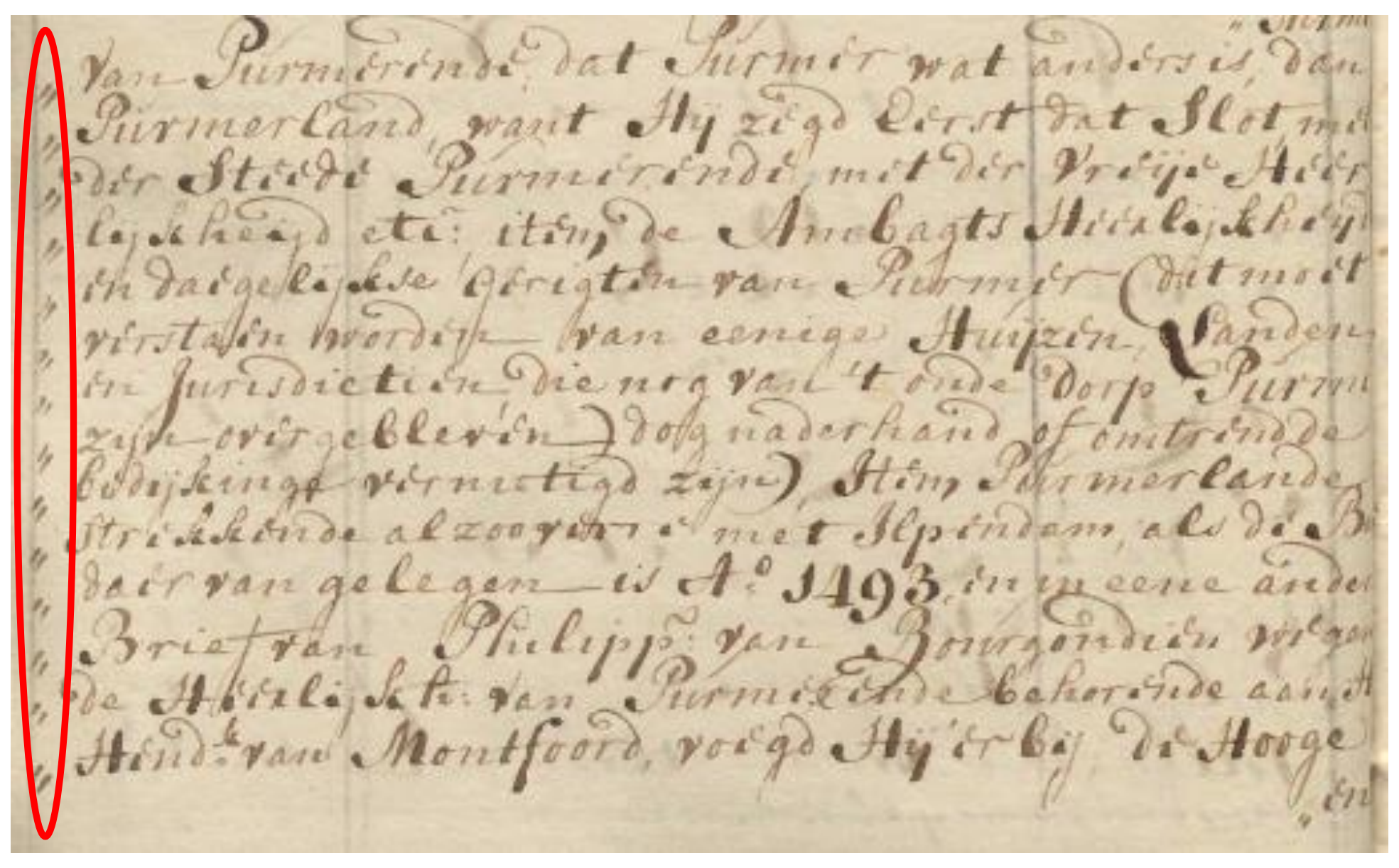
Identifying Copied Fragments in a 18th Century Dutch Chronicle

Eleanor L. T. Smith, Lianne Wilhelmus, Erika Kuijpers, Alie Lassche, Roser Morante

BACKGROUND

- Chronicles are manuscripts written to record events and phenomena which were considered important. They use a wide range of sources.
- Not much is known about the reception of news and media in the 18th century Low Countries.
- We use stylometric techniques to research the use of sources in one 18th century Dutch chronicle which exhibits use of inverted commas



van Purmerende dat Purmer wat anders is, dan
Purmerland want Hy zegd Eerst dat Slot met
der Steede Purmerende met der Vrye Heer
lykheid etc: item de Ambagts Heerlykheyd
en dagelykse Gerigten van Purmer (dit moet
verstaen worden van eenige Huyzen Landen
en Jurisdictien die nog van 't oude Dorp Purm
zyn overgebleven doch naderhand of omtrende
bedykinge vernietigd zyn) Item Purmerlande
strekkende alzoo [illegible] met Ilpendam, als de B
daer van gelegen is Aº 1493 in een eene ander
Brief van Philipps van Bourgondien wegen
de Heerlykh: van Purmerende behorende aan H
Hendk van Montfoord, voegd Hy er by de Hooge
En

Sample of the Chronicle with inverted commas circled in red

RESEARCH QUESTIONS

To what extent are computational authorship verification techniques useful to differentiate between text by the author in his own words and text copied from other sources, in an 18th century Dutch chronicle?

a) How reliable are the author's quotation marks to differentiate between copied and non copied text?

b) Is it useful to compare fragments by the author with fragments by external sources?

DATA

We define two types of chronicle fragment:

- Copy- Text found within inverted commas in the original manuscript
- No-copy- Text found outside of inverted commas in the original manuscript

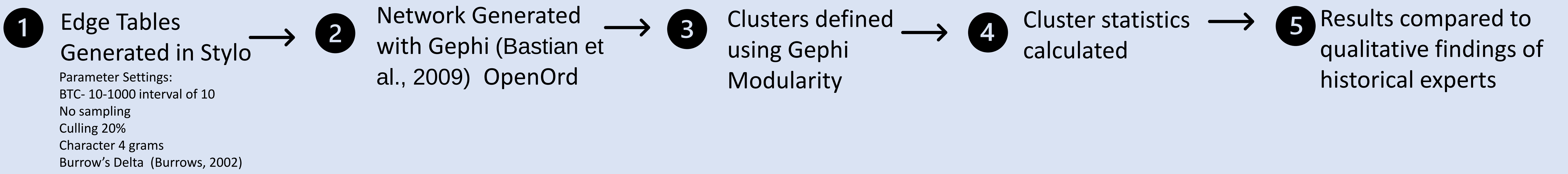
Text Type	Abbreviation	# Fragments	# Tokens
Chronicle Section 1	*Chr-1*	516	157,200
Chronicle Section 2	*chr-2*	257	171,230
Other Texts by Author	*Author*	18	4,296
Historical Sources	*source-hist*	37	88,742
Contemporary Sources	*source-contemp*	21	29,622
All	-	849	451,090

Text types and amount of material

Data Set	Chronicle Data	Author Data	Source Data
External Data 1	-	author	source-hist
Dataset1	chr-1	author	source-hist
Outliers1	chr-1 outliers	author	source-hist
External Data 2	-	author	source-contemp
Dataset2	chr-2	author	source-contemp

Composition of data sets

METHOD



1 Edge Tables Generated in Stylo →

Parameter Settings:
BTC- 10-1000 interval of 10
No sampling
Culling 20%
Character 4 grams
Burrow's Delta (Burrows, 2002)

2 Network Generated with Gephi (Bastian et al., 2009) OpenOrd →

3 Clusters defined using Gephi Modularity →

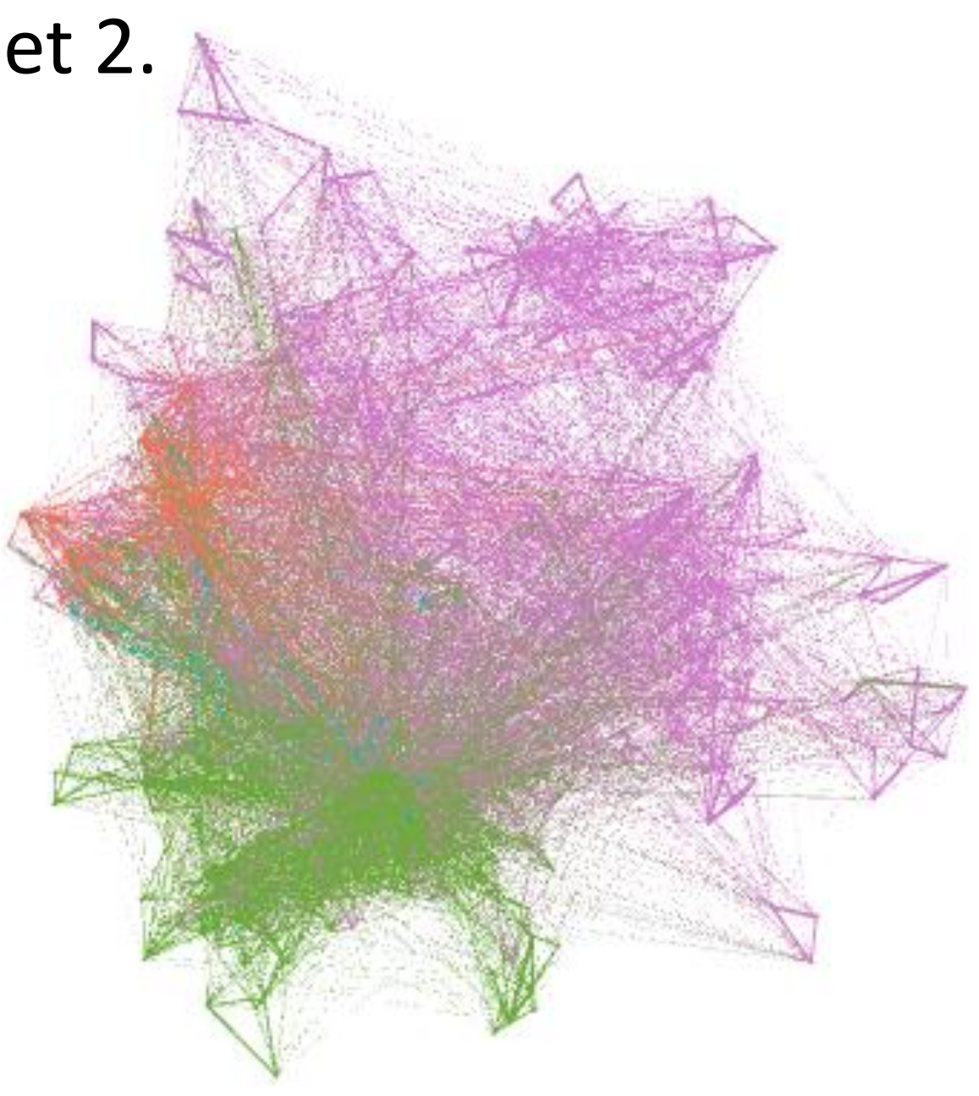
4 Cluster statistics calculated →

5 Results compared to qualitative findings of historical experts

RESULTS

Experiments with external data sets 1 & 2 showed that the method could successfully separate 100% of the data of known origin. Below are the results of the experiments with Dataset 1, Outliers 1 and Dataset 2.

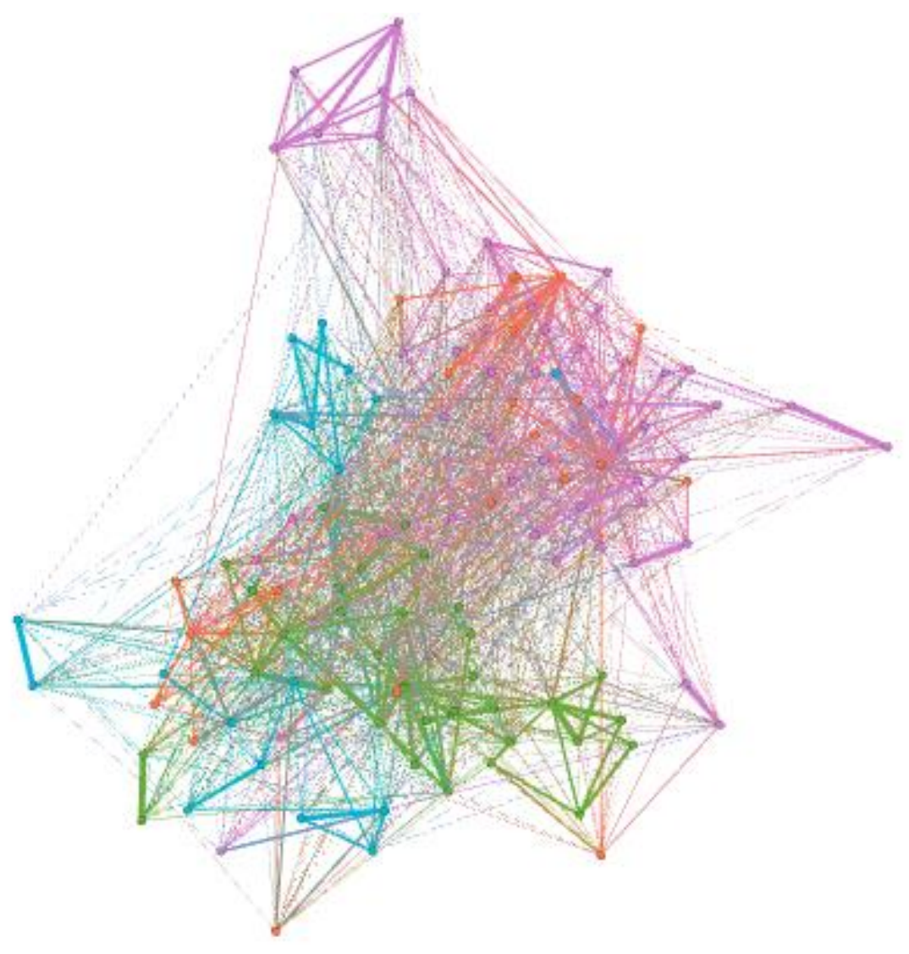
Colour	Data Type
Green	Chr-1 no-copy
Pink	Chr-1 copy
Orange	Source-hist
Blue	Author



Dataset 1

- Strong classification for 84.5% of the chronicle fragments in chr-1
- 15.5% chronicle fragments classed as outliers

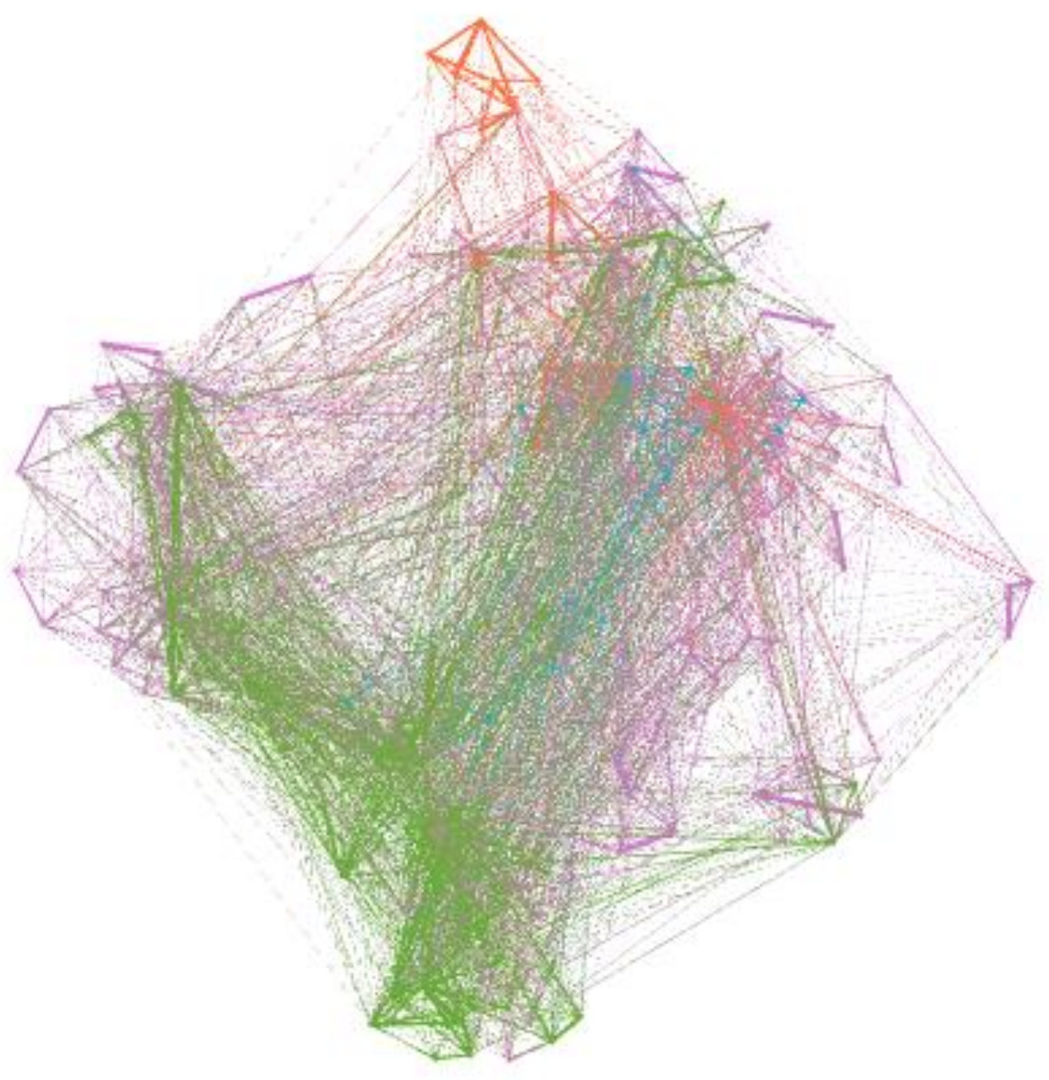
Colour	Data Type
Green	Chr-1 no-copy
Pink	Chr-1 copy
Orange	Source-hist
Blue	Author



Outliers 1

- 83.75% of the outlier fragments from Dataset 1 experiment strongly classified
- Combined with Dataset 1 results, provides strong classification for 97.5% of all chronicle fragments in chr-1

Colour	Data Type
Green	Chr-2 no-copy
Pink	Chr-2 copy
Orange	Source-contemp
Blue	Author



Dataset 2

- More mixed clusters
- Strong classification for 63.8% of chronicle fragments in chr-2

DISCUSSION

- The stylometric methods used can provide an authorship hypothesis for the majority of chronicle fragments
- The results of the Dataset 1 & 2 experiments show that the inverted commas present in the chronicle are a strong indication of source use
- The outlier experiments with Outliers 1 show the importance of collaboration with historical experts
- The presence and clustering of 'Tale Kanaäns' shows the issue that arises with author independant style
- The higher number of mixed cluster in the Dataset 2 experiment may be explained by the difference in function of chr-1 and chr-2
  - Chr-1: an overview of the history of the local area. Knowledge of these events comes from history books, as the author was not alive when they took place.
  - Chr-2: covers contemporary life and political events in the Netherlands, interweaving the author's experiences
    with quotes from news media and publications by political institutions and authorities.

FUTURE WORK

- Testing outliers from Dataset 2 experiment
- Test method on a chronicle which does not use inverted commas

Eleanor.Smith@uantwerpen.be

https://chroniclingnovelty.com/