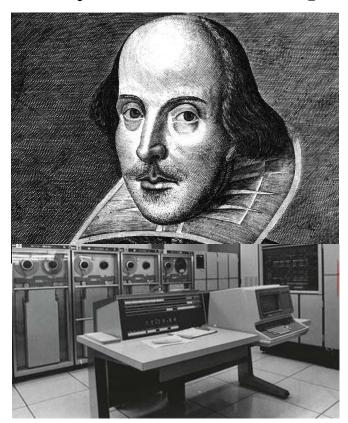
Stylometry and authorship attribution





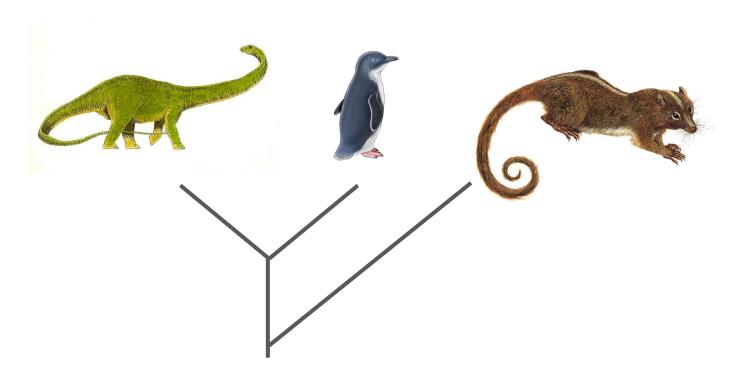
How similar? How different?



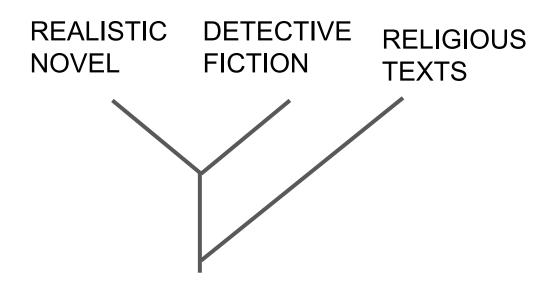
Think in terms of distances: what evolutionary path each species has undergone



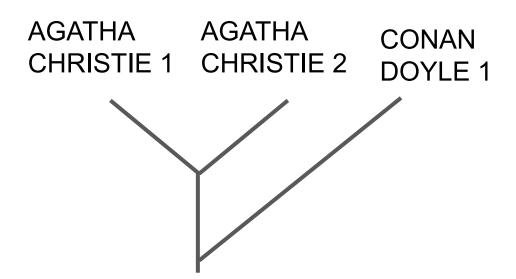
Think in terms of distances: what evolutionary path each species has undergone



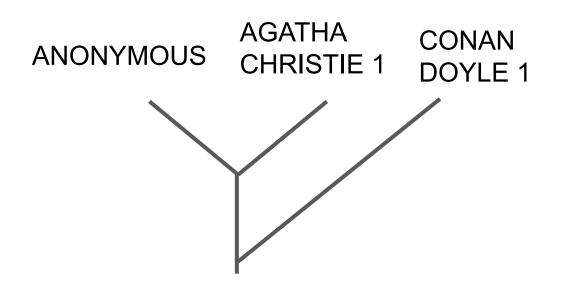
Continue thinking in terms of distances



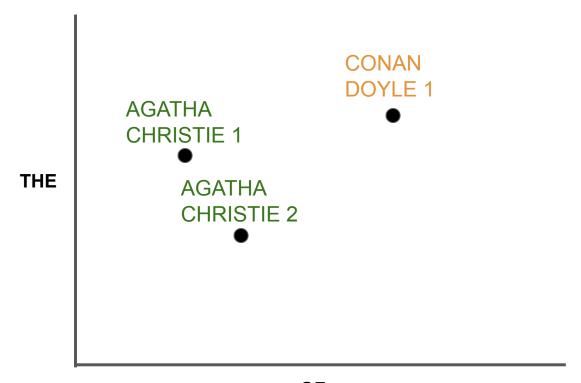
Continue thinking in terms of distances



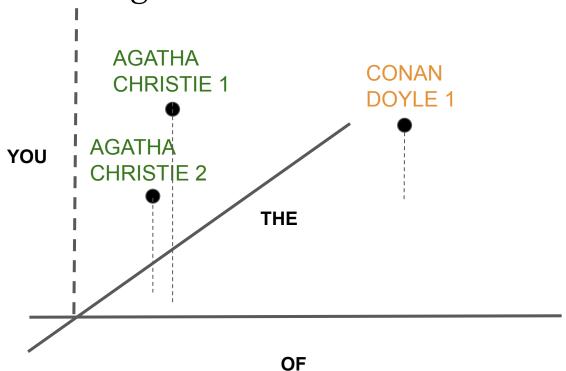
Continue thinking in terms of distances



Continue thinking in terms of distances - 2D

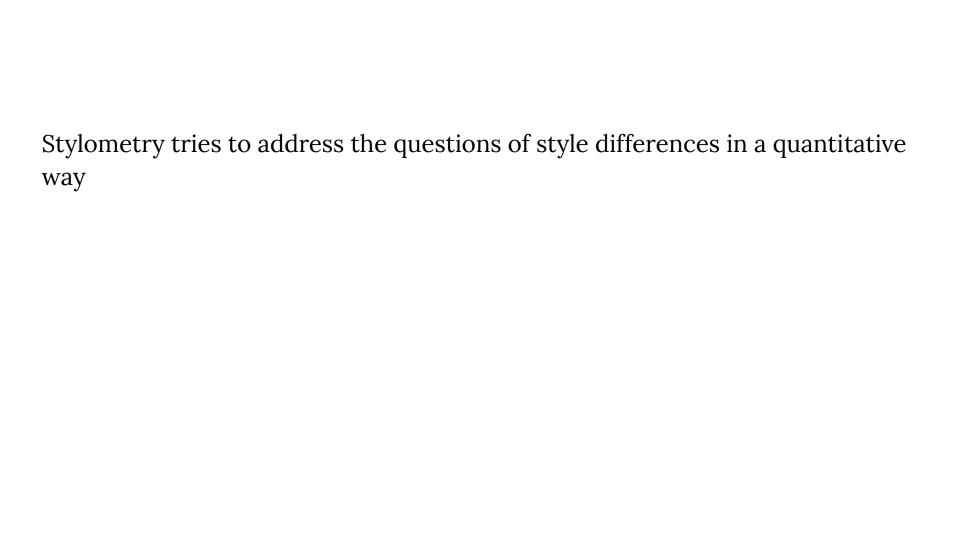


Continue thinking in terms of distances - 3D



http://everynoise.com/

an ente	VV CL V C	demoscene	N W	ambeat	cumbia paragu
african metal oshare kei metropopolis	I.	new	ack smooth	slow come ti	urntablism
rab metal esband	/I	rock ska espanol	new jack	swing	
pop punk anime dance- j-metal punk urbano thai ir	punk bolivian	polish	reggae	dutch hip h	OD
j-metal ' punk urbano thai ir christian rock	die	danish po	p j-rap	nz hip hop	bay area h
core rock		teen pop	litch beats	hip pop	
Undarian metal		brazilian ska		chinese hip hop	dirty texa
deep pop emo	alent show	brazilian ska ntempo fusion vedish idol pop		albanian h	ip hop
thrash-groove metal	ai rock down	tempo fusion		nerdcore	scratch
vegas indie	SW	vedish idol pop	musik anak-	anak t	rap queen
etal spanish modern rock	TI	innish dance pop	h grime	pop rap	
shalternative rock alternative DOD rock				trap n	nusic
rap motal canadian conte	emporary country		kaneka	old school ned	erhon
j-indie rock cristiano		french indietronica		zouglou	gangster r
wrestling modern cour	ntry rock	antiviral pop	israe	li hip hop	jazz boon
ash metal madchester chillwave		deep rai		danish hip ho	Ρ
onic metal	new wave	рор	. Р	ost-disco rap	traditional
j-indie rock cristiano wrestling modern cour ash metal madchester chillwave nic metal rock christian alternative rock nintendocore	australian	POP deep turkish po	op pinoy	nip nop	jazz rap
nintendocore	Vanor non	shibuya-kei		southorn s	spanish regga oul blues
post-grunge deep chiptune modern alternative rock	deep german indie		urban cont	emporary	Jui Diues
glam modern alternative rock	spanish po	op rock	czech hin	hop south	orn hip hop
canadian rock alberta cour	itry	bulgarian rock	malaysian hip h	OD din	ty south mp
J-poppunk lussball new v	vave	volkstumliche m	usik zim hip	hop pative	cy south rap
rock turkish metal	gauze pop t	urbo folk	h	orrorcore bulge	arian hin hon
post-grunge deep chiptune modern alternative rock glam canadian rock alberta cour j-poppunk fussball new v l christian punk rap meta rock turkish metal gothic post-punk progressive metal al swedish alternative rock I metal northern irish indie alternative metalcore trio batak modern uplift witch hou	macedonian p	redneck .		finnish hip hop	man mp nop
progressive metal		dangdut koplo	polynesi	an pop	azonto
al swedish alternative rock		vapor soul	norwegi	an pop rap	
metal northern irish indie		belgian pop	drill	argentine reggae	deep
modern unlift witch hou	ukrainia	n indie austr	alian hip hop	romanian	hip hop
i-punk medieval rock paraguay	an rock	antideutsche te	cnobrega	ia pop trap m	exicano
c metalcore rock chapin		swedi	sn soul	thai hip hop	
alt-indie rock scottish new way	e latin arona pop	neue deutsche weile	indie p	oop rap sv	wedish reggae
I metal swedish alternative rock I metal northern irish indie alternative metalcore trio batak modern uplift witch hot j-punk medieval rock paraguay c metalcore rock chapin alt-indie rock scottish new way	iaun arena pop	south afr	ican non	arabic hin ho	n



Stylometry tries to address the questions of style differences in a quantitative way

A "style" in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Stylometry tries to address the questions of style differences in a quantitative way

A "style" in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Many features were proposed: word lengths, sentence lengths, word frequencies, character n-grams, part-of-speech tags, versification features like rhythm and rhyme patterns...

Stylometry tries to address the questions of style differences in a quantitative way

A "style" in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Many features were proposed: word lengths, sentence lengths, word frequencies, character n-grams, part-of-speech tags, versification features like rhythm and rhyme patterns...

Authorship attribution is a sub-problem of stylometry: **how unique patterns of textual features help to catch author identity**

Stylometry and the 19th century positivism

- New Shakespeare Society (feminine endings in Shakespeare's verse)
- T.C. Mendenhall (style and spectral analysis)
- Dating *Dialogues* of Plato (German school and V. Lutoslawski)

Stylometry as a quest for the "fingerprint", "individual handwriting".

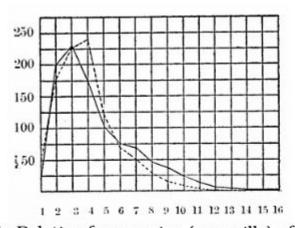
Sounds not too much like literary analysis, right?

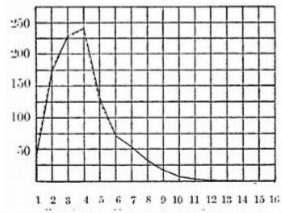
Because it is not: more criminalistics and forensics



T.C. Mendenhall (1841-1924)
The characteristic curves of composition (1887)

T.C. Mendenhall: word lengths



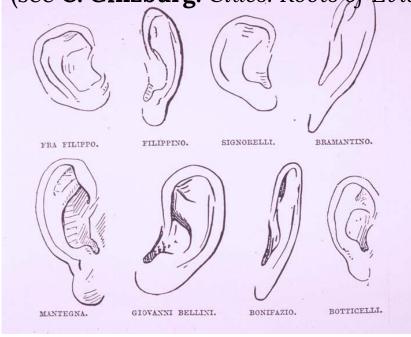


works of W. Shakespeare (dashed) and F. works of W. Shakespeare (dashed) and C. (facsimile).

FIG. 1: Relative frequencies (per mille) of word- FIG. 2: Relative frequencies (per mille) of wordlengths measured by number of characters in lengths measured by number of characters in Bacon (full line). Source: Mendenhall 1901: 104 Marlowe (full line). Source: Mendenhall 1901: 105 (facsimile).

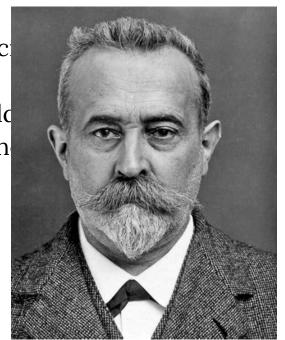
Giovanni Morelli and study of authorship in paintings

(see C. Ginzburg. Clues: Roots of Evidential Paradigm)



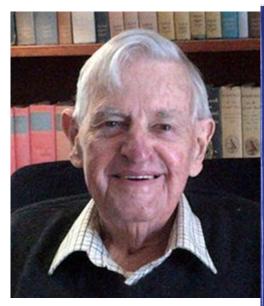
"Anthropometrics" and fingerprints

- "Anthropometrics" of Alphonse Bertillon: very preconseveral sets of measurements.
- One dimension of measurements (e.g. height) could we combine 3 of them? 6? 12? Hundreds or even the modern stylometry actually do

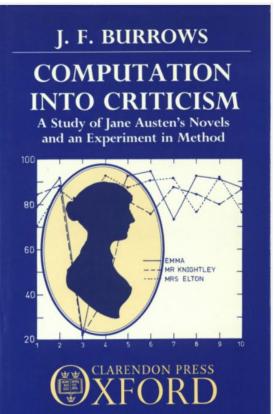


Alphonse Bertillon (1853-1914)

Distances between texts: Burrow's delta (2002)



John Burrows



Delta-distance

$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - z(y_i)|}{n}$$

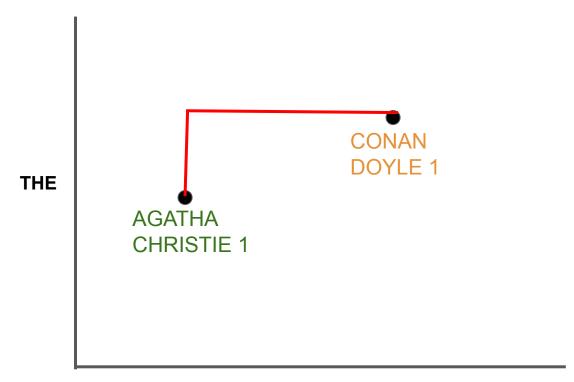
On the distances

http://versologie.cz/talks/2017chicago/

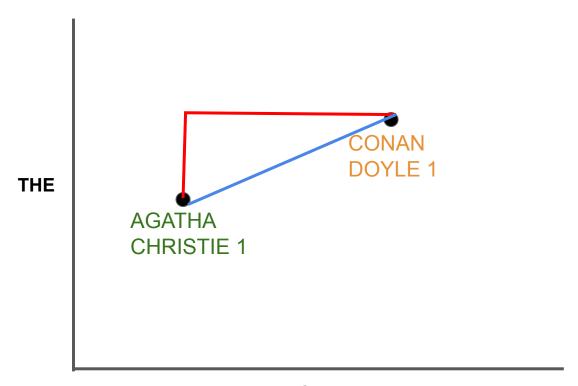


Petr Plecháč (Czech Academy of Sciences)

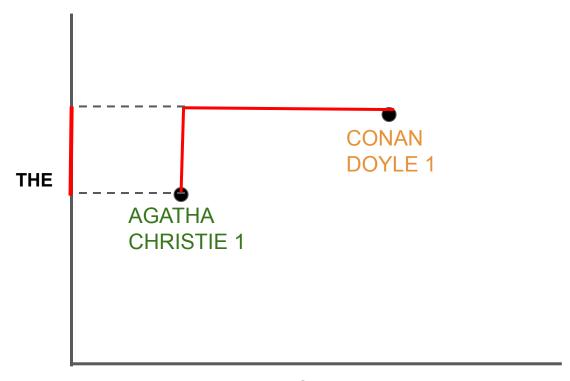
Manhattan (also Delta)



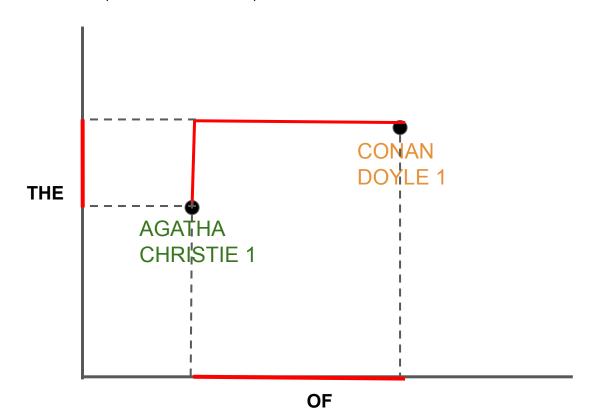
Manhattan + Euclidean

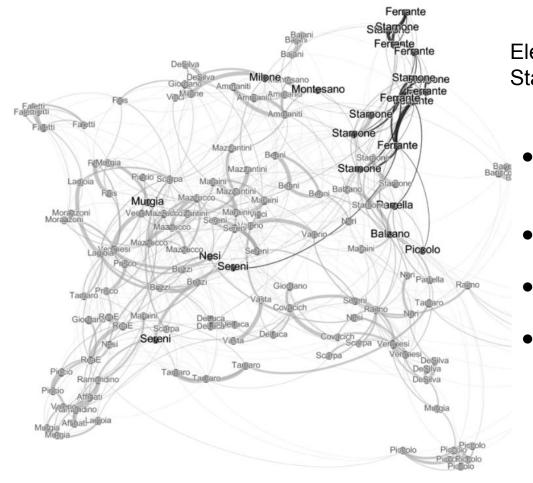


Manhattan (also Delta)



Manhattan (also Delta)





Elena Ferrante / Anita Raja / Domenico Starnone?

- Bootstrap Consensus Network based on Delta measure. 3 nearest neighbours of each novel, connected
- 150 Italian novels late 20th early 21st century
- From 100 to 1000 MFWs (bootstrap, increment 100)
- List of closest candidates (Starnone the closest)

Eder 2018

"stylo" package for R

Eder, M., Rybicki, J. and Kestemont, M. (2016).

Stylometry with R: a package for computational text analysis. R Journal 8(1): 107-121.



Delta-distance

$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - (z(y_i))|}{n}$$

Normalization

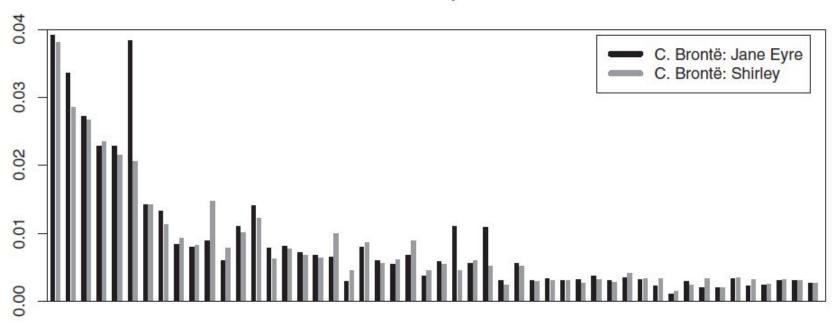
• To overcome this effect a scaling of values is performed. We compute standardised **z-scores** for relative frequencies

$$z(x_i) = rac{x_i - \mu(x)}{\sigma(x)}$$

• ...which means z-score of some frequency = this frequency **minus** mean frequency of all observations (μ) **and divide** by standard deviation

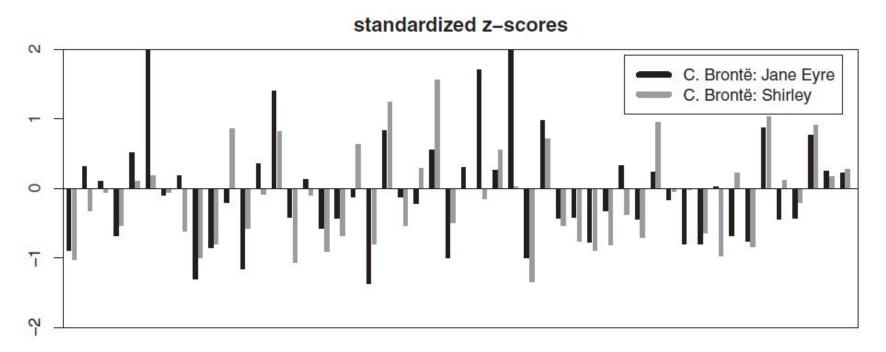
Relative frequencies curve (Zipf's curse)

relative frequencies



Evert et al. 2017

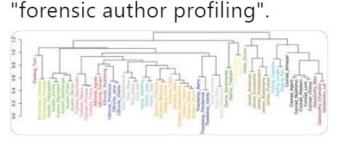
Normalized frequencies



some deep stats!!



Based upon our statistical analysis of vowed language used in the New York Times anonymous Op Ed, the author is likely to be an older (58%), conservative (92%) male (66-87%). Sources should protect themselves by consulting "adverserial stylometry" and



6:49 PM - 6 Sep 2018

Opinion

I Am Part of the Resistance Inside the Trump Administration

I work for the president but like-minded colleagues and I have vowed to thwart parts of his agenda and his worst inclinations.

Robert Lee @downtownrob88 · Sep 6

Replying to @wikileaks

Great profiling! I expected it to be a young, black female with a girlfriend and membership in the Communist Party of America. I'm shocked, Sherlock.

5

(

17

16

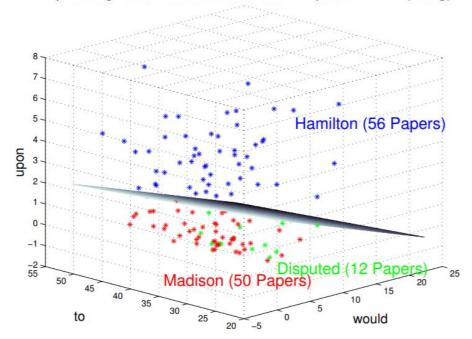
C

40

3 more replies

12 disputed Federalists papers

Separating Plane for the Federalists Papers – 1788 (Fung)

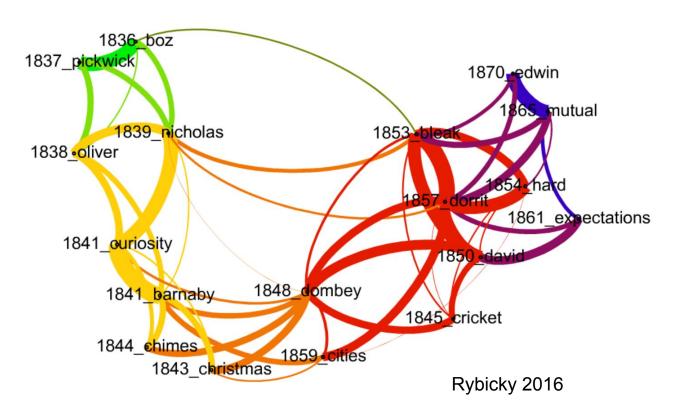


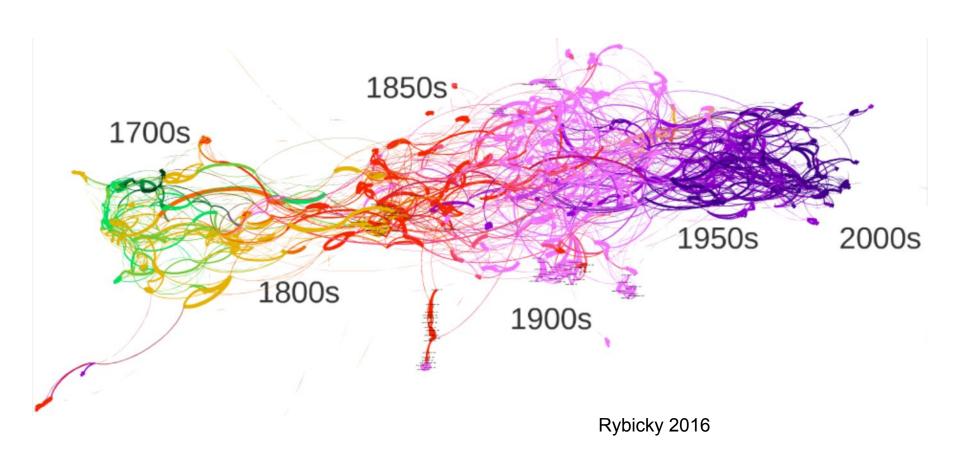
Glenn (2003). The disputed federalist papers: SVM feature selection via concave minimization

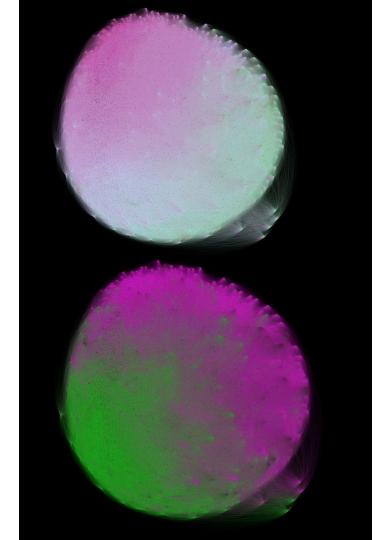
Fig. 3. Obtained Hyperplane in 3 dimensions

- Language provide enough variability for largely unconscious and unique patterns of writing
- We basically are constantly "emanating" our fingerprint and not only through natural languages
- Chat rooms (Inches et al. 2013), programming code, even texts written in artificial languages (elven, dothraki, klingon) exhibit strong-enough authorial signal to figure out probable authors
- Beware of "forensic authorial profiling"!

What lexical frequencies signal us?







Global slow changes?

- 1 m. connections of novels, based on distance measurements using most frequent words & topics (themes)
- Colors: 1) time, 2) gender
- Jockers Macroanalysis (2013)
- plots redone in 2019

Resources (corpora)

From stylo guys: https://computationalstylistics.github.io/resources/

English, Latin, Russian, Italian, Shakespeare etc.

.txtLab **Novel450** dataset: https://txtlab.org/data-sets/

150 ger, 150 fr, 150 eng