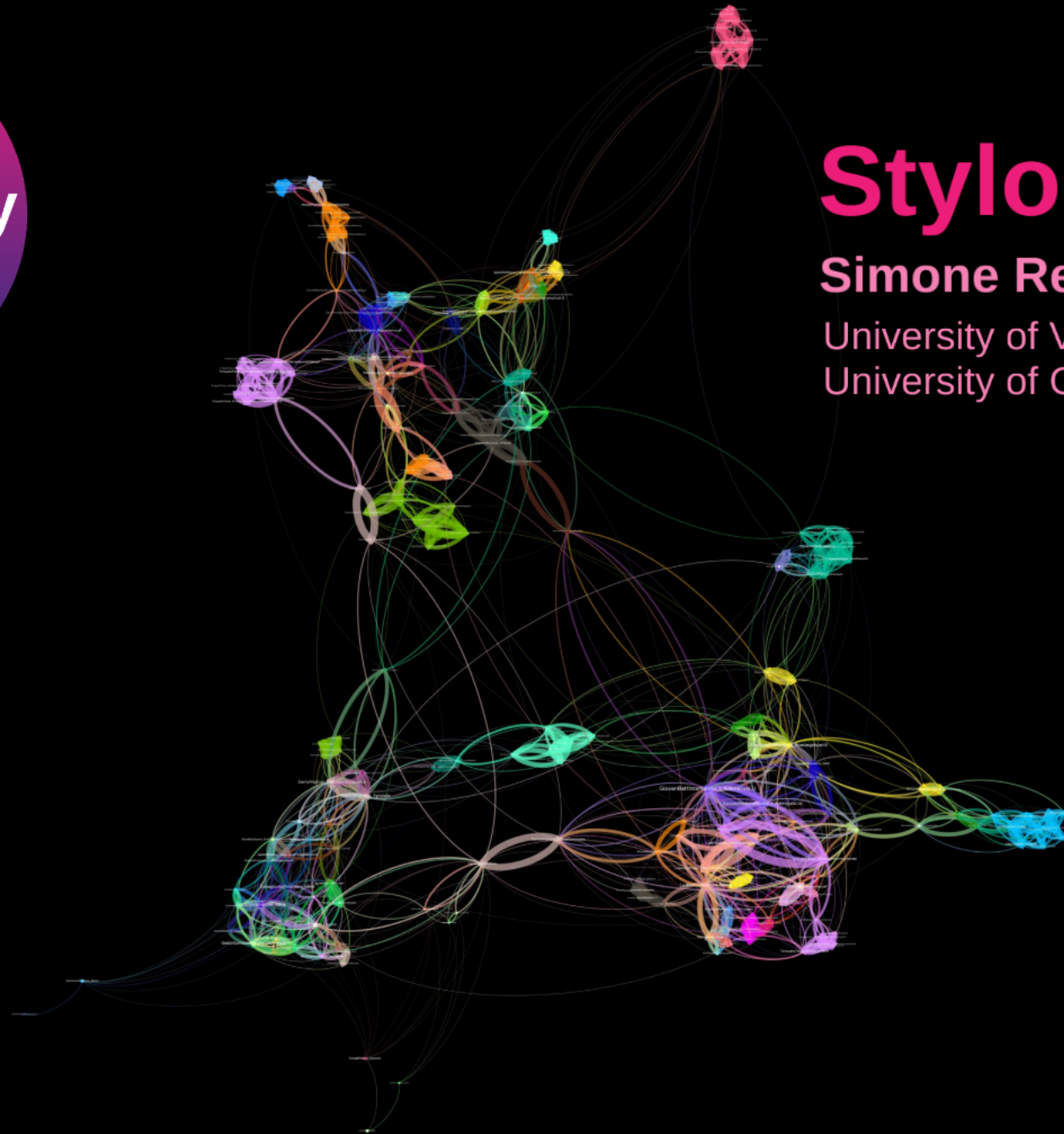


Stylometry

**Authorship
Attribution**

**Network
Analysis**



Stylometry

Simone Rebora

University of Verona

University of Göttingen

Stylometry

=

"measuring
style"

The
Origins

The
Revolution

The Origins

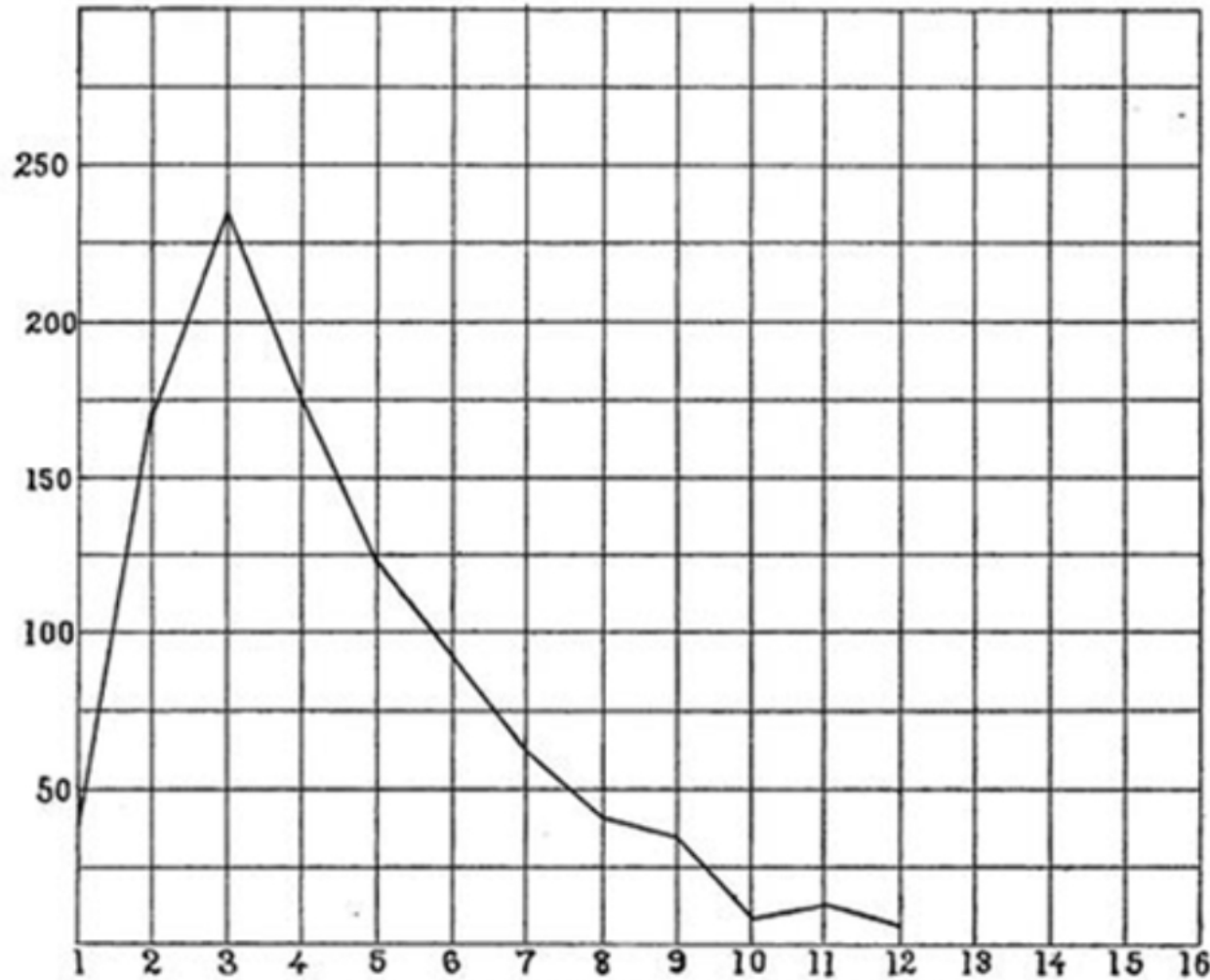


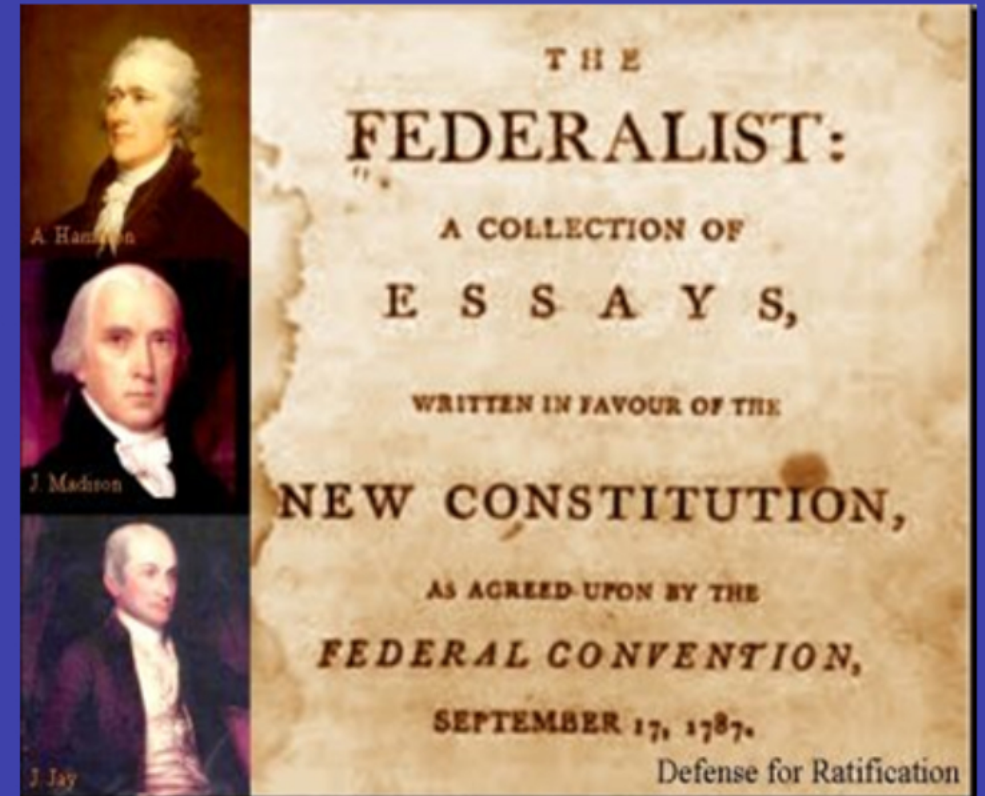
FIG. 1. — FIRST ONE THOUSAND WORDS IN 'OLIVER TWIST.'

Mendenhall, T. C.
(1887). "The
Characteristic Curves of
Composition". *Science*.
IX (214): 237–248

A history of successes...

- 3 authors (A. Hamilton, J. Madison, J. Jay)
- 85 articles and essays written in 1787-1788, under the pseudonym "Publius"
- frequency of 165 words (mainly functional)

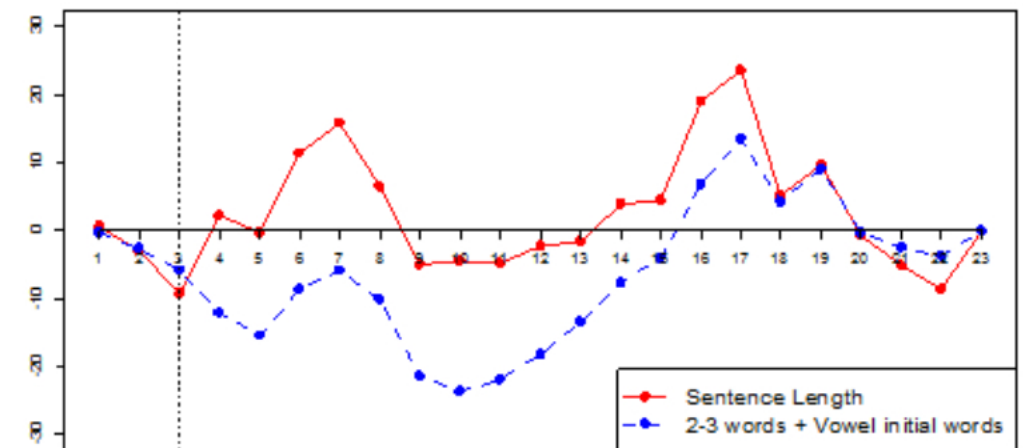
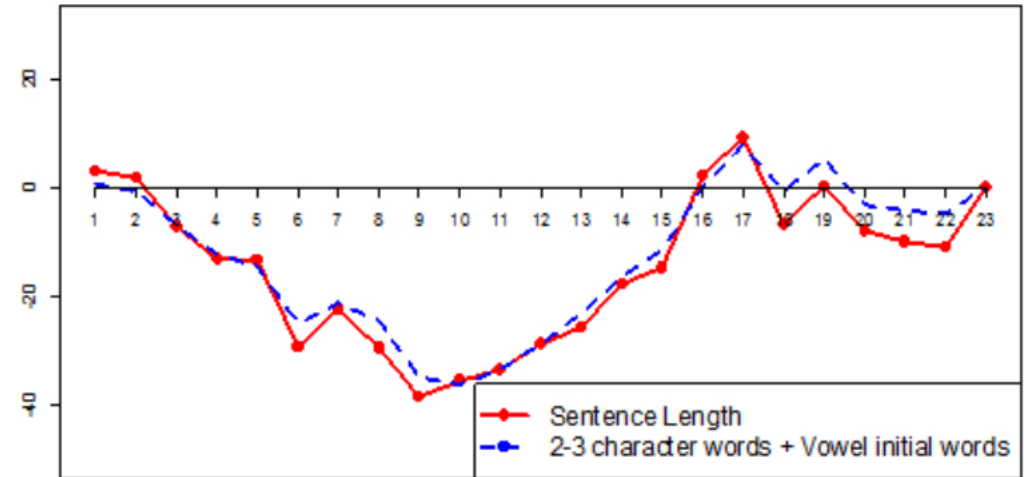
	enough	while	whilst	upon
Hamilton	0.59	0.26	0	2.93
Madison	0	0	0.47	0.16
Disputed texts	0	0	0.34	0.08
Co-authored texts	0.18	0	0.36	0.36



Mosteller &
Wallace (1964)

...and Epic Failures

- Andrew Morton in the early '60 adapted Cumulative Sum – CUSUM or QSUM (a method which originally was used in the industrial quality control) to be used in texts.
- BBC live show (1993)
 - Documents of convicted criminals were attributed to ... the Secretary of State for Justice!!!



The Revolution

'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship¹

John Burrows
University of Newcastle, Australia

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

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

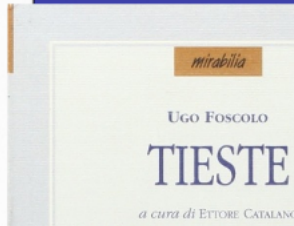


1. e
2. che
3. di
4. la
5. a
6. il
7. non
8. l
9. in
10. per
11. le
12. si
13. con
14. i
15. è
16. un
17. del
18. da
19. più
20. d
21. gli
22. ma



ALESSANDRO MAN
ADELCHI

0.13
0.11
0.09
...



mirabilia
UGO FOSCOLO
TIESTE
a cura di ETORE CATALANO

0.10
0.14
0.08
...



Alfieri
Tragedie

0.11
0.12
0.10
...



Goldoni
La locandiera
Introduzione di Giorgio Strehler

0.13
0.10
0.07
...

$$\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i} \right|$$



	A	B	C	D	E	F
1		AlessandroManzoni_Adelchi	AlessandroManzoni_IIContediCarmagnola	AlessandroManzoni_InniSacri	AlessandroManzoni_Odi	AlessandroManzoni_Poesiegio
2	AlessandroManzoni_Adelchi	0	0,481290655	0,666926925	0,738545533	0,5688
3	AlessandroManzoni_IIContediCarmagnola	0,481290655	0	0,746348745	0,814261157	0,6543
4	AlessandroManzoni_InniSacri	0,666926925	0,746348745	0	0,633663965	0,6348
5	AlessandroManzoni_Odi	0,738545533	0,814261157	0,633663965	0	0,7338
6	AlessandroManzoni_Poesiegiovanili	0,568820863	0,654375023	0,634854567	0,733827682	
7	CarloGoldoni_Gl'Innamorati	0,980786338	0,936018177	1,013723738	1,101305203	0,9504
8	CarloGoldoni_IICampiello	1,016924762	1,031300757	1,018625104	1,092680684	0,9293
9	CarloGoldoni_IIServitorediduePadroni	0,94860233	0,926662976	0,976288639	1,080804722	0,9181
10	CarloGoldoni_IITeatrocomico	0,915941412	0,896367382	0,971870697	1,085346366	0,8984
11	CarloGoldoni_IIVentaglio	1,011953514	1,00041649	1,074888328	1,131792245	0,9972
12	CarloGoldoni_IRusteghi	1,089096895	1,124315967	1,047451935	1,1240649	0,9778
13	CarloGoldoni_LaBottegadelcaff�	0,997940632	0,980781404	1,069965126	1,139058754	0,9938
14	CarloGoldoni_LaFamigliadell'Antiquario	0,97647637	0,968110166	1,038499373	1,080510085	0,9530
15	CarloGoldoni_LaLocandiera	0,97946604	0,952399004	1,052505983	1,110322738	0,9561
16	CarloGoldoni_LeBaruffechiozzotte	1,051753673	1,103993387	1,018834132	1,082447143	0,9423
17	CarloGoldoni_LeFemminepuntigliose	0,940334542	0,938723973	1,008461186	1,076438004	0,9179
18	CarloGoldoni_LeSmanieperlaVilleggiatura	1,023938091	0,964832878	1,056736183	1,148650567	1,0072
19	CarloGoldoni_UnadelleultimeserediCarnovale	1,045847956	1,085480986	1,047945641	1,10681856	0,9481
20	VittorioAlfieri_Agamennone	0,684514153	0,743793265	0,829452563	0,905939302	0,7011
21	VittorioAlfieri_Antigone	0,73781244	0,801189414	0,824156384	0,91495815	0,7219
22	VittorioAlfieri_Brutosecondo	0,675393312	0,675937144	0,830722082	0,910174086	0,6681
23	VittorioAlfieri_Filippo	0,69672213	0,73856813	0,806194725	0,93419818	0,6694
24	VittorioAlfieri_MariaStuarda	0,693145931	0,715015202	0,806081448	0,948928306	0,6738
25	VittorioAlfieri_Merope	0,735463235	0,783055974	0,855979157	0,971583955	0,7091
26	VittorioAlfieri_Mirra	0,76329317	0,819104452	0,864045202	0,9659327	0,7605
27	VittorioAlfieri_Oreste	0,70530237	0,777981376	0,829335057	0,930970217	0,7154
28	VittorioAlfieri_Ottavia	0,762895099	0,791949819	0,874379901	0,96265065	0,7225
29	VittorioAlfieri_Saul	0,645417404	0,735038238	0,760393582	0,871007648	0,6668
30						

Dendrograms

Frequent Collocations and Authorial Style

David L. Hoover

New York University, New York, USA

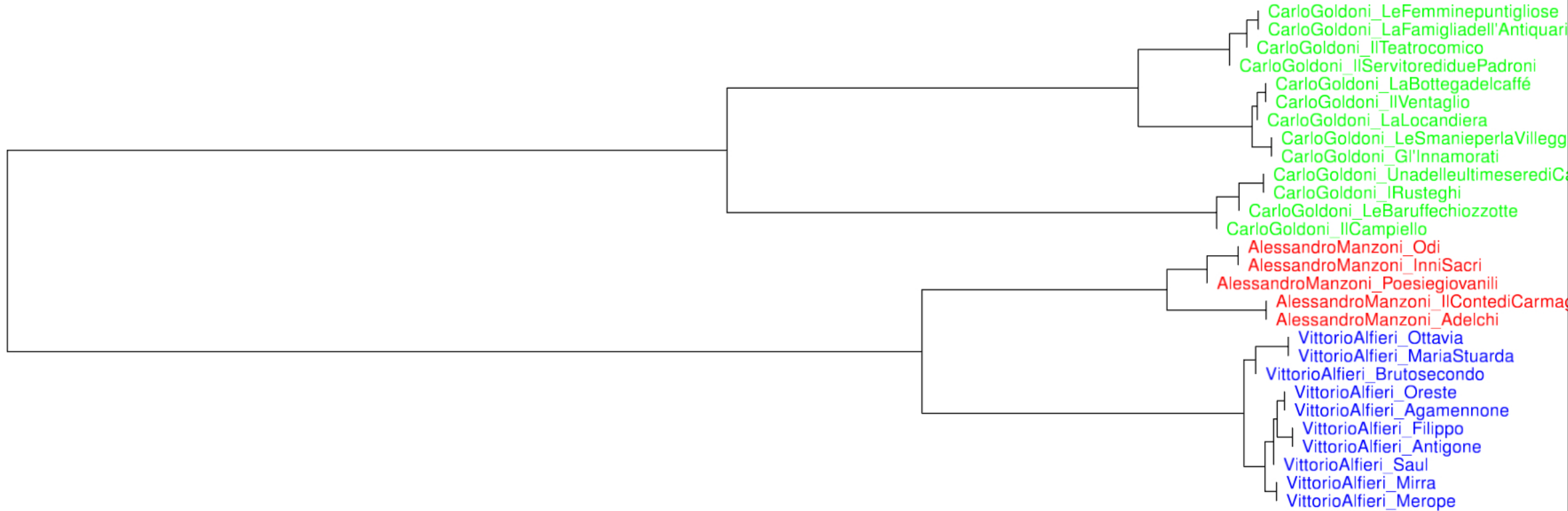
Abstract

This paper examines the effectiveness of multivariate analysis of the frequencies of frequent collocations in characterizing authorial style. Cluster analyses of collocations over various spans, types, and linkages are performed on groups of texts by known authors to determine how well the frequencies of those collocations correctly attribute the texts to their authors and distinguish them from texts by other authors. In each case the results are compared with those based on the frequencies of frequent words and the frequencies of frequent sequences of words. Cluster analyses based on frequent words and sequences ascribe many of the texts to their correct authors. However, analyses based on frequent collocations are more accurate for several groups of texts, sometimes producing more completely correct attributions than analyses based on either words or sequences and sometimes producing the only completely correct attributions. They also produce results for small groups of problematic novels and critical texts extracted from the larger corpora that are often superior to those based on

"Literary and
Linguistic
Computing"
18, no. 3
(2003): 261–
83



Letteratura Italiana Cluster Analysis



8

6

4

2

0

100 MFW Culled @ 0%
Classic Delta distance



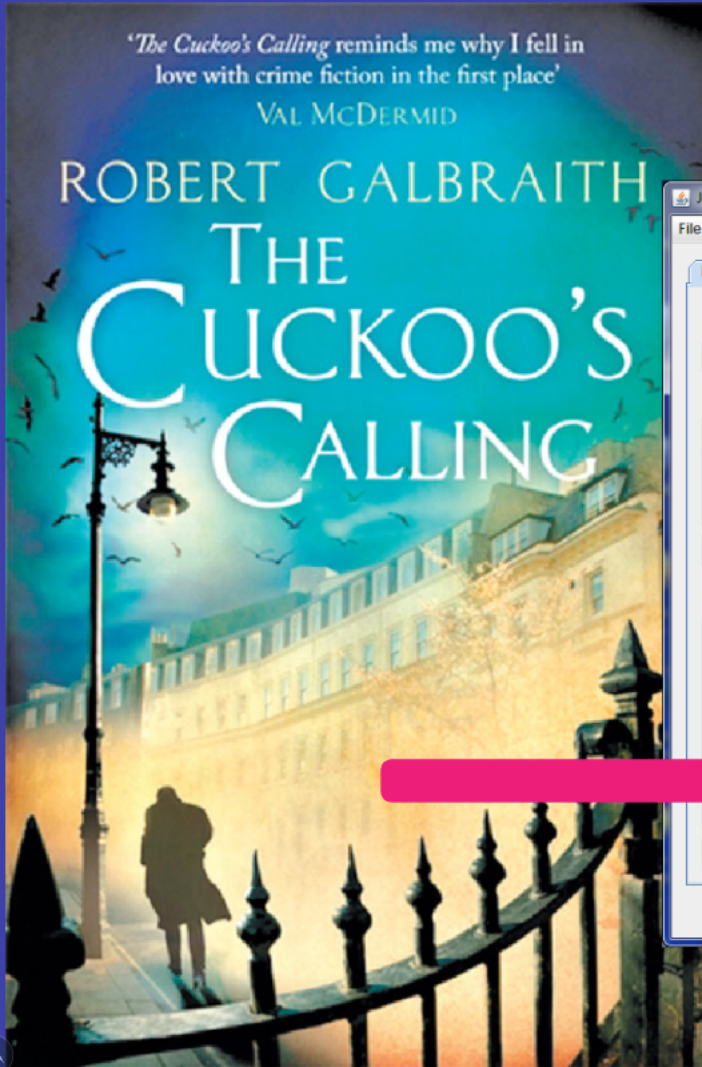
Authorship Attribution

Juola vs.
Rowling

Musil Project
(in Verona)



Juola vs. Rowling



JGAAP

File Help

Documents | Canonizers | Event Drivers | Event Culling | Analysis Methods | Review & Process

Language Notes

English

Unknown Authors

Title	Filepath
2012 Climate Strategy.pdf	C:\Users\Anthony\Desktop\heartland leaked docs\PDF files\2012 Climat...

Add Document Remove Document

Known Authors

- Authors
 - Z. McElrath
 - Joe Bast

Add Author Edit Author Remove Author

Finish & Review Next →



Musil Project (in Verona)

Die „Tiroler Soldaten-Zeitung“ soll den Tiroler Landesverteidigern die neuesten Nachrichten über die militärische Lage, ferner über einzelne militärische Begebenheiten auf den Kriegsschauplätzen sowie über sonstige Angelegenheiten, die das Interesse der Armee oder Einzelner betreffen, vermitteln.

„Für Gott, Kaiser und Vaterland!“

Die „Tiroler Soldaten-Zeitung“ wird wöchentlich 3 mal wöchentlich im Standorte des Landesverteidigungskommandos erscheinen und im Wege der Abfertigung zur weiteren Verbreitung den Kommanden, Gruppen und Anstalten unentgeltlich zugestellt werden. — Von Zivilpersonen ist für 1 Exempl. der Zeitung 10 h zu entrichten.

Tiroler

Soldaten



-Zeitung



Das Reinerträgnis ist einem Hilfsfonds zu Gunsten

verwundeter Tiroler Landesverteidiger gewidmet

Nummer 1

2. Juni 1915

8 Uhr morgens

Inhalt: Geleitwort Sr. Exzellenz des Landesverteidigungskommandanten in Tirol, G. d. K. Dankl . Italiens Krieg . Unbedingte Schweigepflicht im Kriege . Trentino, Triest und Italien . Das glänzende Ergebnis der Kriegsanleihe . Schlachtenberichte . Sonstige militärische Nachrichten . Die Absperrung Englands . Politische Nachrichten . Volkswirtschaftliche Nachrichten . Sonstige Mitteilungen . Letzte Nachrichten.

Soldaten, Kameraden!

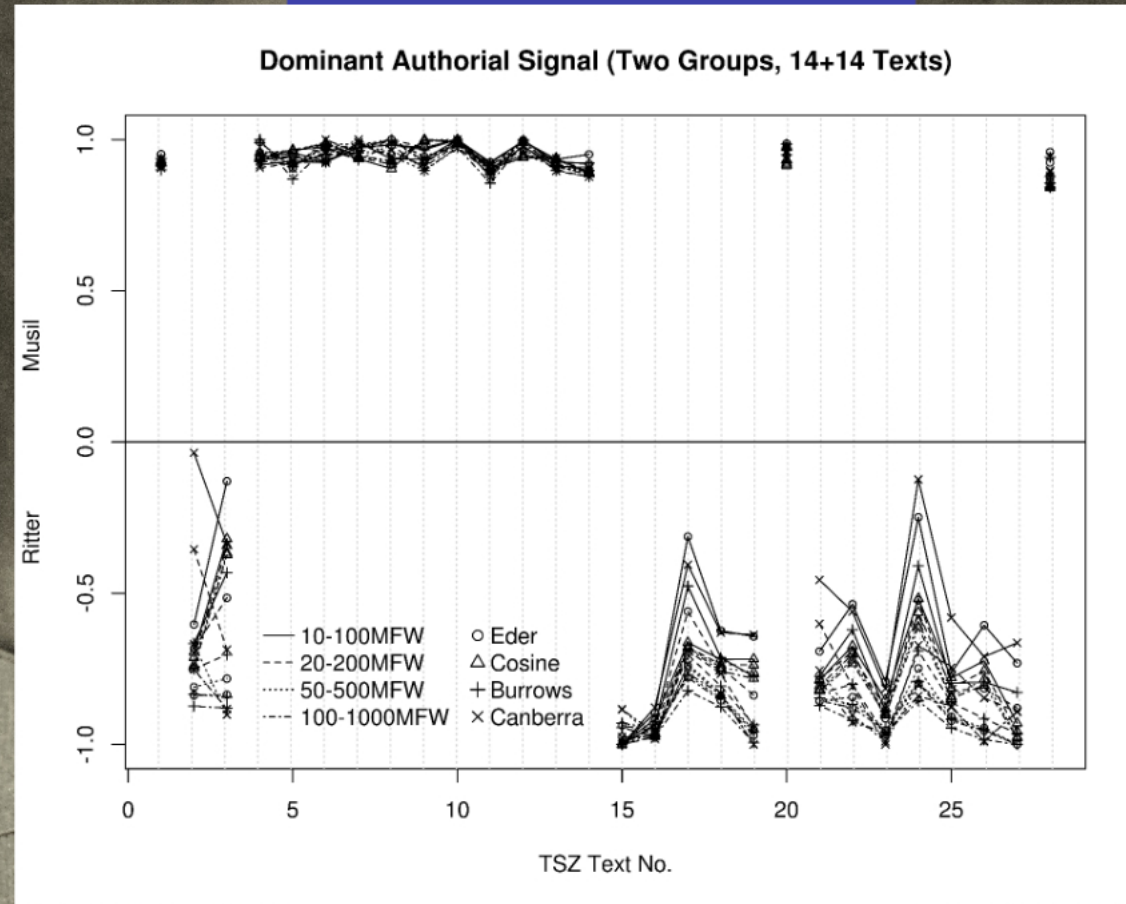
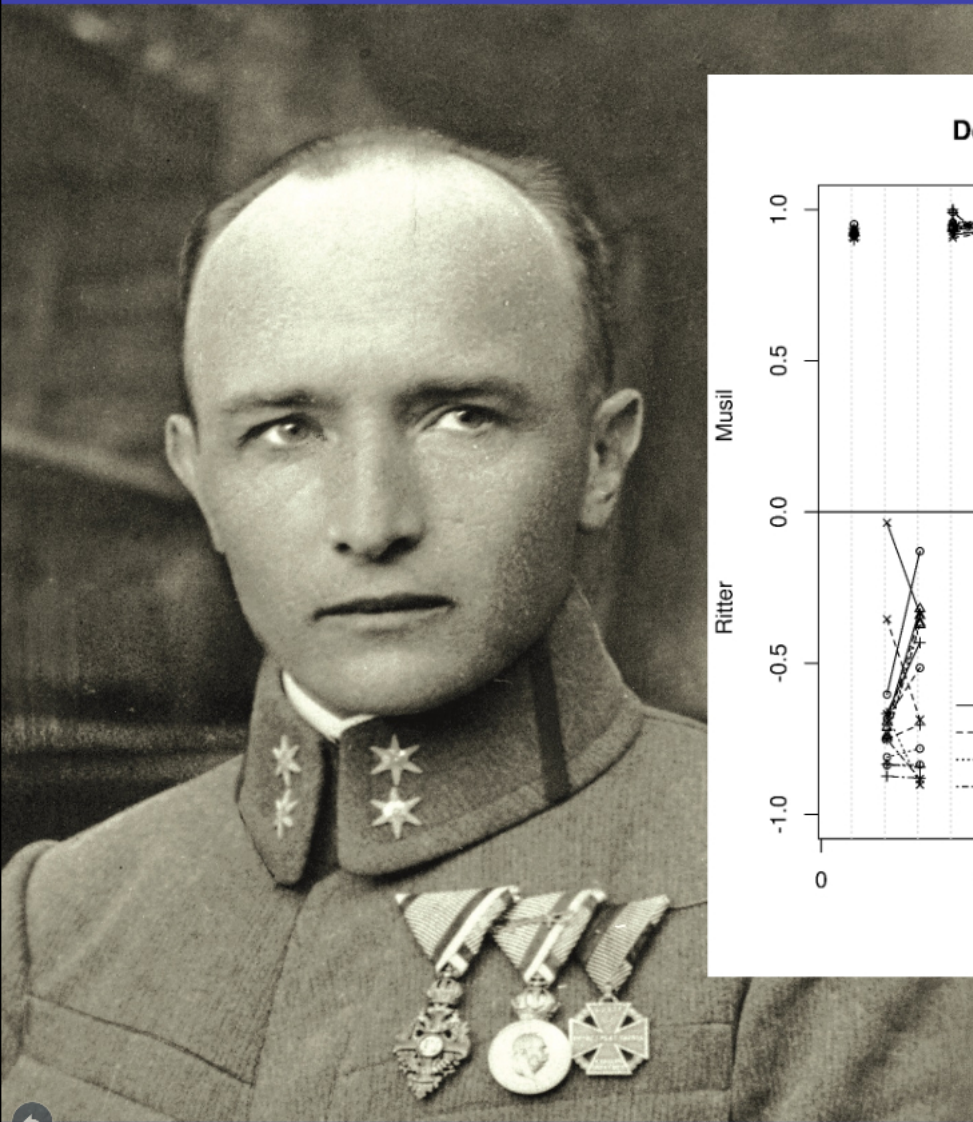
Italien, das durch mehr als ein Menschenalter hindurch im „Dreibunde“ reich und kräftig geworden, hat uns schmählich verraten und meuchlerisch überfallen, es hat die Treue gebrochen, die es uns im Bündnisvertrage zugesagt. Se. Majestät unser erhabener Kaiser und Herr hat, um seinen Völkern diesen neuen Krieg zu ersparen, Konzessionen gemacht, wie sie solche der kühnste Irredentist nicht hätte träumen lassen. Aber die Italiener wollen durchaus den Krieg. Sie wollen kriegerische Lorbeeren holen, wollen

nie gewagt. Ihr Vorgehen ist daher nicht nur treulos und heimtückisch, sondern auch feige. Mit kalter Ruhe und Entschlossenheit stehen wir diesem ekeligen Treiben gegenüber. Wir vertrauen auf unseren Herrgott, auf unser Recht und auf uns selbst, wir vertrauen auf Tirol, dem ruhmreichen Lande Andreas Hofers, das so oft schon seine Feinde mit blutigen Köpfen hinausgeworfen. Unerschütterlich wie die Berge dieses herrlichen Landes werden wir im Vereine mit unserem kampferprobten, tapferen deutschen Verbündeten

Robert Musil



Albert Ritter

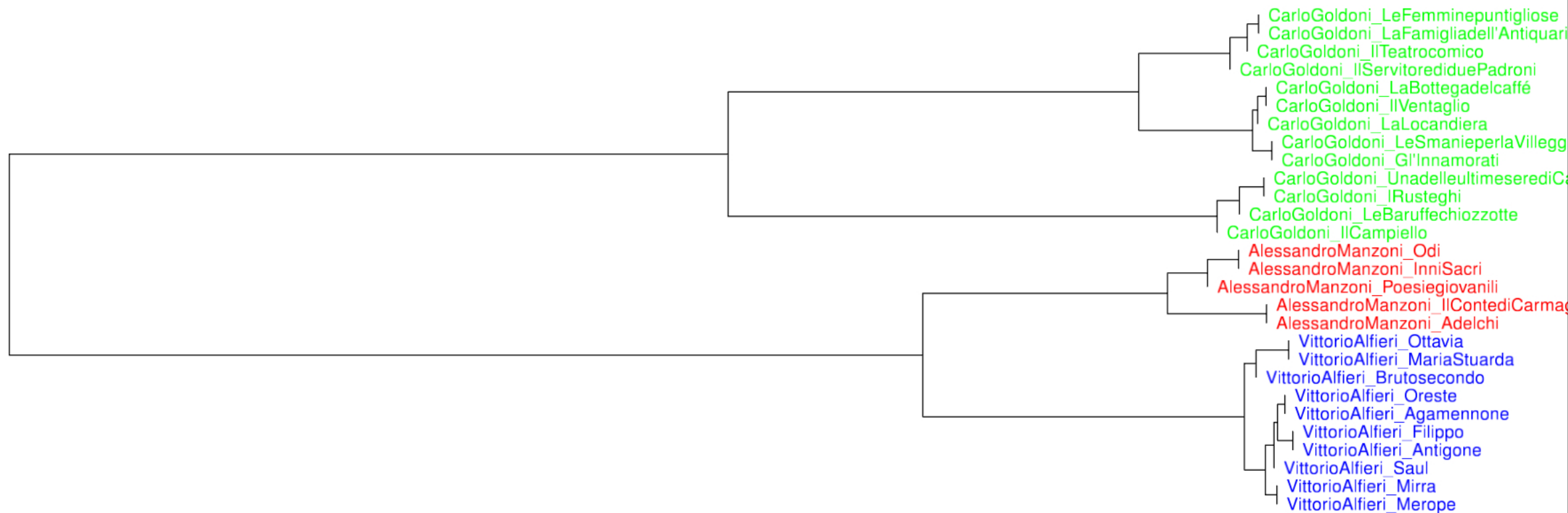


Network Analysis

**not only
authorship
attribution...**



Letteratura Italiana Cluster Analysis



8

6

4

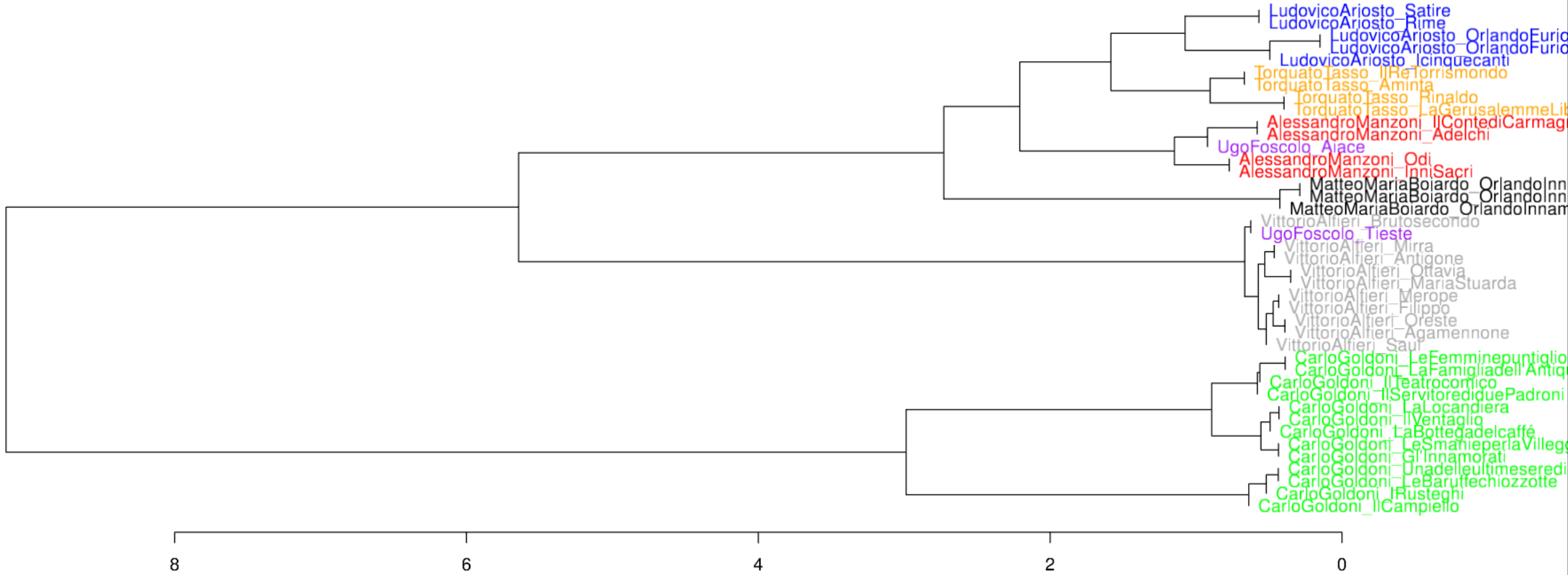
2

0

100 MFW Culled @ 0%
Classic Delta distance



Letteratura Italiana Cluster Analysis



8

6

4

2

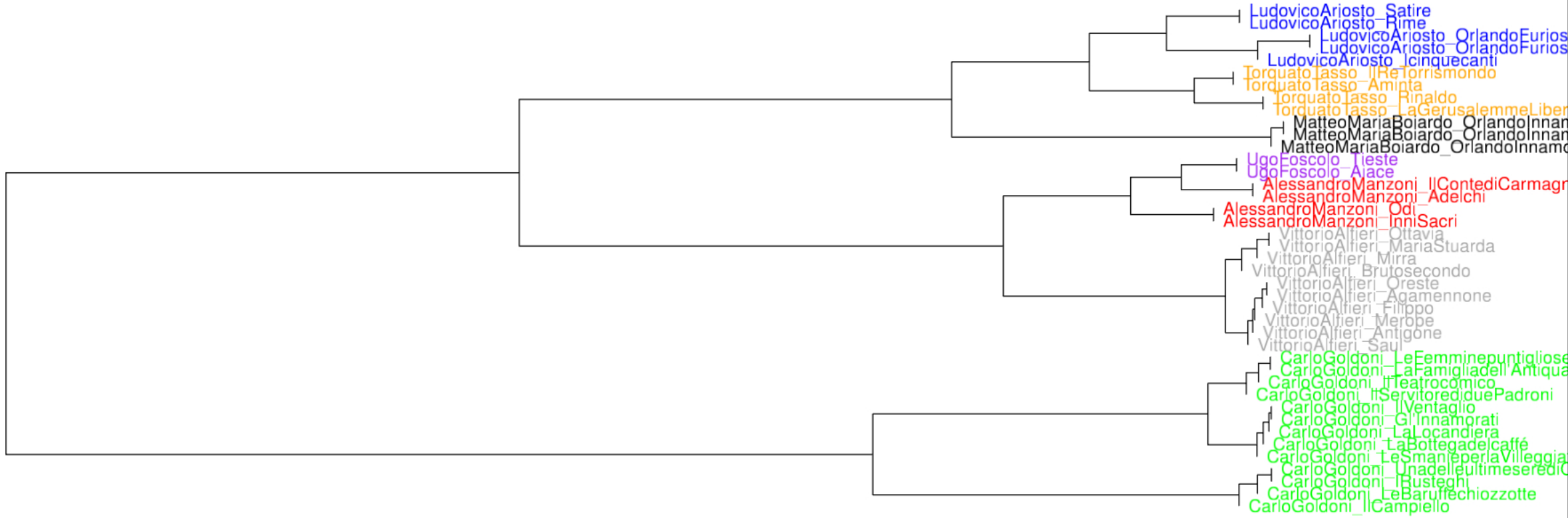
0

100 MFW Culled @ 0%
Classic Delta distance





Letteratura Italiana Cluster Analysis

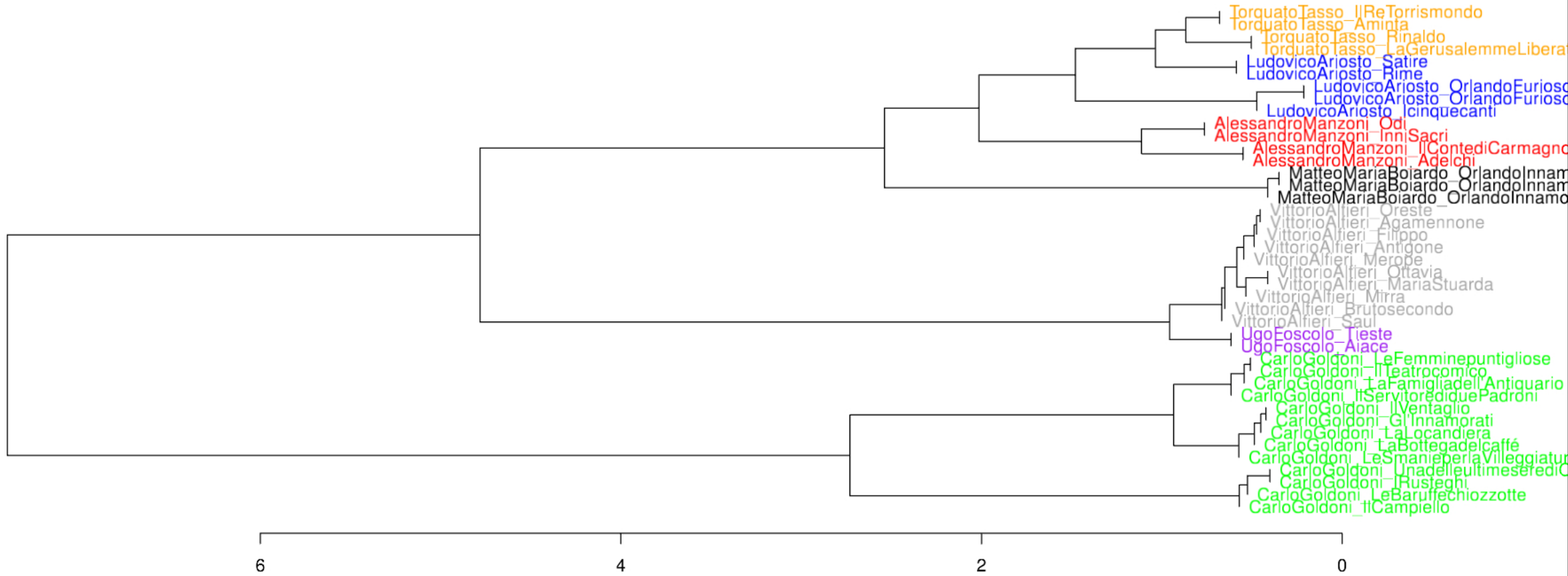


8 6 4 2 0

200 MFW Culled @ 0%
Classic Delta distance



Letteratura Italiana Cluster Analysis



300 MFW Culled @ 0%
Classic Delta distance



Stylometry with R: A Package for Computational Text Analysis

by Maciej Eder, Jan Rybicki and Mike Kestemont

Abstract This software paper describes ‘Stylometry with R’ (*stylo*), a flexible R package for the high-level analysis of writing style in stylometry. Stylometry (computational stylistics) is concerned with the quantitative study of writing style, e.g. authorship verification, an application which has considerable potential in forensic contexts, as well as historical research. In this paper we introduce the possibilities of *stylo* for computational text analysis, via a number of dummy case studies from English and French literature. We demonstrate how the package is particularly useful in the exploratory statistical analysis of texts, e.g. with respect to authorial writing style. Because *stylo* provides an attractive graphical user interface for high-level exploratory analyses, it is especially suited for an audience of novices, without programming skills (e.g. from the Digital Humanities). More experienced users can benefit from our implementation of a series of standard pipelines for text processing, as well as a number of similarity metrics.

Introduction

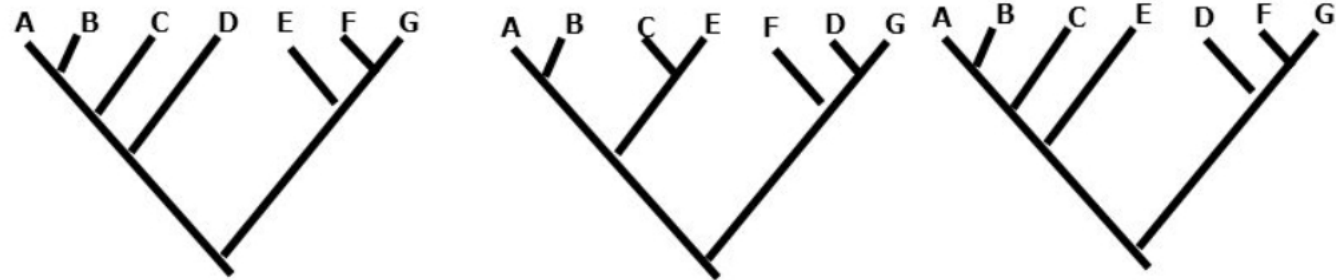
Authorship is a topic which continues to attract considerable attention with the larger public. This claim is well illustrated by a number of high-profile case studies that have recently made headlines across the popular media, such as the attribution of a pseudonymously published work to acclaimed

The R Journal
Vol. 8/1, Aug.
2016

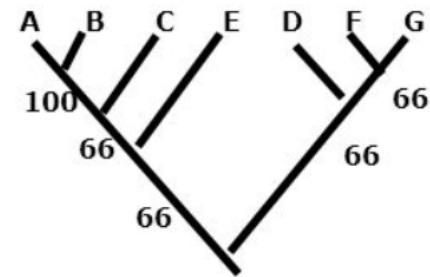


Consensus Trees

Majority rule consensus



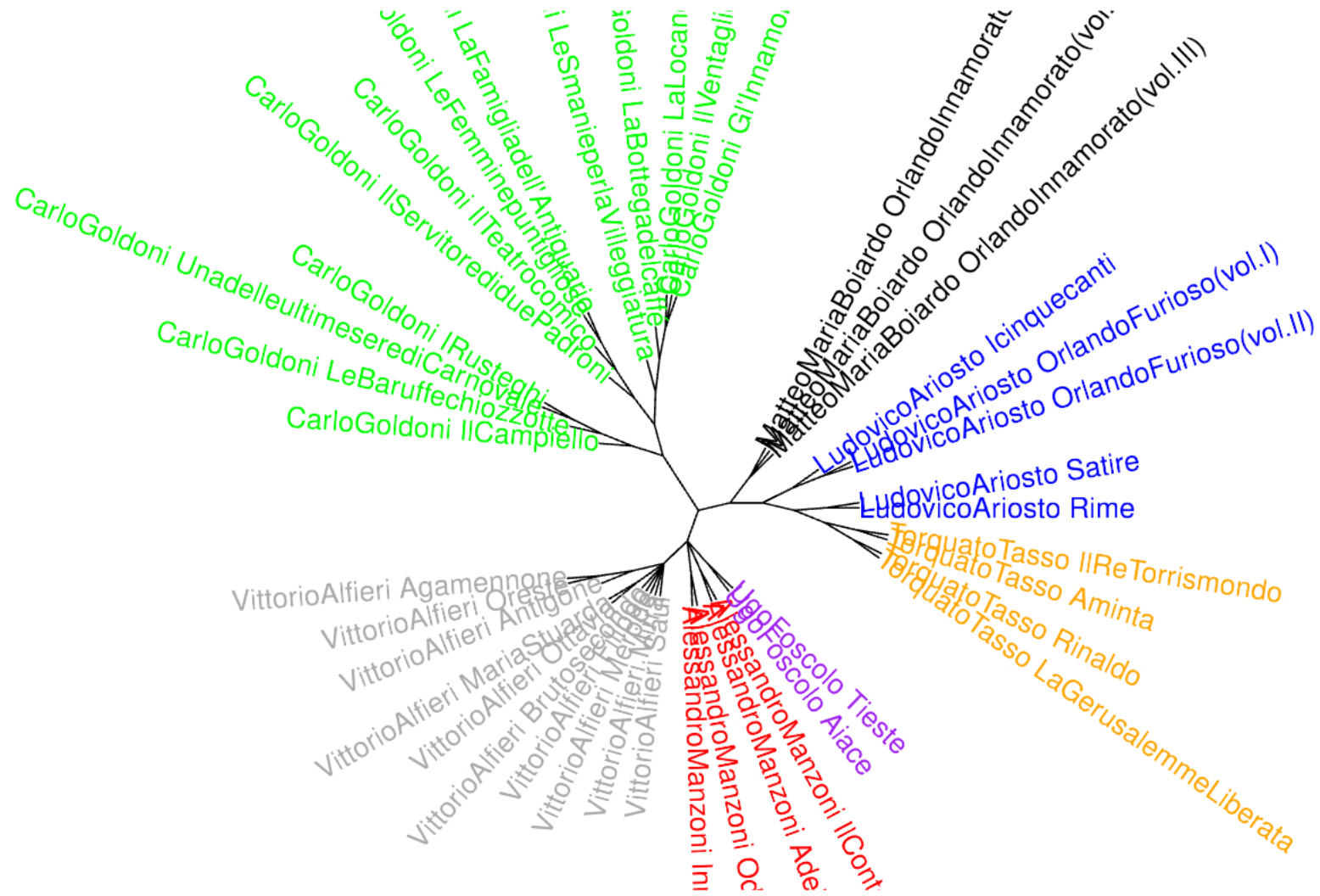
Numbers indicate frequency of clades in the fundamental trees



MAJORITY-RULE CONSENSUS TREE

Letteratura Italiana

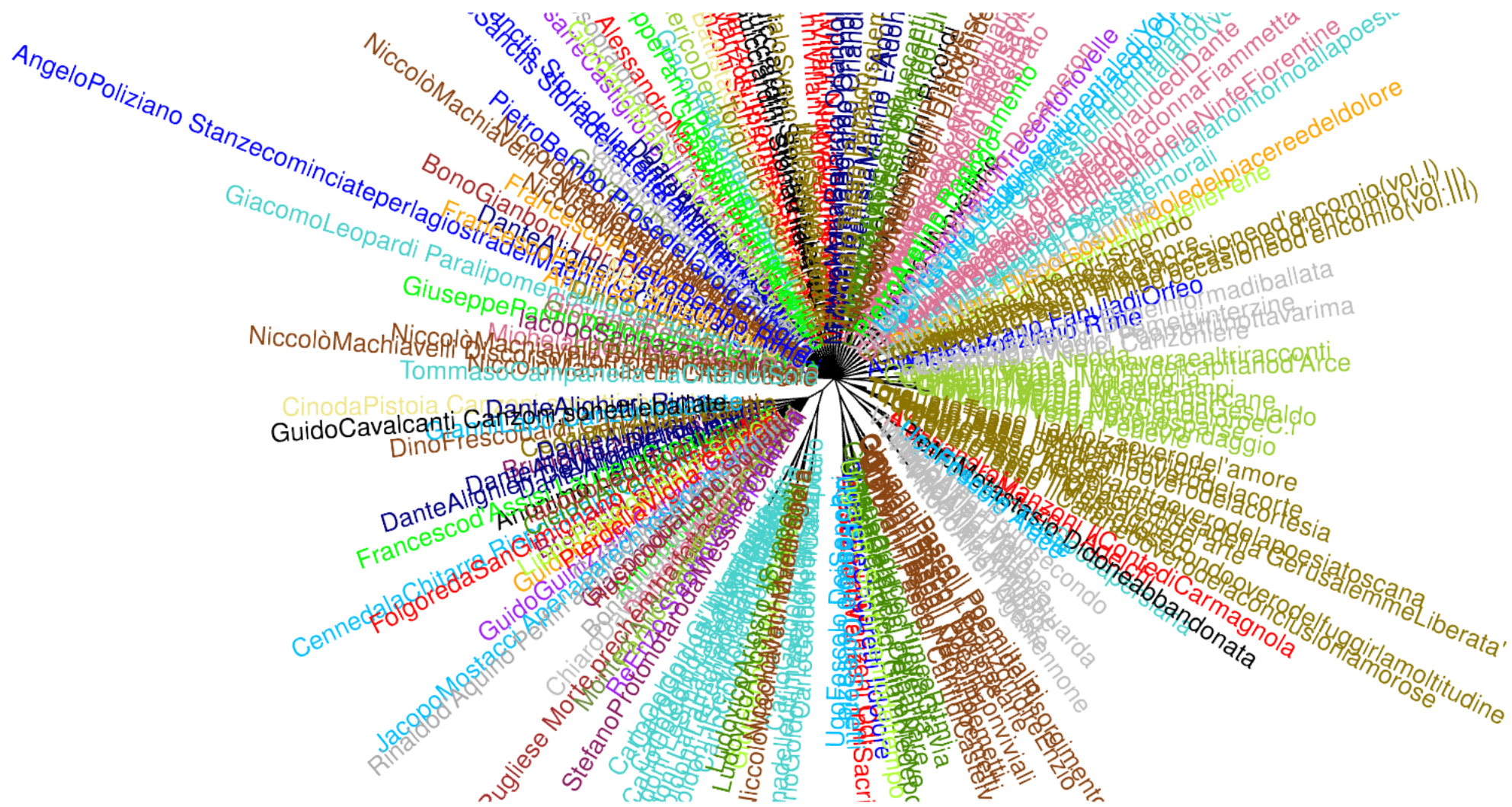
Bootstrap Consensus Tree



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



Visualization in stylometry: Cluster analysis using networks

Maciej Eder

Pedagogical University of Kraków, Poland

Institute of Polish Language, PAS

Abstract

The aim of this article is to discuss reliability issues of a few visual techniques used in stylometry, and to introduce a new method that enhances the explanatory power of visualization with a procedure of validation inspired by advanced statistical methods. A promising way of extending cluster analysis dendrograms with a self-validating procedure involves producing numerous particular 'snapshots', or dendrograms produced using different input parameters, and combining them all into the form of a consensus tree. Significantly better results, however, can be obtained using a new visualization technique, which combines the idea of nearest neighborhood derived from cluster analysis, the idea of hammering out a clustering consensus from bootstrap consensus trees, with the idea of mapping textual similarities onto a form of a network. Additionally, network analysis seems to be a good solution for large data sets.

Correspondence:

Maciej Eder, Institute of Polish Studies, Pedagogical University of Kraków, ul. Podchorążych 2, 30-084 Kraków, Poland.

E-mail:

maciejeder@gmail.com

1 Introduction

Most of the computational methods used in stylometry have been originally introduced to solve authorship attribution problems. This fact had an

algorithms, suitable for classification tasks, derived mostly from the field of biometrics, nuclear physics, or software engineering, that could be easily adopted to authorship attribution. They include naïve Bayes classification, support vector machines,

"Digital
Scholarship in
the Humanities",
Vol. 32, No. 1,
2017



M. Eder

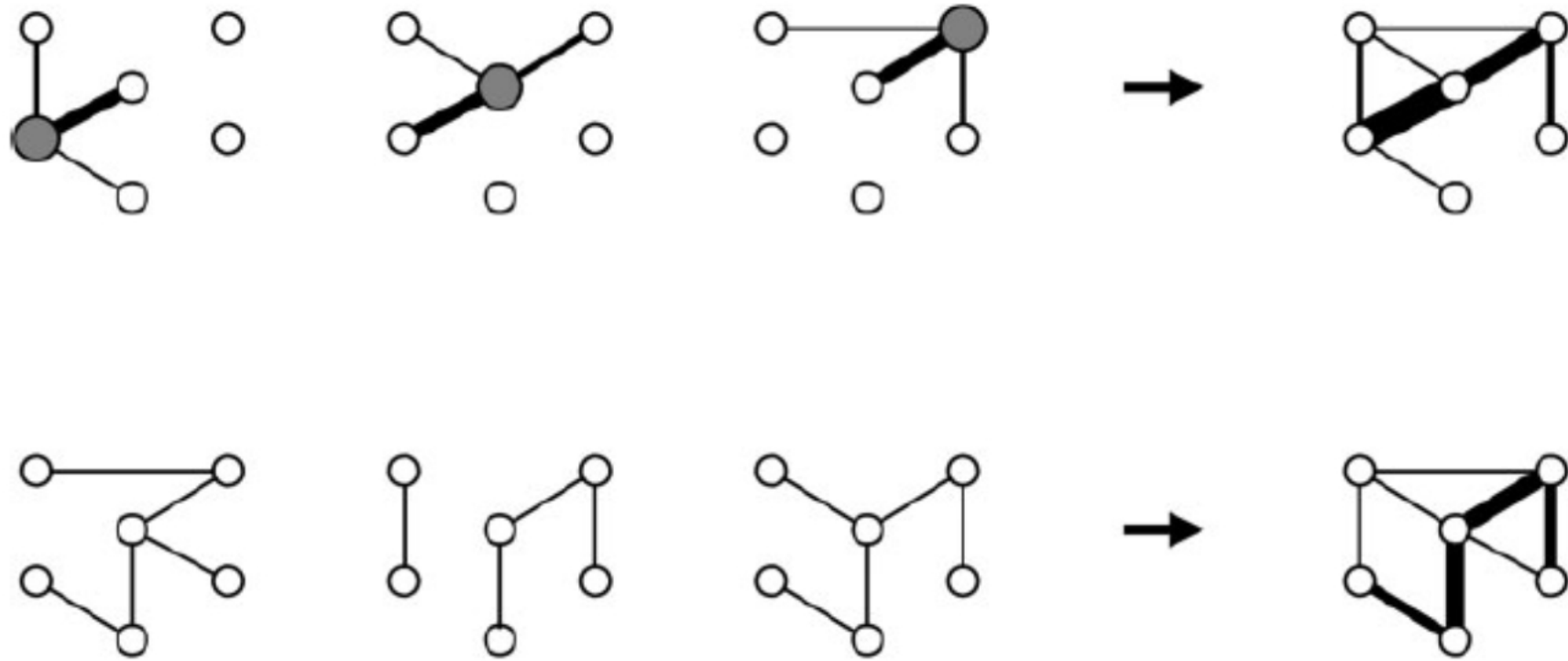
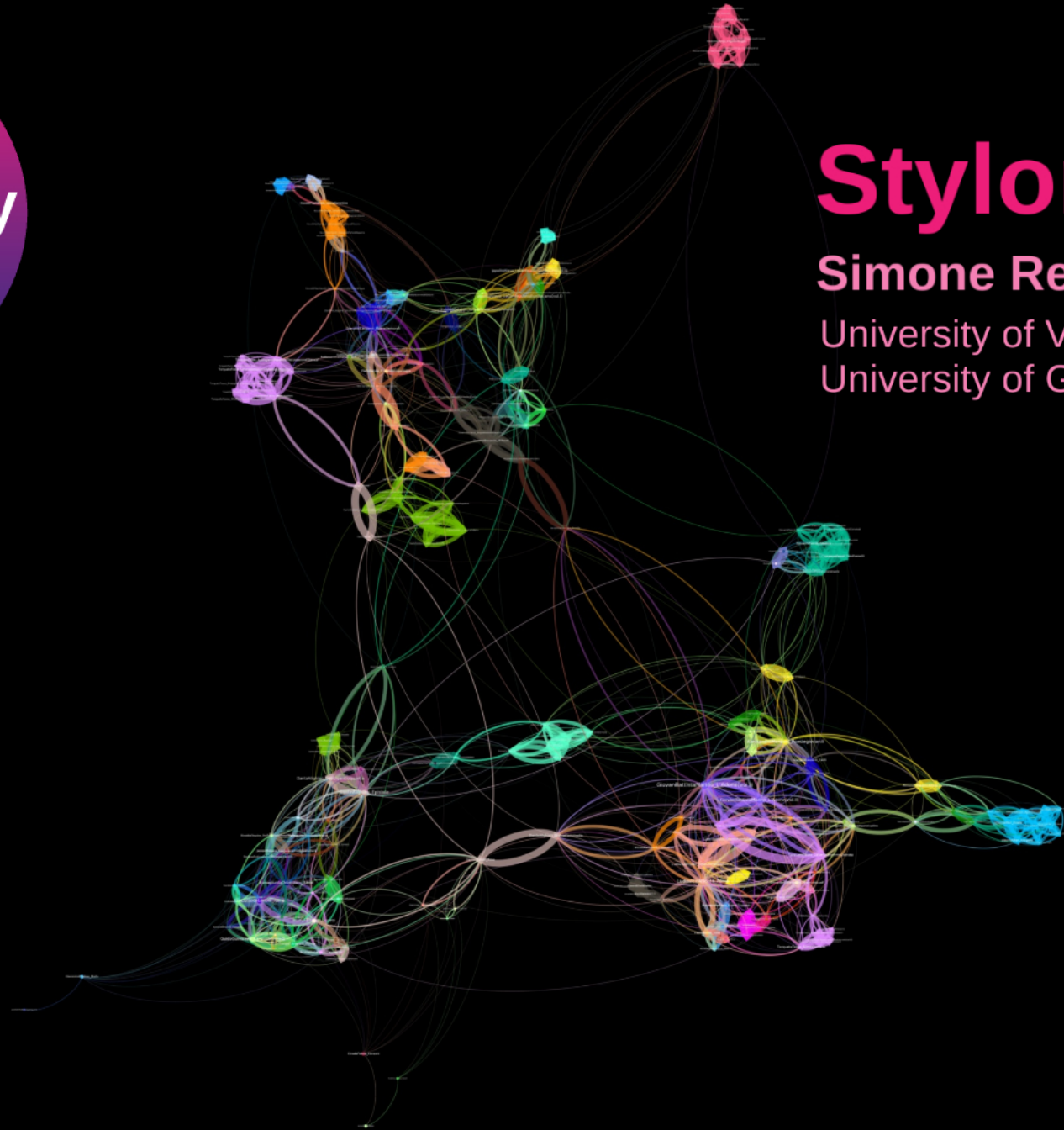


Fig. 6. Two algorithms of mapping textual relations: establishing weighted links to a nearest neighbor and two runners-up (top); producing a consensus network (bottom).

Stylometry

**Authorship
Attribution**

**Network
Analysis**

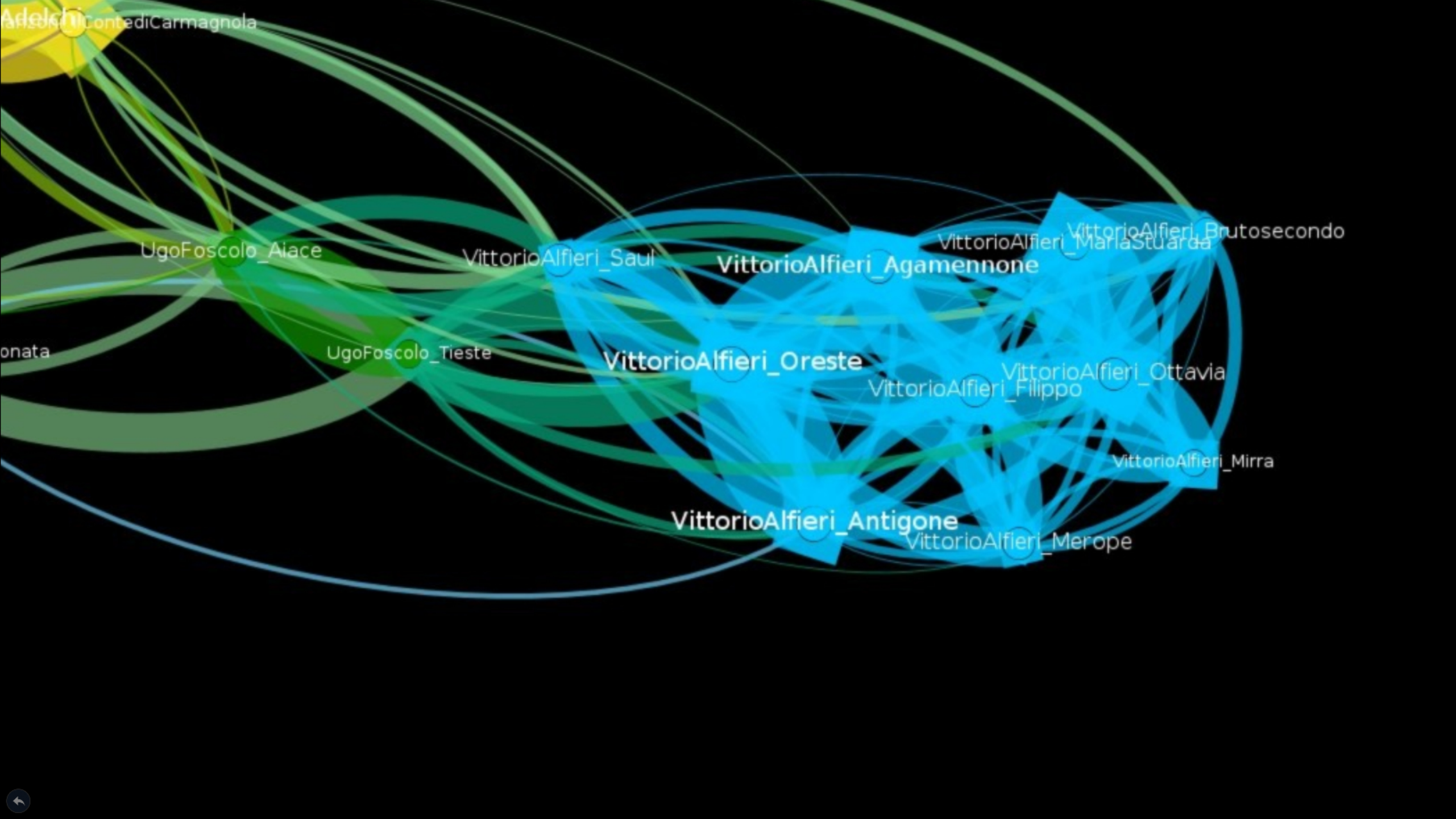


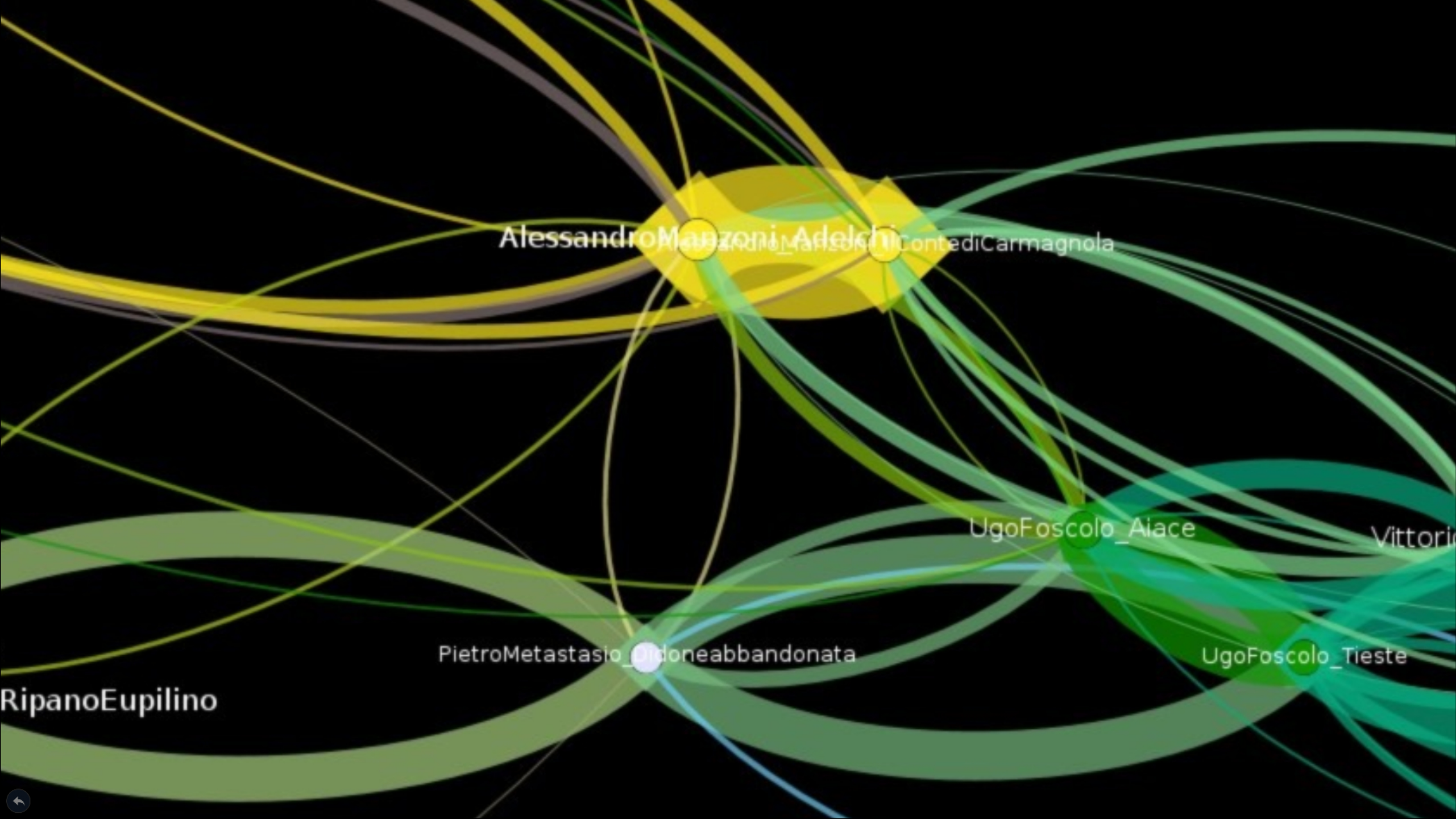
Stylometry

Simone Rebora

University of Verona

University of Göttingen





Alessandro Manzoni Adelchi ContediCarmagnola

UgoFoscolo_Aiace

Vittorio

PietroMetastasio Iloneabbandonata

UgoFoscolo_Tieste

RipanoEupilino





AlessandroManzoni_Poesiegiovanili

GiacomoLeopardi_Canti

GiacomoLeopardi_Poesievarie

UgoFoscolo_LeGrazie

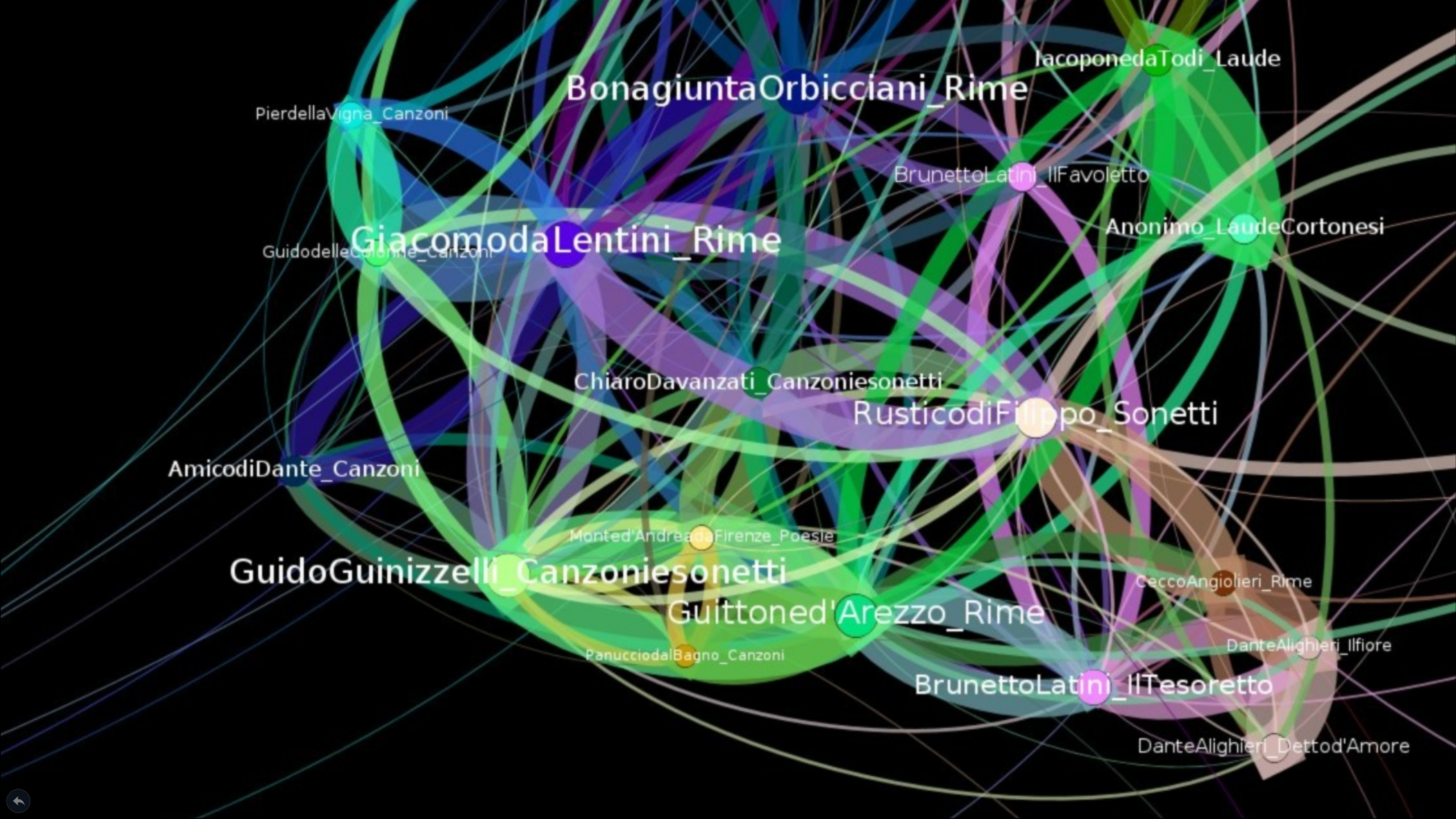
UgoFoscolo_DeiSepolcri

UgoFoscolo_OdieSonetti

GiuseppeParini_LeOdi

GiuseppeParini_IlGiorno

AngeloPoliziano_StanzecominciateperlagiostradelMagnificoGiulianode'Medici



BonagiuntaOrbicciani_Rime

Iaconedatodi_Laude

PierdellaVigna_Canzoni

BrunettoLatini_IlFavoletto

GiacomodaLentini_Rime

Anonimo_LaudeCortonesi

GuidodelleColonne_Canzoni

ChiaroDavanzati_Canzoniesonetti

RusticodiFilippo_Sonetti

AmicodiDante_Canzoni

Monted'AndreadaFirenze_Poesie

GuidoGuinizzelli_Canzoniesonetti

CeccoAngiolieri_Rime

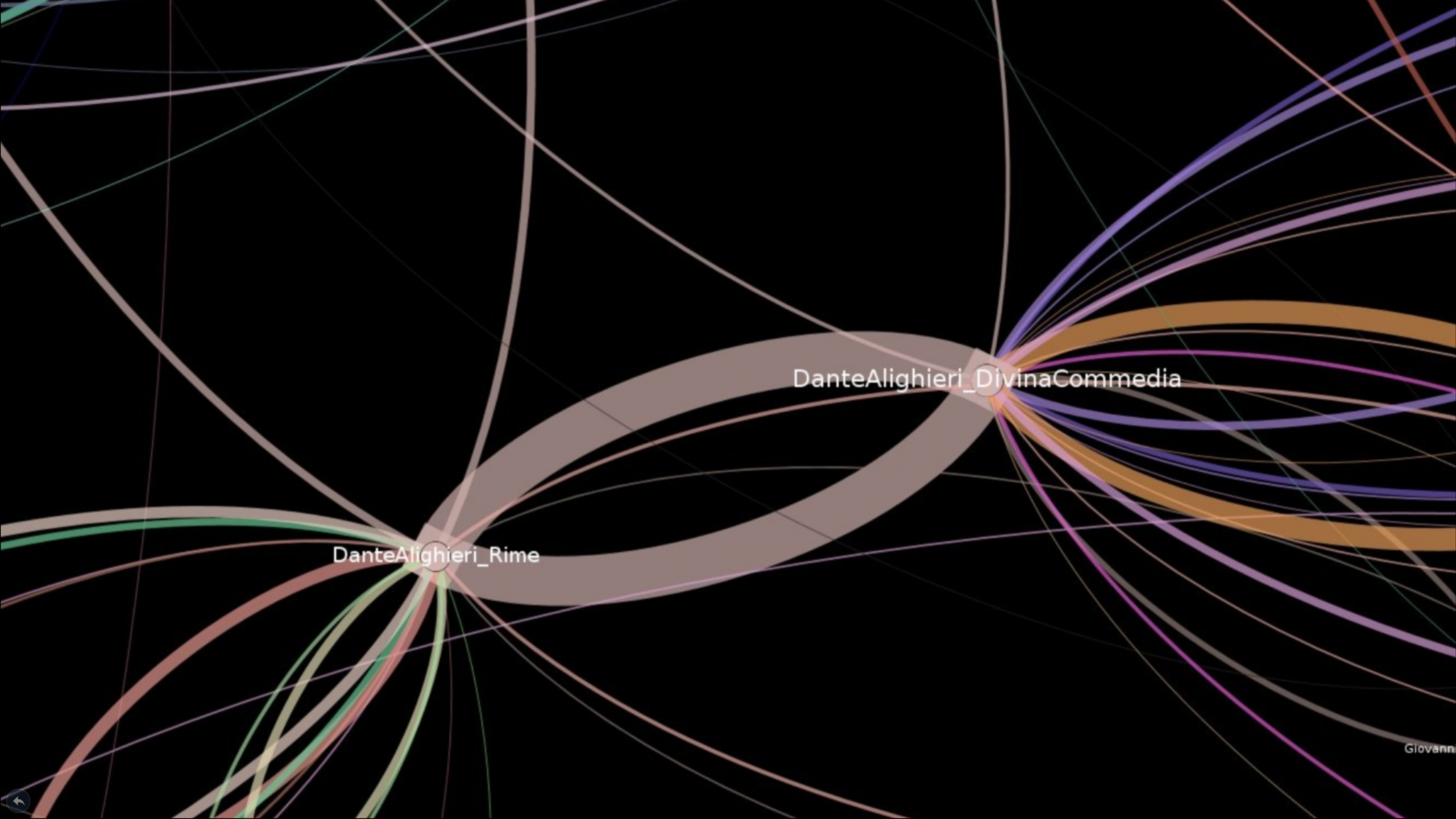
Guittone'dArezzo_Rime

DanteAlighieri_Ilfiore

PanucciodalBagno_Canzoni

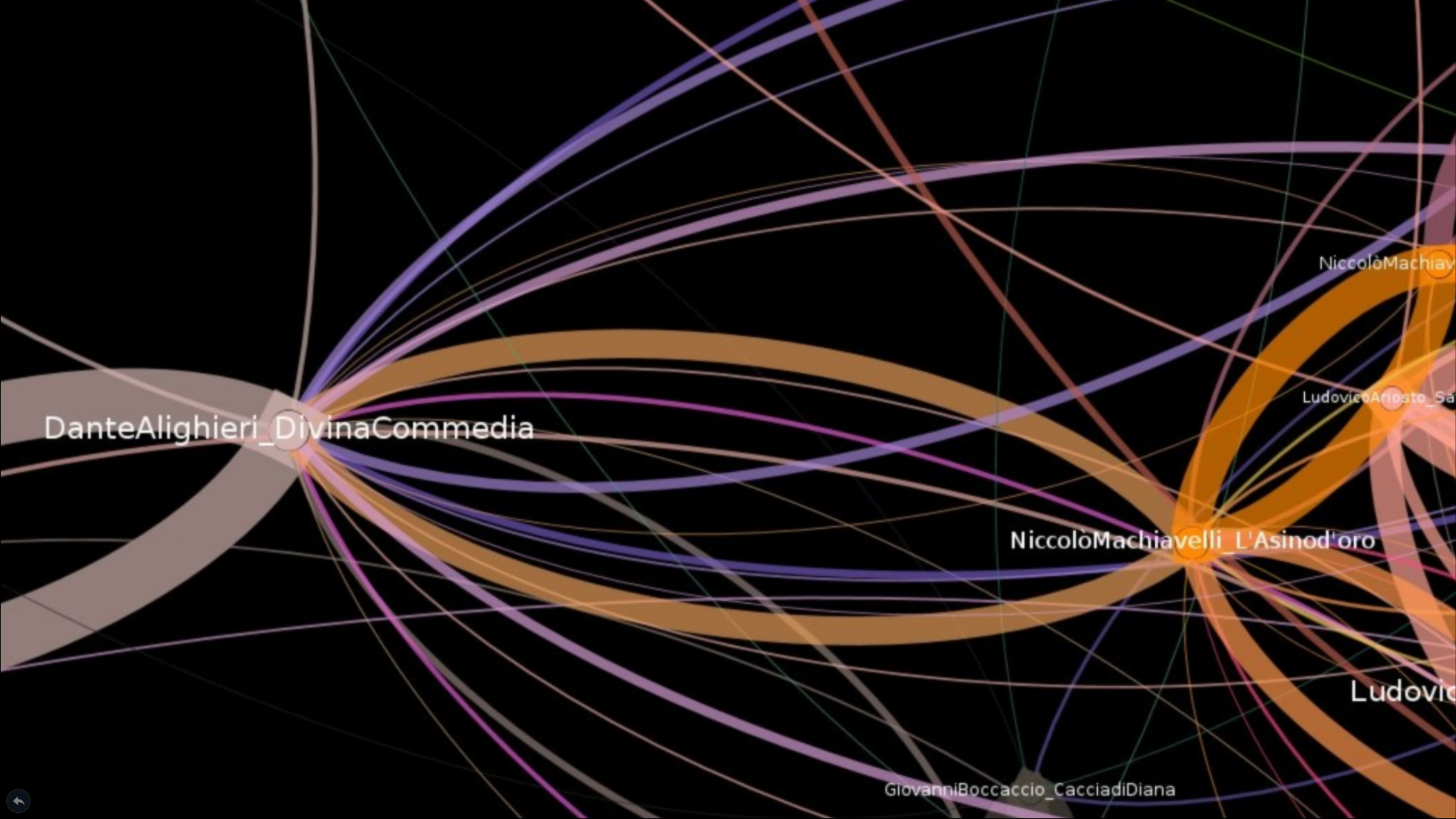
BrunettoLatini_IlTesoretto

DanteAlighieri_Dettod'Amore



DanteAlighieri_Rime

DanteAlighieri_DivinaCommedia



DanteAlighieri_DivinaCommedia

NiccolòMachiavelli_L'Asinod'oro

NiccolòMachiavelli

LudovicAriosto_Satire

Ludovic

GiovanniBoccaccio_CacciadiDiana

