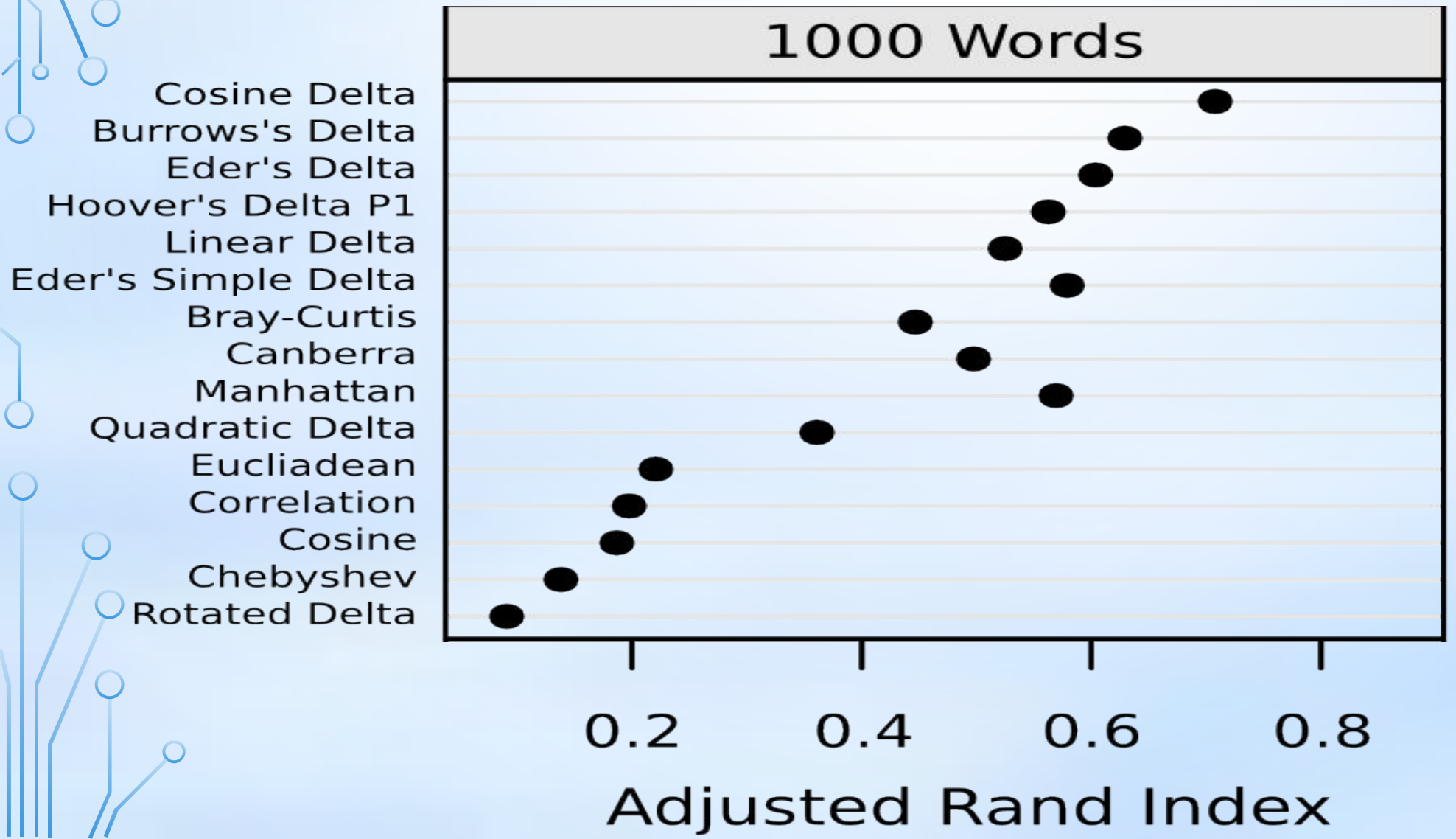
"It has been noted that the switch from content words to function words in authorship attribution studies has an interesting historic parallel in art-historic research. […] Giovanni Morelli (1816-1891) was among the first to suggest that the attribution of, for instance, a Quattrocento painting to some Italian master, could not happen based on 'content' […] Morelli thought it better to restrict an authorship analysis to discrete details such as ears, hands and feet" (Kestemont 2014)

CORPUS SELECTION (TEXT DIMENSIONS)



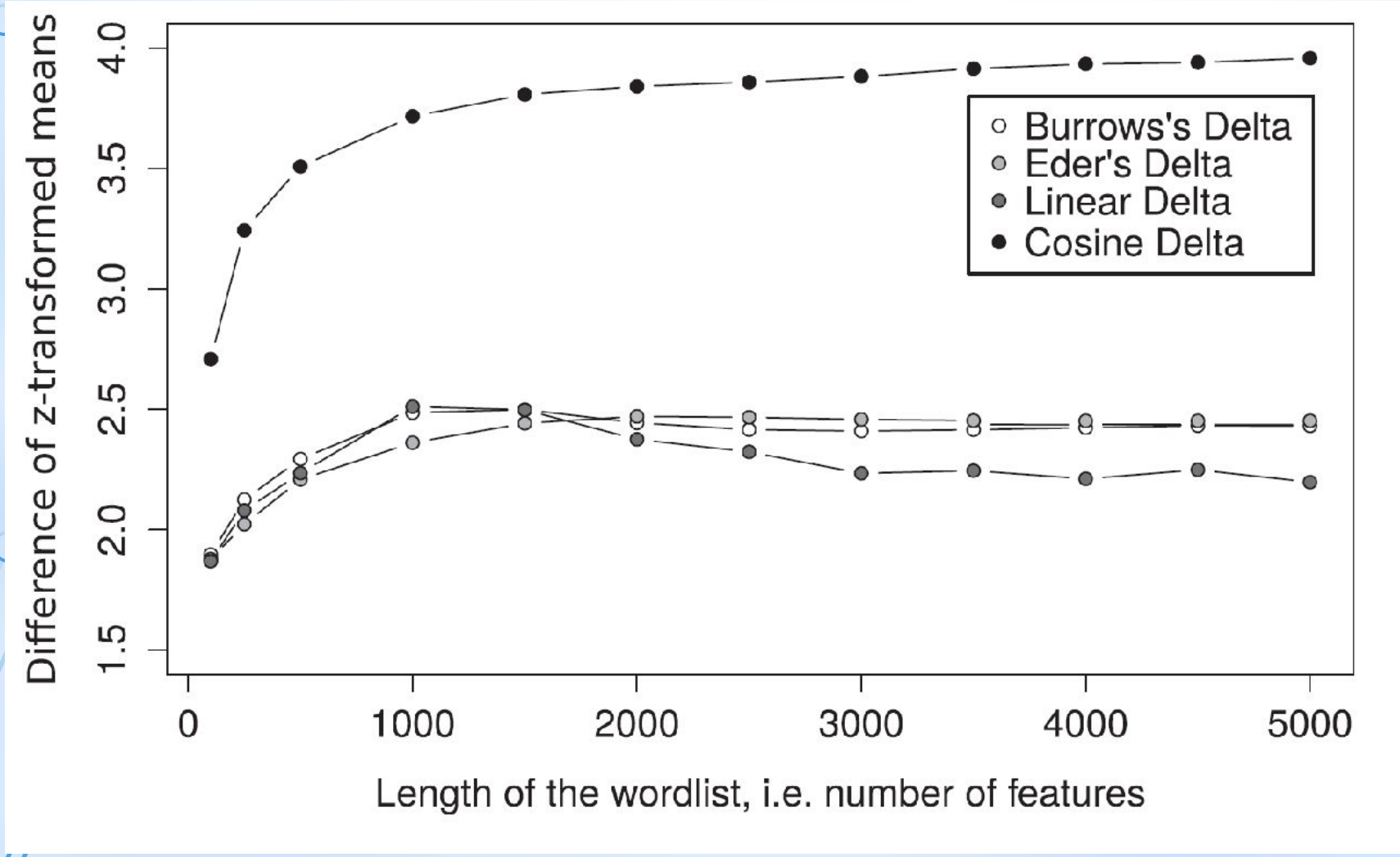
Minumun text length for a reliable stylometric analysis is about 5,000 words (Eder 2015)

FEATURES SELECTION (DISTANCE MEASURE)



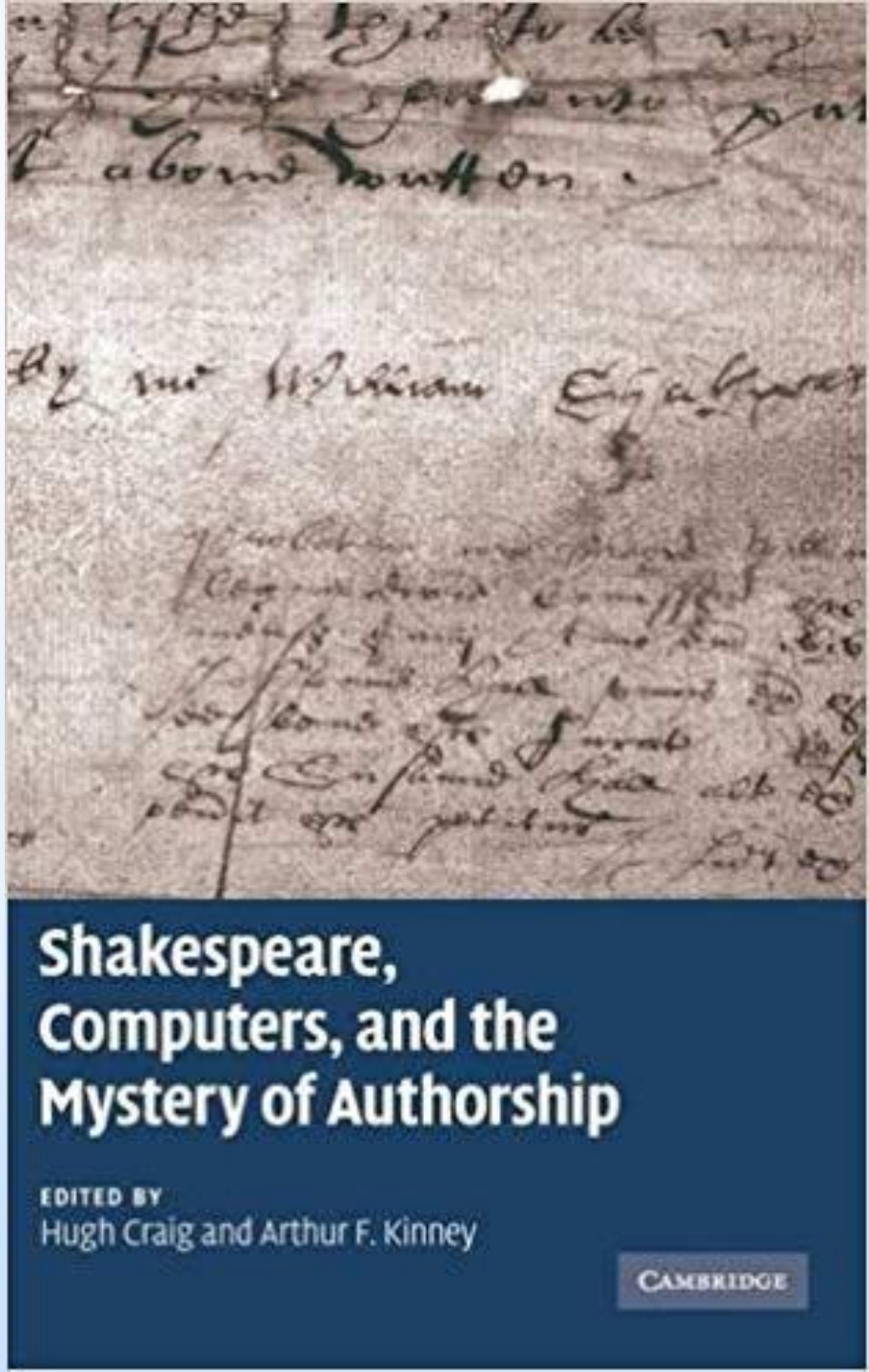
Cosine Delta is the best performing distance (Evert et al. 2017)

FEATURES SELECTION (MFW)



About
1,000-2,000
MFW produce the
best results
(Evert et al. 2017)

THE «ZETA» METHOD



WHAT MAKES THESE THREE SENTENCES «SHAKESPEAREAN»?

I meant indeed to pay you with this, which if like an ill venture it come unluckily home, I break, and you, my gentle creditors, lose. Here I promised you I would be, and here I commit my body to your mercies. Bate me some, and I will pay you some, and (as most debtors do) promise you infinitely.

2 Henry IV

But since you have made the days and nights as one,
To wear your gentle limbs in my affairs,
Be bold you do so grow in my requital
As nothing can unroot you.

All's Well that Ends Well

This is a sleepy tune. O murd'rous slumber!
Layest thou thy leaden mace upon my boy,
That plays thee music? Gentle knave, good night;
I will not do thee so much wrong to wake thee.

Julius Caesar

WHAT MAKES THESE THREE SENTENCES «SHAKESPEAREAN»?

I meant indeed to pay you with this, which if like an ill venture it come unluckily home, I break, and you, my gentle creditors, lose. Here I promised you I would be, and here I commit my body to your mercies. Bate me some, and I will pay you some, and (as most debtors do) promise you infinitely.

2 Henry IV

But since you have made the days and nights as one,
To wear your gentle limbs in my affairs,
Be bold you do so grow in my requital
As nothing can unroot you.

All's Well that Ends Well

This is a sleepy tune. O murd'rous slumber!
Layest thou thy leaden mace upon my boy,
That plays thee music? Gentle knave, good night;
I will not do thee so much wrong to wake thee.

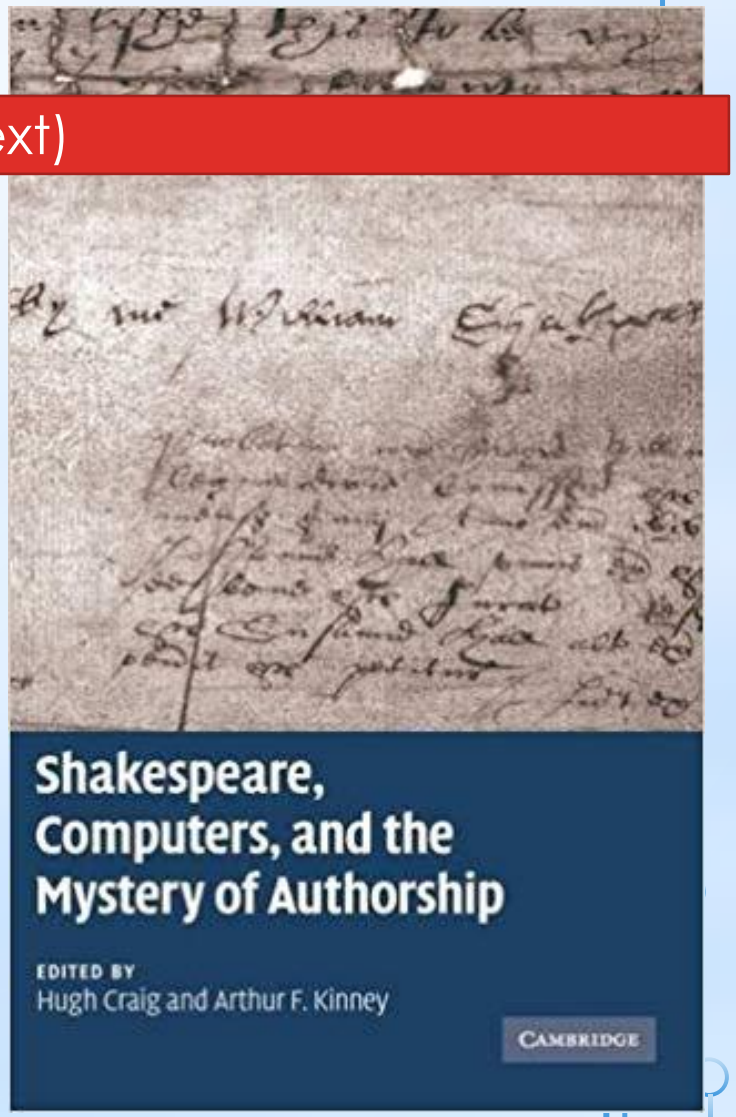
Julius Caesar

STYLOMETRY AND SHAKESPEARE

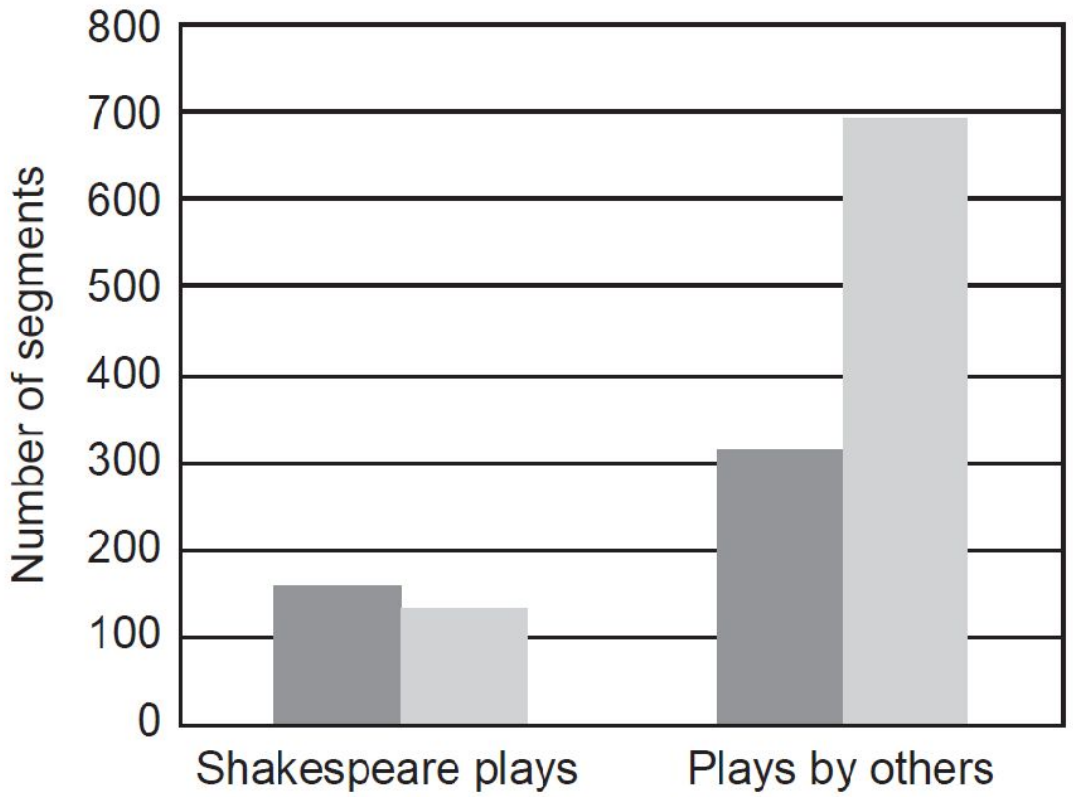
Shakespeare's works (as a single string of text)

3,000 words
3,000 words
3,000 words
etc.

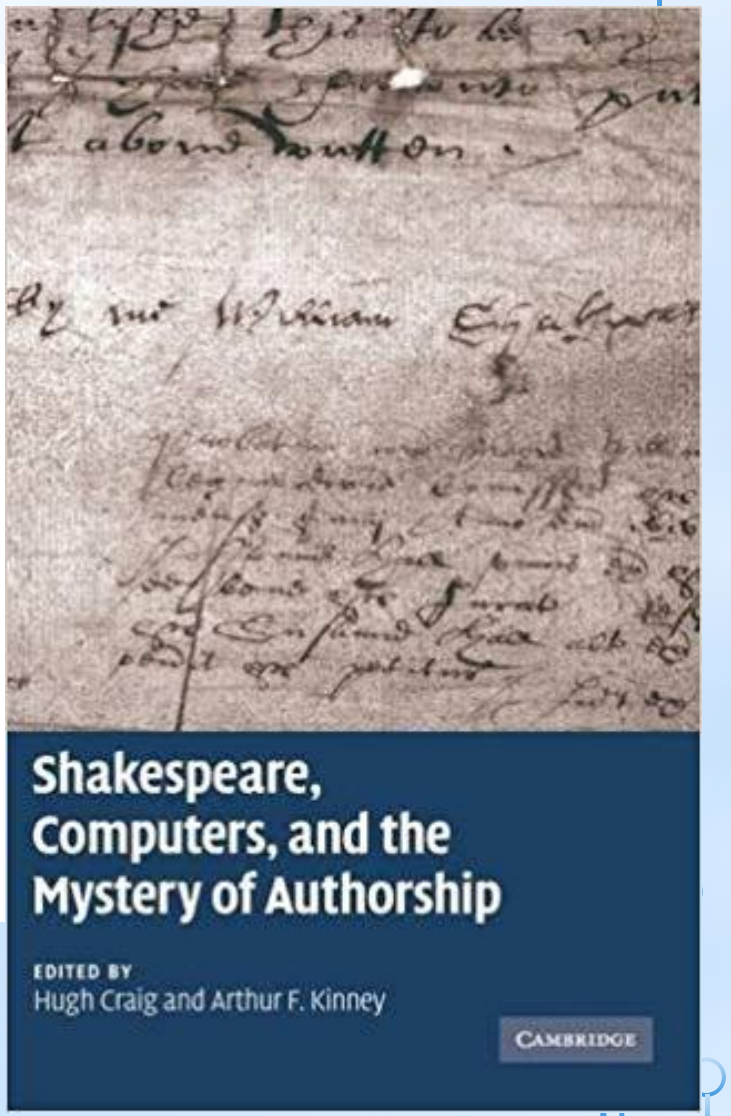
How many of the slices contain the word «Gentle»?



STYLOMETRY AND SHAKESPEARE

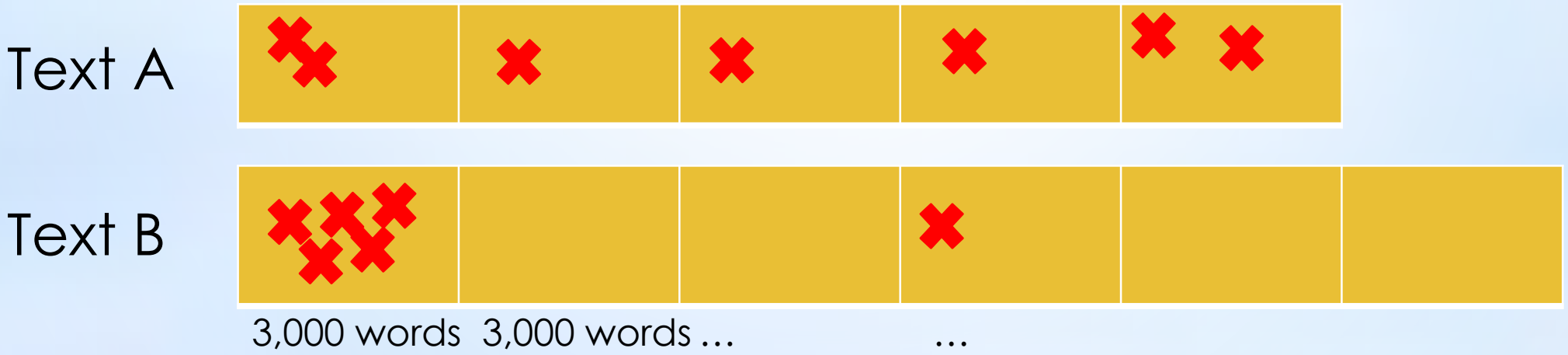


- Segments in which ‘gentle’ appears
- Segments in which ‘gentle’ does not appear



THE "ZETA" METHOD

Pick up a word:
«gentle» (for example)

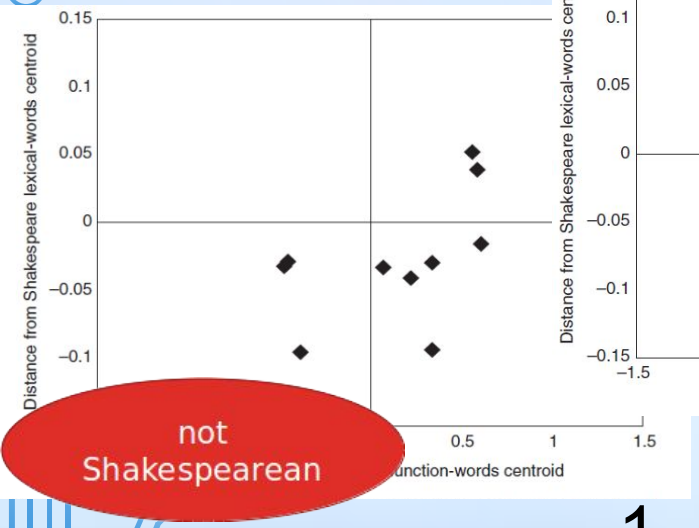


- Count in how many slices of the text the word «gentle» appears
- Calculate the proportion
  Text A: 1 (100%); text B: 0.33 (33%)
- Subtract the two values
  (so the word «gentle» has Zeta = 0.66 for Text A)
- Repeat the operation for all the words in the two texts

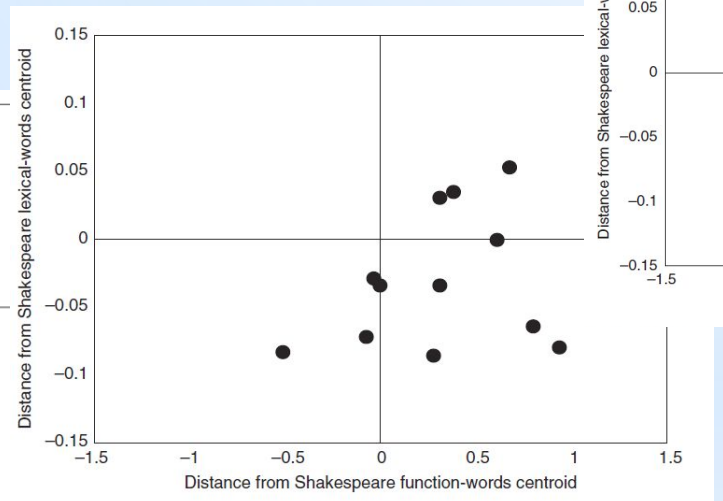
Score
-1.0
-0.5
0.0
0.5
1.0
0
10
20
e
fate
only
hopes
than
amongst
perhaps
court
enjoy
reward
yes
h
lets
sure
its
because
country
expect
wealth
somewhat
above
lose
fit
joys
wait
gentle
ore
bid
st
morrow
beseech
ouer
lord
yea
pluck
note
duke
between
toward
hoa
therefore
bloody
spoke
heavy
borne
bear
wherefore
answer
mark
comes
Shakes

APPLICATIONS IN LITERARY STUDIES

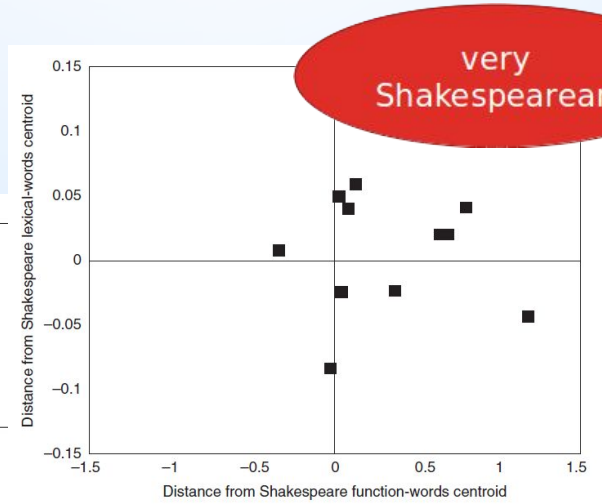
Stylometry and the three parts of *Henry VI* (Craig and Kinney 2009)



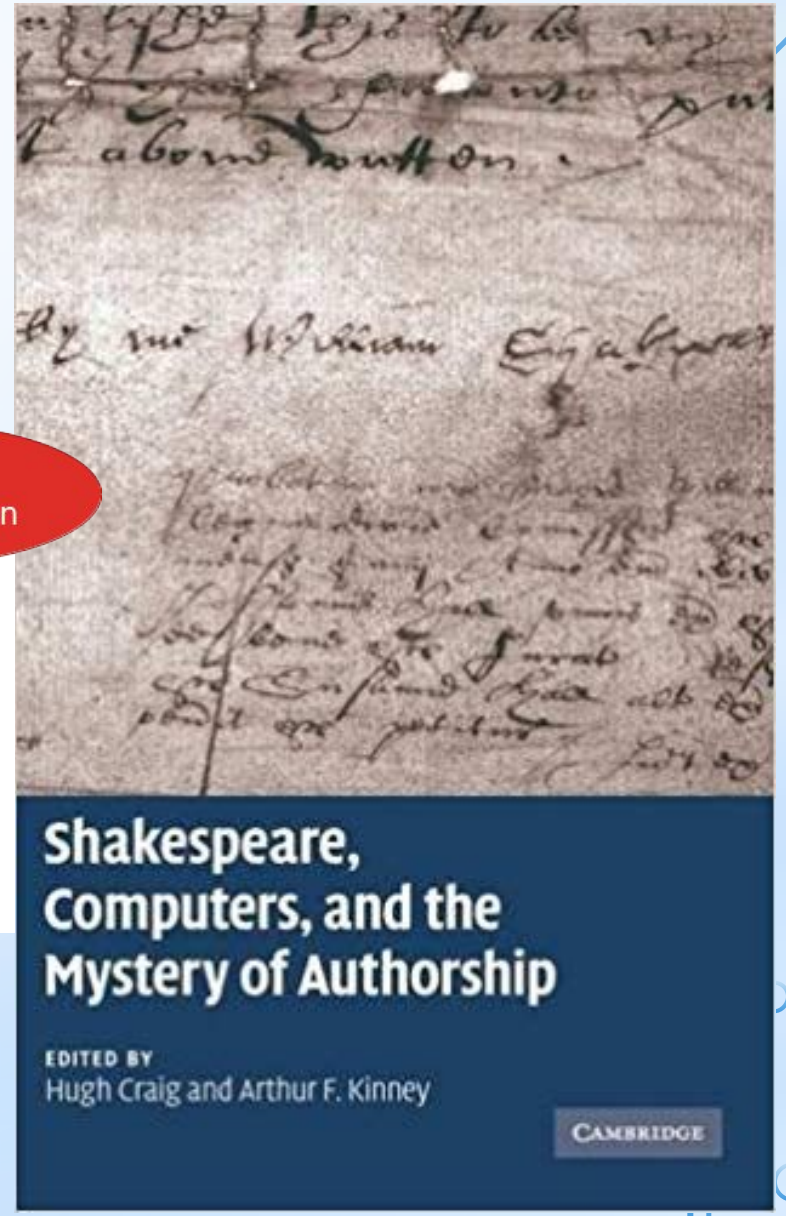
1



2

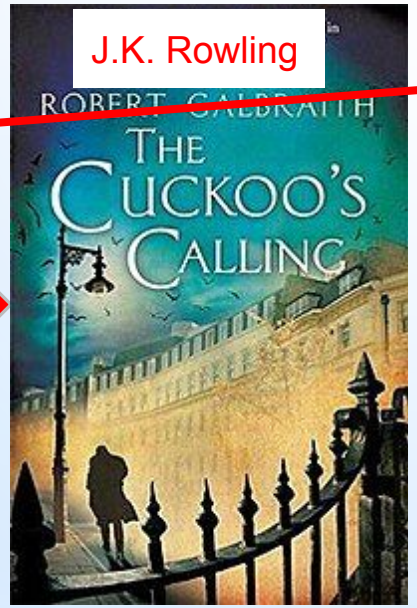


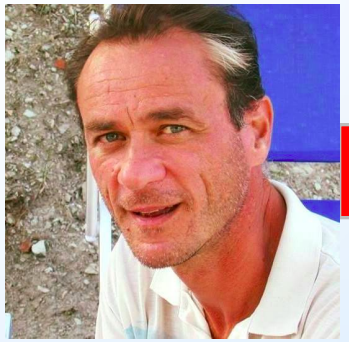
3



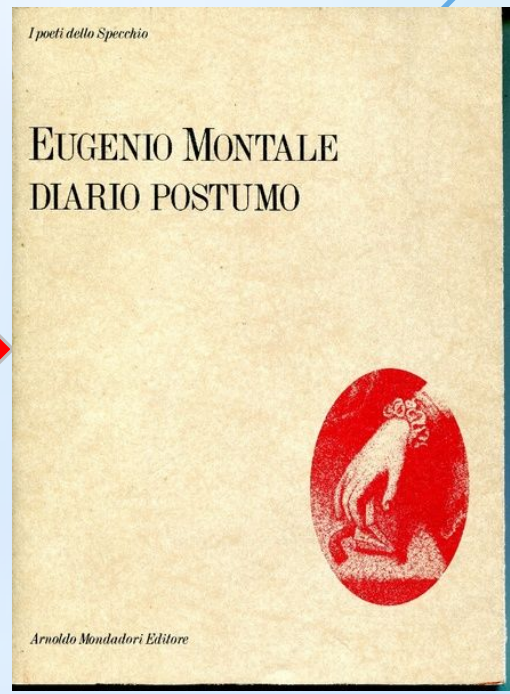
APPLICATIONS IN LITERARY STUDIES

(Patrick Juola)

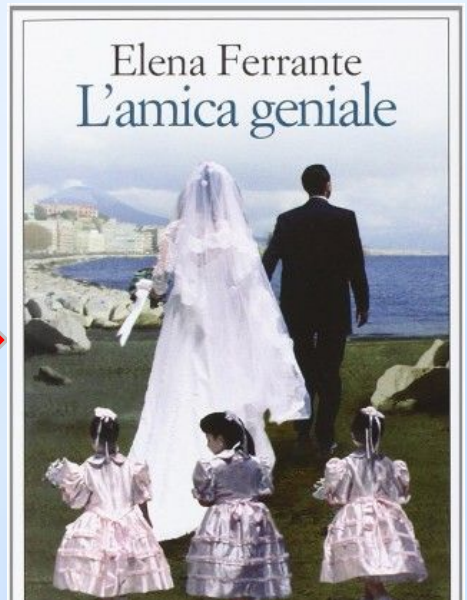




(Paolo Canettieri)

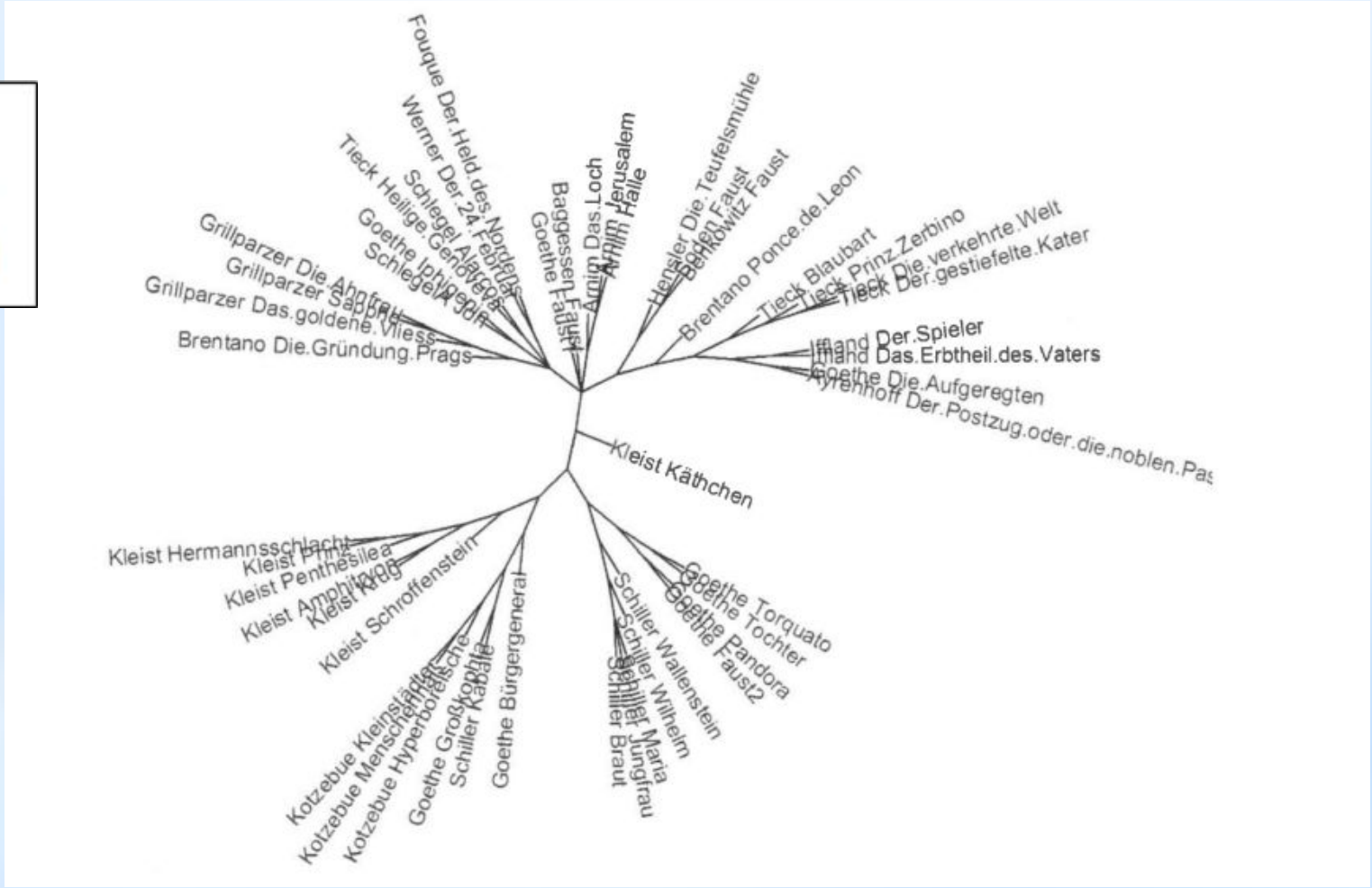


(Arjuna Tuzzi)



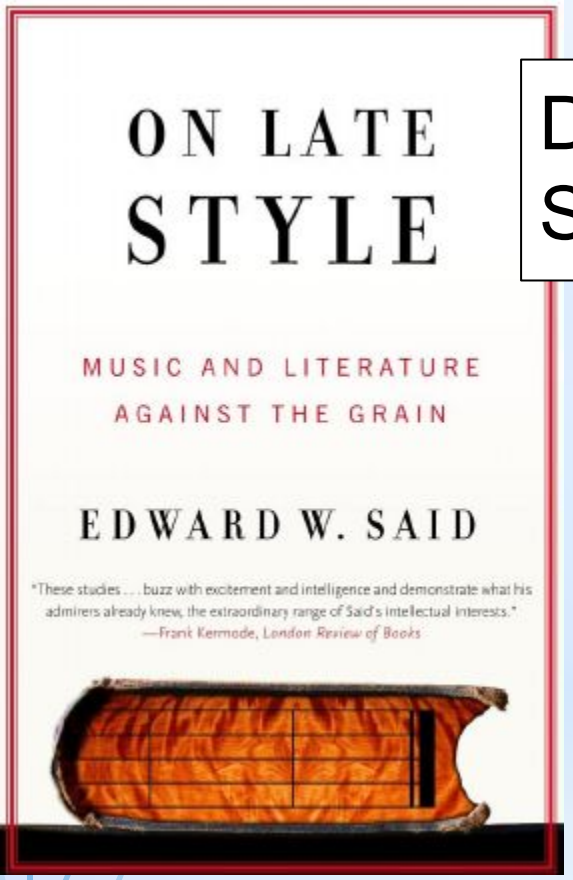
APPLICATIONS IN LITERARY STUDIES

Is Kleist a classicist or a romantic?



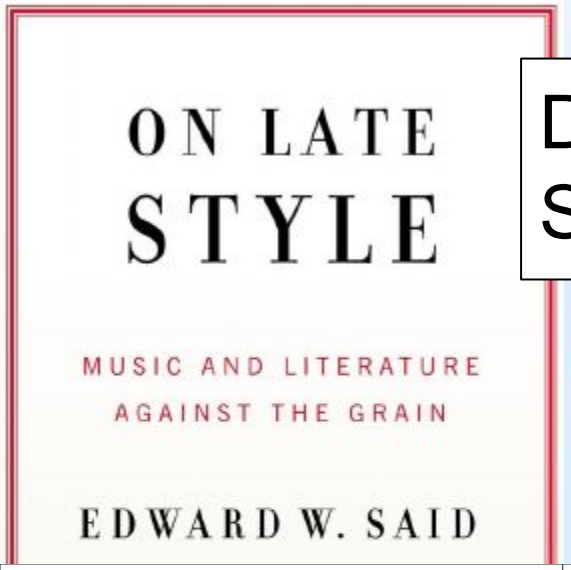
(Jannidis and Lauer, 2014)

APPLICATIONS IN LITERARY STUDIES



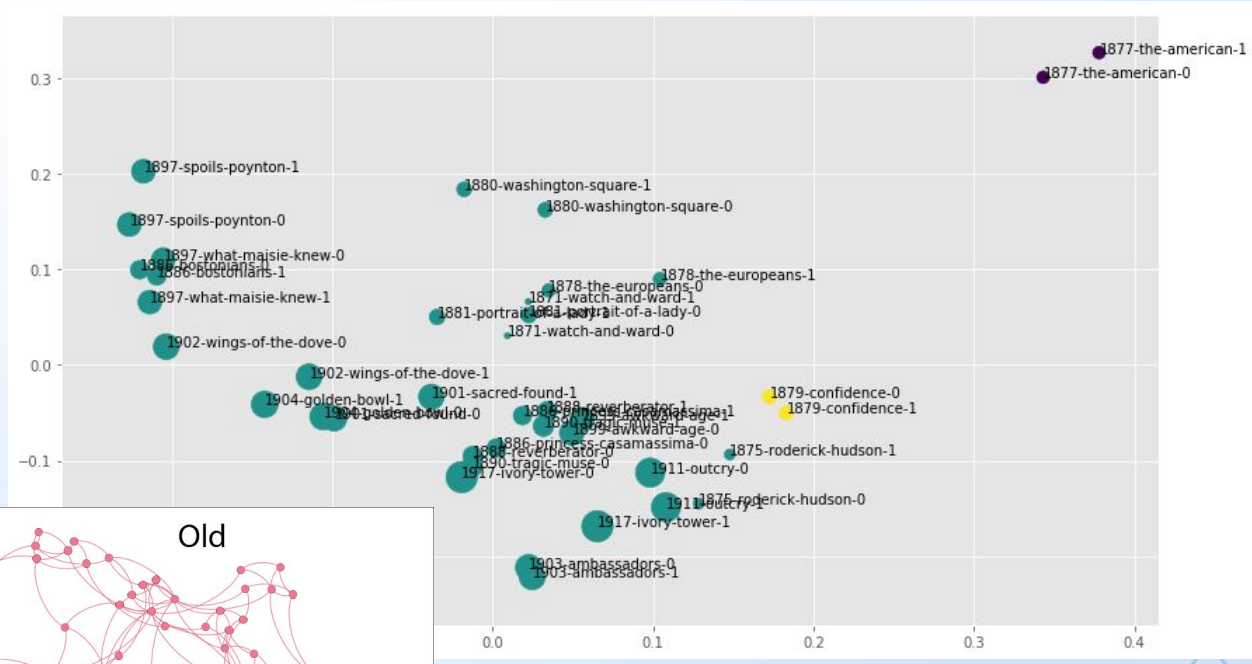
Does Late Style Exist?

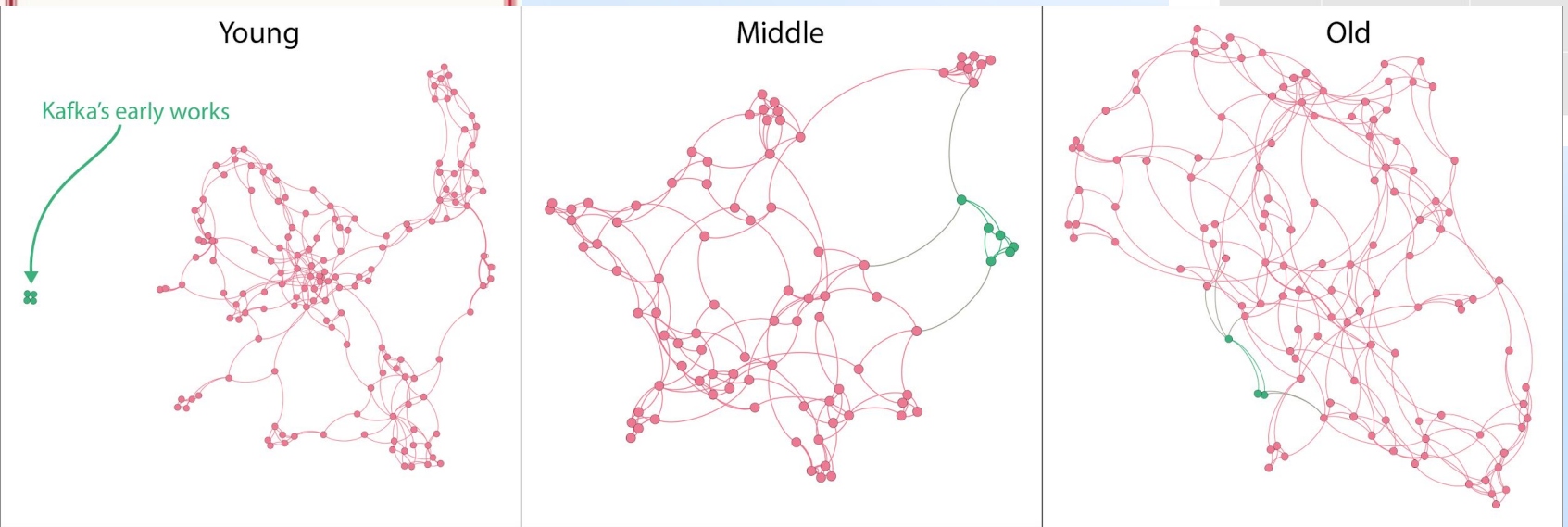
APPLICATIONS IN LITERARY STUDIES



Does Late Style Exist?

(Reeve, 2018)





(Rebora and Salgaro, 2018)

HANDS-ON!!!