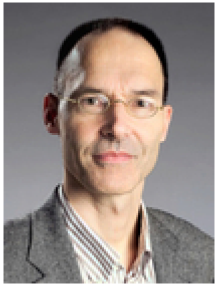
A new Resource for Analyzing Collaborative Writing Styles and One-Sidedness

Scientific Authorship and Peer Review: Between a Means of Governance and Structural Meaning?

01. – 02.09.2022, HU Berlin

Erik
Körner

Harry
Scells



Benno
Stein

Bauhaus-
Universität
Weimar

Webis

Outline

① Multi-Authorship Identification

② Collaborative Writing Styles

③ SMAuC - The Scientific Multi-Authorship Corpus

④ Researching Algorithmic Bias

Multi-Authorship Identification

Introduction

- Multi-Authorship Identification/Analysis an important variant of the vanilla (Single) Authorship Identification problem
  - Single-Author: “Who is the author?” of a letter, an article, or a book
  - Multi-Authorship Identification: questions and issues about documents written by a group of authors

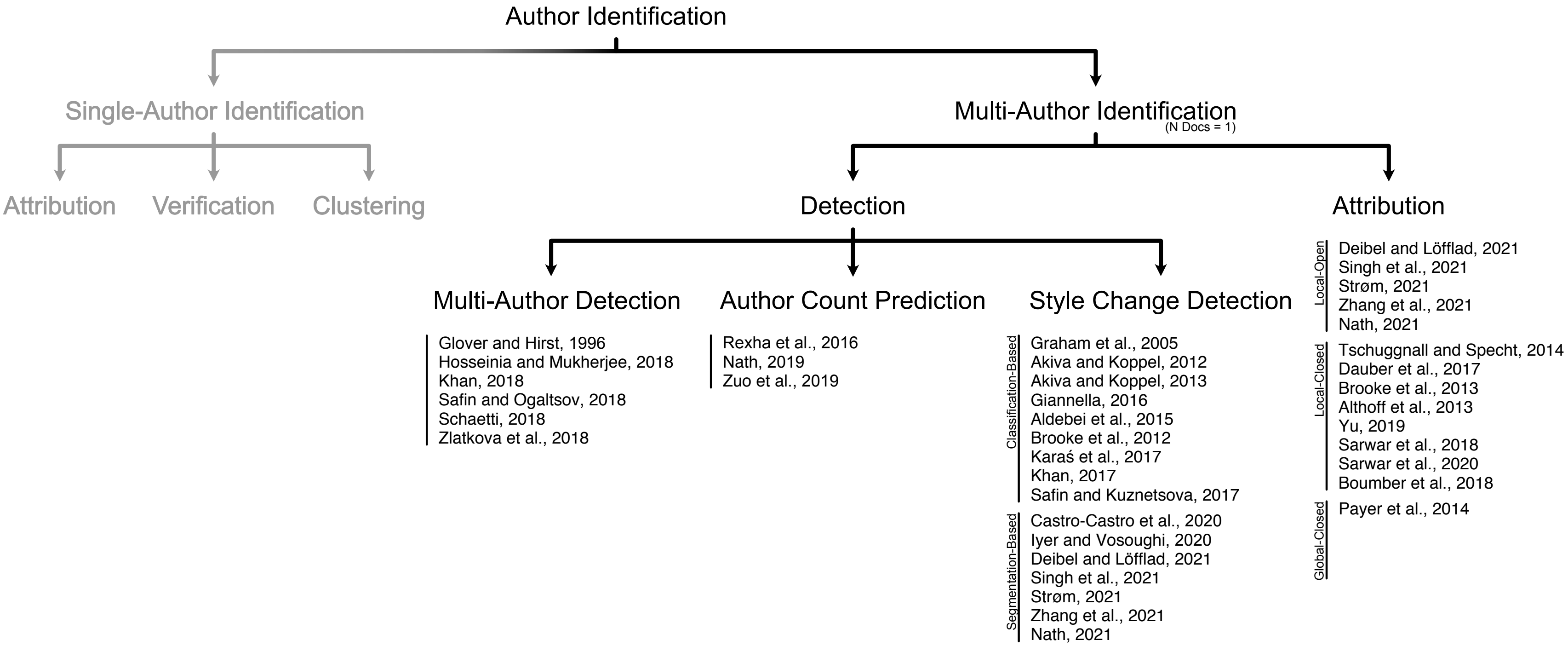
- Authorship in academia ➜ often multiple authors
  - Intentional, collaborative writing
  - Text reuse, plagiarism, . . .

- Increased attention and application of Multi-Authorship Identification
  - Numerous tasks, datasets and methods over the years
  - PAN, various shared task and datasets

  - However, style of collaborative writing mostly the same
  - Very little (public) academic datasets, affects and hinders comparability of approaches against each other

Multi-Authorship Identification

Author Identification Tasks in Literature



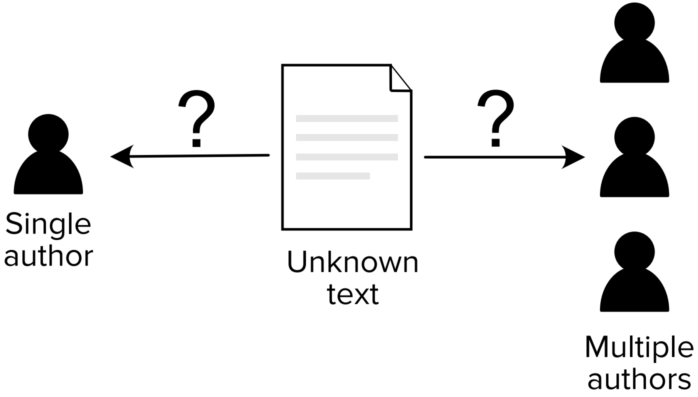
Author Identification

- Single-Author Identification
  - Attribution
  - Verification
  - Clustering
- Multi-Author Identification (N Docs = 1)
  - Detection
    - Multi-Author Detection
      - Glover and Hirst, 1996
      - Hosseinia and Mukherjee, 2018
      - Khan, 2018
      - Safin and Ogaltsov, 2018
      - Schaetti, 2018
      - Zlatkova et al., 2018
    - Author Count Prediction
      - Rexha et al., 2016
      - Nath, 2019
      - Zuo et al., 2019
    - Style Change Detection
      - Classification-Based
        - Graham et al., 2005
        - Akiva and Koppel, 2012
        - Akiva and Koppel, 2013
        - Giannella, 2016
        - Aldebei et al., 2015
        - Brooke et al., 2012
        - Karaś et al., 2017
        - Khan, 2017
        - Safin and Kuznetsova, 2017
      - Segmentation-Based
        - Castro-Castro et al., 2020
        - Iyer and Vosoughi, 2020
        - Deibel and Löfflad, 2021
        - Singh et al., 2021
        - Strøm, 2021
        - Zhang et al., 2021
        - Nath, 2021
  - Attribution
    - Local-Open
      - Deibel and Löfflad, 2021
      - Singh et al., 2021
      - Strøm, 2021
      - Zhang et al., 2021
      - Nath, 2021
    - Local-Closed
      - Tschuggnall and Specht, 2014
      - Dauber et al., 2017
      - Brooke et al., 2013
      - Althoff et al., 2013
      - Yu, 2019
      - Sarwar et al., 2018
      - Sarwar et al., 2020
      - Boumber et al., 2018
    - Global-Closed
      - Payer et al., 2014

- ❑ Single-Author: [Thomas Corvin Mendenhall 1887]
- ❑ Multi-Author: [Glover and Hirst 1996]

Multi-Authorship Identification

Multi-Author Detection

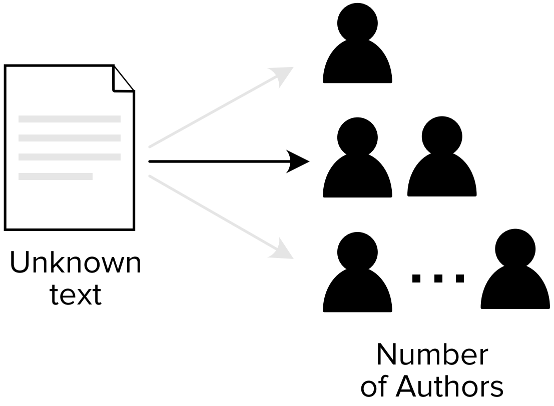


- **Task**: Single author or multiple authors?

- Only very few studies that solely address this problem
- Often as consequence or reduction of more complex result, e.g. author count, style changes
- Many datasets with assumption that texts are multi-authored and then just application of more ‘sophisticated’ methods

Multi-Authorship Identification

Author Count Prediction

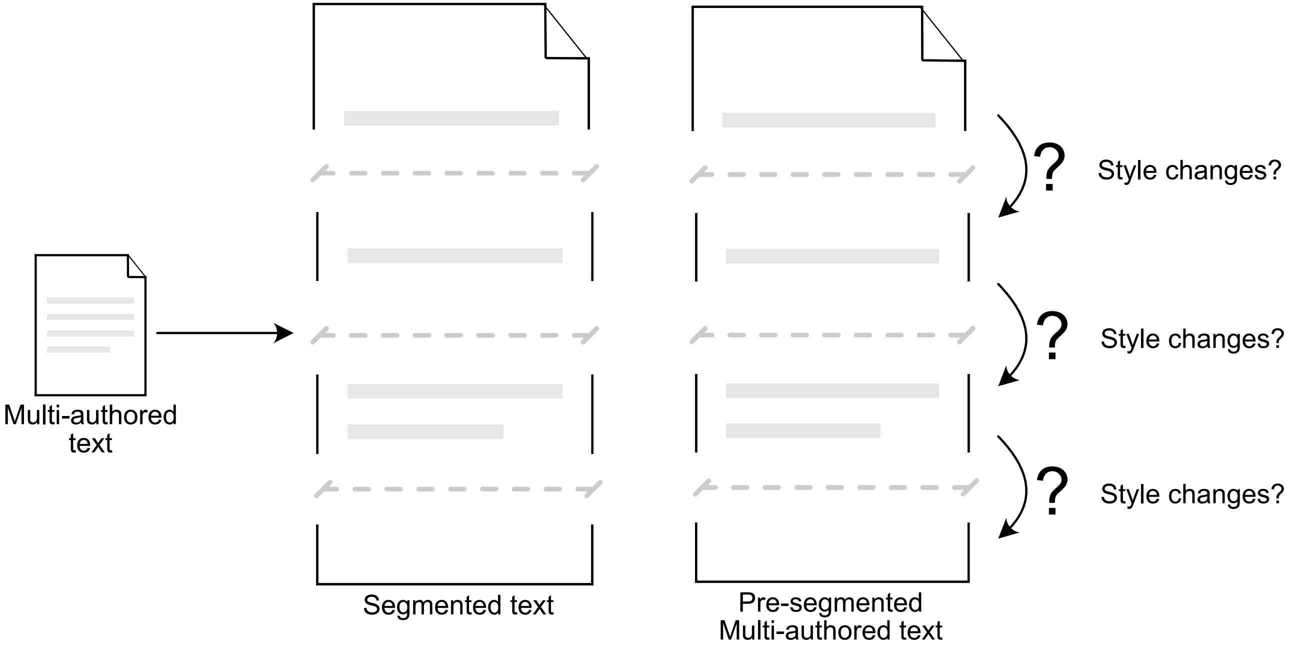


- ❑ **Task**: Number of authors?

- ❑ Fundamental multi-author identification task
- ❑ Application not limited to human-readable texts, e.g. compiled binary software

Multi-Authorship Identification

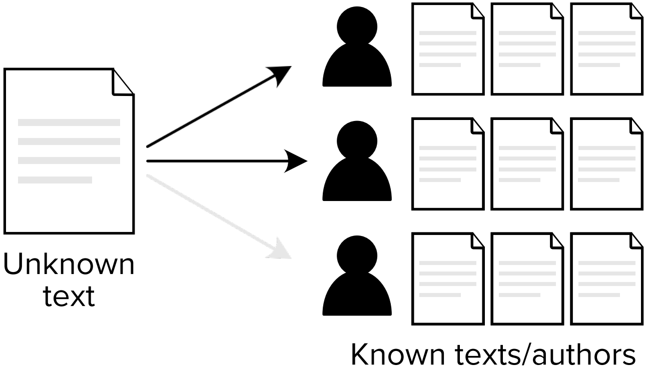
Style Change Detection



- ❑ **Task**: Identify boundary where style of text changes.
- ❑ Sub-tasks that require to first segment the text vs. pre-segmented texts

Multi-Authorship Identification

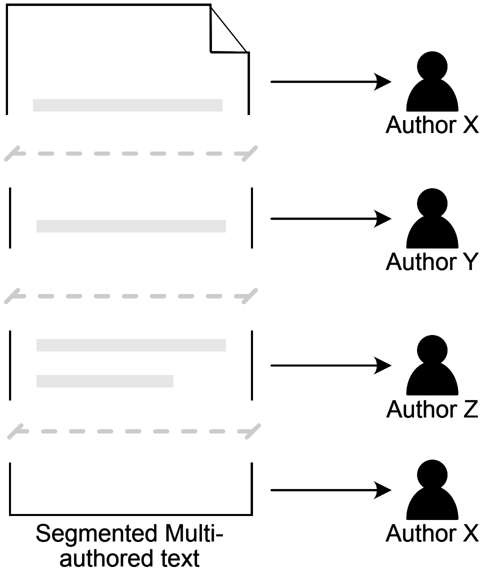
Multi-Author Attribution

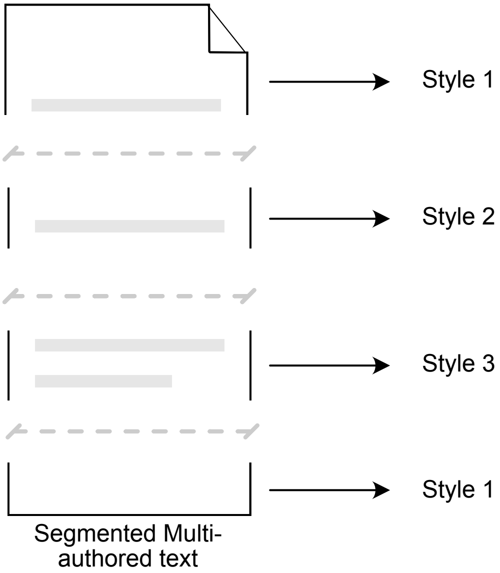


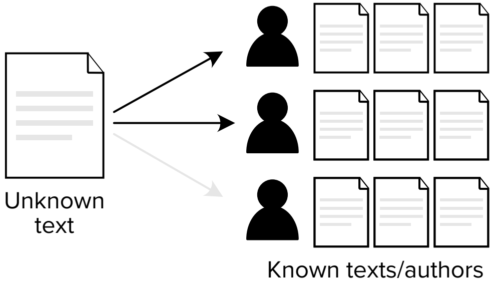
- **Task**: Attribution of text segments
- Sub-tasks:
  - Authors: closed-set vs. open-set
  - Document: local vs. global

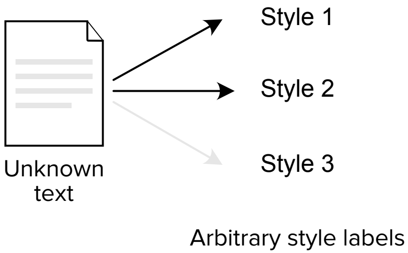
Multi-Authorship Identification

Multi-Author Attribution Sub-Tasks









Collaborative Writing Styles

Introduction

- At PAN, multi-authorship identification datasets have been constructed so far by combining texts that are written by single authors into a single, multi-authored text.
- *Multi-Author Attribution* research often only focuses on the metadata of the text, e.g. author list of journal articles.
- But *Collaborative Writing Styles* are not really taken into account when developing methods to address *Multi-Author Identification*.

- What are the different types a text can be written collaboratively?
- Where does research (currently) happen?

Collaborative Writing Styles

Types of Writing Styles

❑ Sequential

❑ Group Single

❑ Horizontal Division

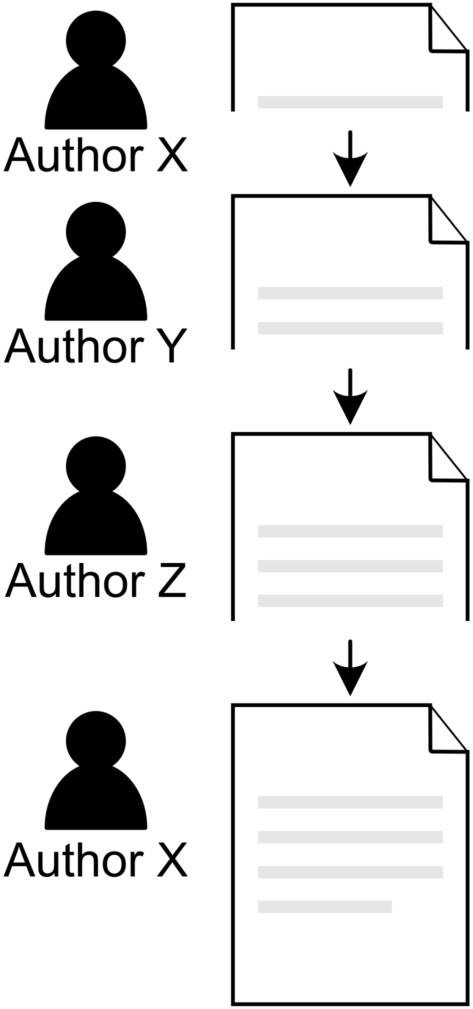
❑ Stratified Division

❑ Reactive

Building a Taxonomy and Nomenclature of Collaborative Writing to Improve Interdisciplinary Research and Practice [Lowry et al. 2004]

Collaborative Writing Styles

Sequential Writing Style



Collaborative Writing Styles

Sequential Writing Style

Characteristics

- Each author writes a section of the text, sequentially, independently
- Boundaries of authorial style explicitly defined, co-authors are not allowed to edit outside of their section of text

Examples

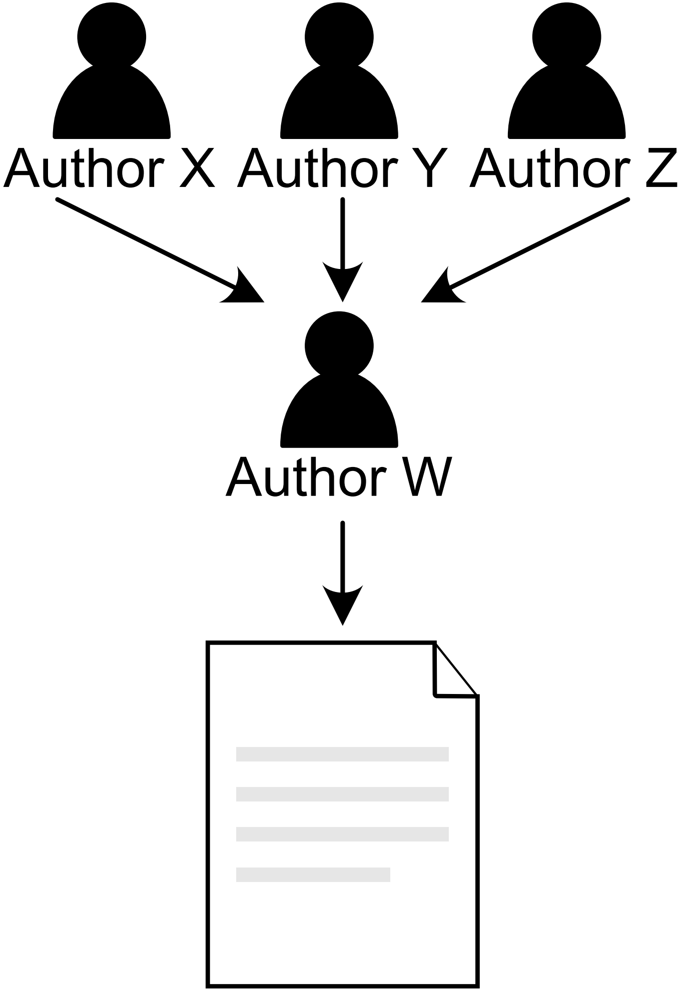
- Collaboration of a PhD student and supervisor on a research paper; supervisor writing the introduction and conclusion, student the content in between

Tasks

- Multi-Author Detection, Style Change Detection, Multi-Author Attribution, . . .

Collaborative Writing Styles

Group Single Writing Style



Collaborative Writing Styles

Group Single Writing Style

Characteristics

- Several authors contribution to the ideation phrase of writing
- Single author compiles these into a single text
- *Consistent authorship style*, yet involvement of multiple authors in creation

Examples

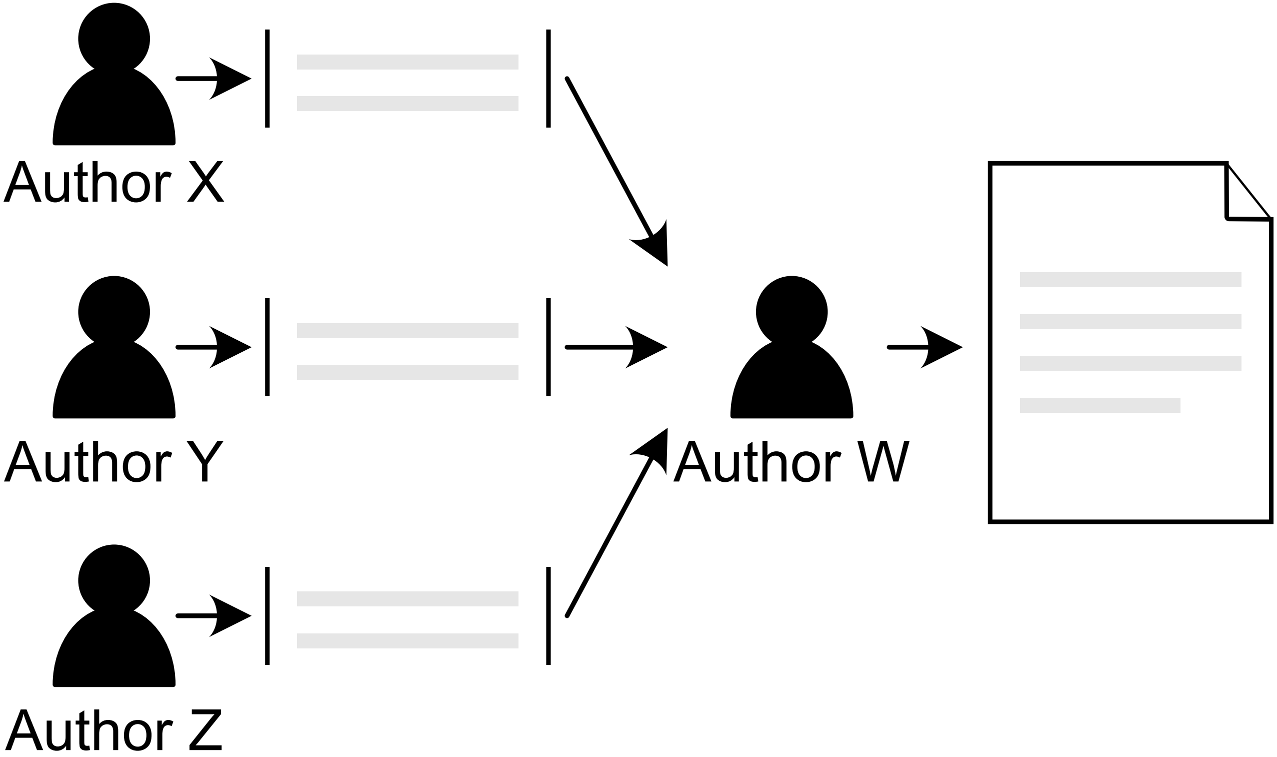
- Grant writing: many principal investigators or collaborators involved in ideation, chief investigator writes proposal document

→ Multi-Authorship Identification methods may not be applicable?

- Are there style boundaries for Style Change Detection?
- Multi-Author Detection may be possible?

Collaborative Writing Styles

Horizontal Division Writing Style



Collaborative Writing Styles

Horizontal Division Writing Style

Characteristics

- Several authors contribute ‘sub-documents’
- Single authors compiles these into a single text
- Compiled text may contain authorship styles of co-authors, depending on the amount of editing applied

Examples

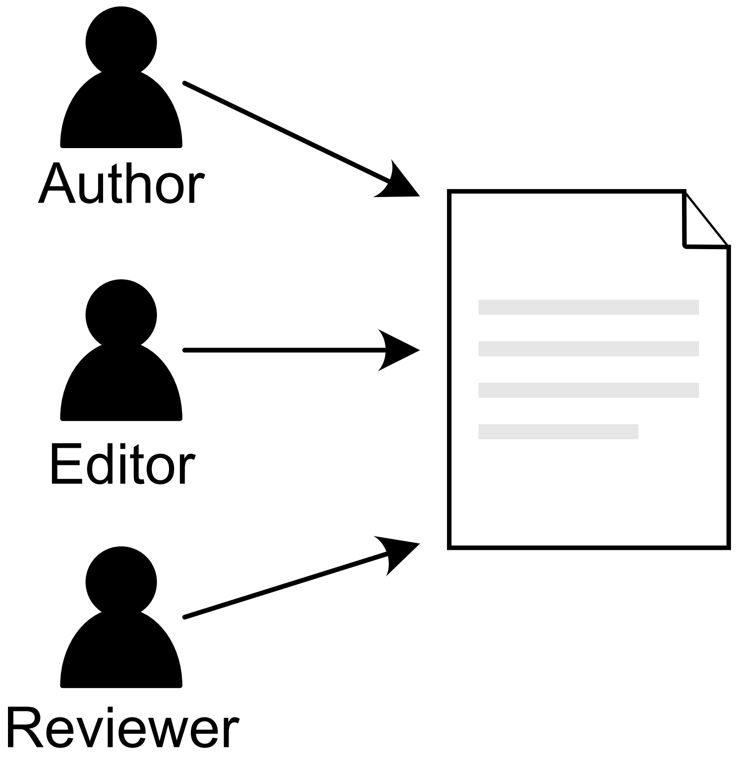
- Academic book: several academics write different chapters, an editor combines them into a cohesive manuscript
- Text Reuse

Notes

- Mainly targets *Style Change Detection* task
- Easiest and most obvious way to create artificial datasets

Collaborative Writing Styles

Stratified Division Writing Style



Collaborative Writing Styles

Stratified Division Writing Style

Characteristics

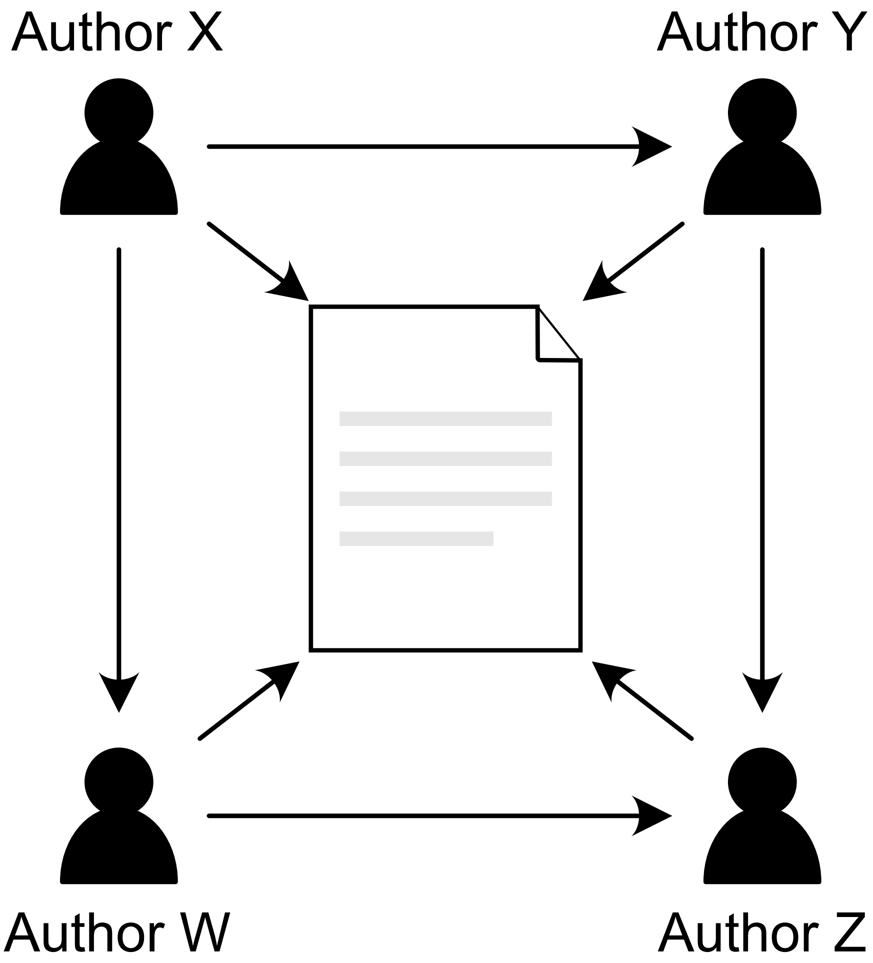
- Similar to *Horizontal Division*
- Each co-author plays a certain role in the creation of a text, e.g. author, editor, reviewer

Examples

- Scholarly article: one author writes majority of text, another author edits the text, an independent reviewer provides critical feedback that feeds back into the creation process

Collaborative Writing Styles

Reactive Writing Style



Collaborative Writing Styles

Reactive Writing Style

Characteristics

- Authors write synchronously on the same text while adjusting the writing of others

Examples

- Several undergraduate students in a group assignment writing a report together
- Collaborative writing platforms, e.g. Overleaf, Etherpad, Google Docs

Notes

- Blurred authorial style boundaries
- Most complex in terms of developing Multi-Author Identification methods

Collaborative Writing Styles

Observations

- Different writing styles may be *easier* or *harder* to apply Multi-Author Identification methods to
- Boundaries are more clearly defined in *Horizontal Division* compared to *Reactive Writing Style*
- Some multi-authorship approaches are impossible to apply, e.g. *Style Change Detection* to *Reactive Writing Style*
- In literature most datasets for Multi-Author Identification are created using *Horizontal Division*!

Collaborative Writing Styles

Observations

- Different writing styles may be *easier* or *harder* to apply Multi-Author Identification methods to
- Boundaries are more clearly defined in *Horizontal Division* compared to *Reactive Writing Style*
- Some multi-authorship approaches are impossible to apply, e.g. *Style Change Detection* to *Reactive Writing Style*
- In literature most datasets for Multi-Author Identification are created using *Horizontal Division*!

➜ Existing methods may not be robust against different Collaborative Writing Styles.

➜ The way in which multi-authored texts are created is fundamental to which tasks are applicable and to the difficulty in applying methods to those tasks.

Collaborative Writing Styles

Overview over Datasets and Methods

Dataset		Dataset	Dataset Source	**Collaborative**	#Docs &	Users
		Task		**Writing Style**	Tr/Va/Te Splits	
MULTI-AUTHOR DETECTION (MAD)						
[Glover et al. 1996]		AV	film summaries	HD	20	***self***
MULTI-AUTHOR ATTRIBUTION (MAA)						
PAN12 AA [Patrick Juola 2012]	🔗	AA/AC	Feedbooks	HD	170	2
[Brooke 2013]		AC	*The Waste Land*, poems	HD	21	***self***
[Althoff et al. 2013]		AA	arXiv	HD, S	594	***self***
[Tschuggnall et al. 2014]		MAD	Gutenberg/FED	HD	75	***self***
[Payer et al. 2014]		AA	conference papers	S	3,516/-/378	***self***
[Dauber et al. 2017]		AA	Wookiepedia	R	-	***self***
[Sarwar et al. 2018]		MAA	Gutenberg/arXiv	HD, S	6,173	***self*** +1
MLPA-400 [Boumber et al. 2018]	🔗	MAA	ML papers	S	400	***self***
[Brian Yu 2019]		MAA	Gutenberg	HD	-	***self***
AUTHOR COUNT PREDICTION (ACP)						
[Rexha et al. 2016]		ACP	PubMed	S	6,144	***self***
[Alrabaee et al. 2019]		ACP	open-source code	HD	31,150	***self***
PAN19 SCD [Zangerle et al. 2019]	🔗	ACP	StackExchange	HD	2,546/1,272/1,210	PAN: 2
STYLE-CHANGE DETECTION (SCD)						
[Graham et al. 2005]		SCD	Usenet	HD	-	***self***
[Brooke et al. 2012]		SCD	*The Waste Land*, poems	HD	51	***self***
[Akiva et al. 2012]		SCD/AC	Biblical/Blogs/NYT	HD	14	***self***
[Akiva et al. 2013]		SCD/AC	Biblical/Blogs/NYT	HD	-	***self*** +2
PAN16 AD [E. Stamatatos 2016]		AD	Webis-TRC-12	HD	174/-/8	PAN: 2
PAN17 SCD [Tschuggnall et al. 2017]	🔗	SCD	Webis-TRC-12	HD	187/-/99	PAN: 3
PAN18 SCD [Kestemont et al. 2018]	🔗	MAD	StackExchange	HD	2,980/1,492/1,352	PAN: 5
PAN20 SCD [Zangerle et al. 2020]	🔗	SCD	StackExchange	HD	11,448/5,732/5,696	PAN: 2
PAN21 SCD [Zangerle et al. 2021]	🔗	SCD	StackExchange	HD	11,200/2,400/2,400	PAN: 5

HD: Horizontal Division (randomly combining text fragments from different authors), *R*: Reactive,
S: scientific papers (combination of *Group-single*, *Stratified Division*, *Reactive*; no stylistic ‘editing’ by dataset creators)

SMAuC - The Scientific Multi-Authorship Corpus

Motivation

- *Scientific writing* as a new and interesting domain for authorship analysis, especially for *Multi-Authorship Analysis*
- Most datasets lack material from science domain or required metadata
- Research often only with small *unpublished* datasets using arXiv preprints, PubMed articles or journal papers
  ➜ Reproduction and comparability difficult due to varying approaches for data proprocessing and dataset curation
- Very few publication that publish their research, e.g. MLPA-400 [Boumber et al. 2018]

➜ Requirement for large, openly accessible dataset of scientific works

SMAuC - The Scientific Multi-Authorship Corpus

Motivation

- *Scientific writing* as a new and interesting domain for authorship analysis, especially for *Multi-Authorship Analysis*
- Most datasets lack material from science domain or required metadata
- Research often only with small *unpublished* datasets using arXiv preprints, PubMed articles or journal papers
  ➜ Reproduction and comparability difficult due to varying approaches for data proprocessing and dataset curation
- Very few publication that publish their research,
  e.g. MLPA-400 [Boumber et al. 2018]

➜ Requirement for large, openly accessible dataset of scientific works

SMAuC - The Scientific Multi-Authorship Corpus 🐲

SMAuC - The Scientific Multi-Authorship Corpus

Dataset Sources

- ❑ CORE database [Knoth et al. 2011] [Knoth and Zdrahal 2012]
  - *Collection of metadata and full texts of open access scientific publications*
  - Dump from 2018-03-01[1]
  - 123M metadata items, 85.6M items w/ abstracts, 9.8M items w/ **full texts**

- ❑ Microsoft Open Academic Graph (OAG) [Sinha et al. 2015]
  - *Openly accessible heterogeneous knowledge graph based on scientific articles, authors, and institutions*
  - Source for identifying and disambiguating authors and fields of study
  - Version 2 of the OAG [Hu et al. 2020][2]
  - 179M nodes, 2B edges

[1] https://core.ac.uk/services/dataset

[2] https://www.microsoft.com/en-us/research/project/open-academic-graph/

SMAuC - The Scientific Multi-Authorship Corpus

Dataset Curation Process

Conditions applied	Number of documents	
CORE	123,988,821	*(100.00%)*
↪ full texts	9,835,064	*(7.93%)*
↪ text language filtering	6,531,442	*(5.27%)*
↪ OAG matching	3,508,509	*(2.82%)*
↪ text quality assurance	**3,356,686**	*(2.70%)*

- High requirements on data quality
  - Multi-step language filtering with fastText
  - Improved mapping of full texts and OAG metadata using DOIs and titles
  - Manual mapping of heterogenous OAG *field of study* ➜ *DFG Classification of Scientific Disciplines and Research Areas* [DFG 2016]
  - Removal of markup, non-ASCII characters; lowercasing, collapsing whitespaces
  - Additional (heuristical) filtering for text quality, e.g. text length, language

SMAuC - The Scientific Multi-Authorship Corpus

Counts for all types of documents and their total

Document Type	Count
Single author w/o multi author	711,471
Single author w/ multi author	261,629
Multi author w/o single author	1,481,106
Multi author w/ single author	894,945
No author information	7,535
Total	3,356,686

SMAuC - The Scientific Multi-Authorship Corpus

Number of documents in the corpus by text length in characters and document type with percentage per row

Length	Total	Single author		Multi author	
≤ 3,000	39,300	13,680	(1.41%)	25,567	(1.07%)
– 5,000	96,067	32,059	(3.29%)	63,832	(2.69%)
– 50,000	2,273,246	467,844	(48.07%)	1,799,435	(75.73%)
– 250,000	771,756	301,975	(31.03%)	468,473	(19.72%)
> 250,000	176,317	157,542	(16.19%)	18,744	(0.79%)
Total	3,356,686	973,100	(100.00%)	2,376,051	(100.00%)

SMAuC - The Scientific Multi-Authorship Corpus

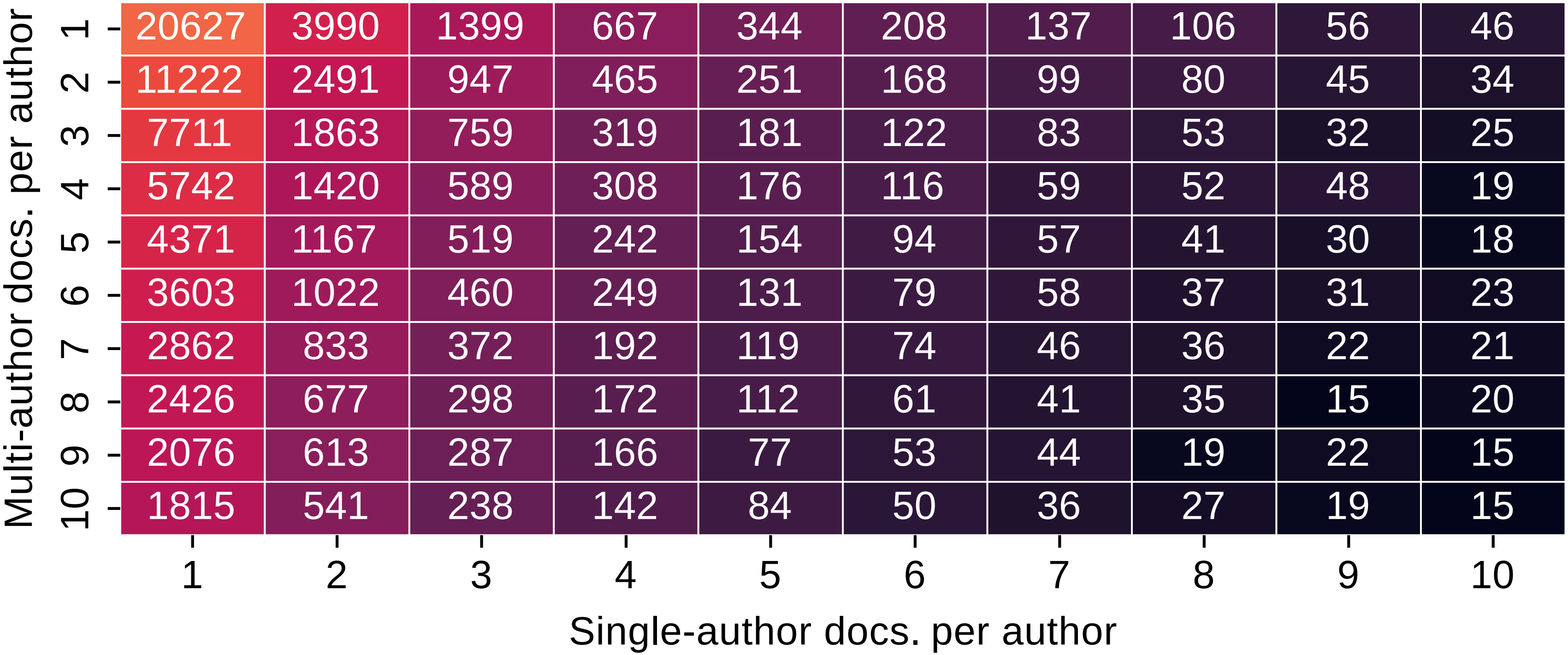
Document counts by research area [DFG 2016]

Research Area	**SA**	**MA**	**A**	**TL**
Engineering Sciences	55,015	375,206	3	28,467
Humanities	58,317	199,926	3	37,224
Life Sciences	48,723	715,218	5	32,616
Natural Sciences	147,024	651,076	3	26,103

Single author documents (**SA**), multi author documents (**MA**), median authors per document (**A**) and median text length (**TL**).

SMAuC - The Scientific Multi-Authorship Corpus

Total author count over the number of single-author and multi-author publications per author



Multi-author docs. per author \ Single-author docs. per author	1	2	3	4	5	6	7	8	9	10
1	20627	3990	1399	667	344	208	137	106	56	46
2	11222	2491	947	465	251	168	99	80	45	34
3	7711	1863	759	319	181	122	83	53	32	25
4	5742	1420	589	308	176	116	59	52	48	19
5	4371	1167	519	242	154	94	57	41	30	18
6	3603	1022	460	249	131	79	58	37	31	23
7	2862	833	372	192	119	74	46	36	22	21
8	2426	677	298	172	112	61	41	35	15	20
9	2076	613	287	166	77	53	44	19	22	15
10	1815	541	238	142	84	50	36	27	19	15

SMAuC - The Scientific Multi-Authorship Corpus

Publication and Access

- Features
  - Full-text extracts, annotated with author metadata
  - Publications from different scientific domains, stylistically diverse texts
  - Monographs and multi-authored documents

- Paper currently under review
  SMAuC - The Scientific Multi-Authorship Corpus

- Dataset will be made accessible via Zenodo, restricted to academia

- Ongoing experiments in context of **multi-authorship** and **algorithmic bias**

Researching Algorithmic Bias

Motivation

Background

- Increasing reliance on *machine learning* processes in various domains, esp.
  - Plagiarism Detection,
  - Authorship Attribution of scientific research,
  - Digital Text Forensics.

Problem

- Detection of Plagiarism or Authorship Attribution may perform worse or fail for (a) one **gender** compared to another, or (b) non-**native speakers** compared to native speakers e.g. in court decisions, job assessment, etc.
- . . .

➜ Unfair advantages, faulty predictions, monetary loss, etc. due to ML model bias

Researching Algorithmic Bias

Focus

- Scientific domain / academia
- Algorithmic bias

Types

- Native Speakers (English)
- Gender

Data

- SMAuC - The Scientific Multi-Authorship Corpus

Researching Algorithmic Bias

Work in Progress

- Manually annotating *gender* and *native language* for authors in SMAuC
- Prototype using *Generalized Unmasking* [Koppel and Schler 2004] [Bevendorff et al. 2019]

Future Plans

- Creating experiment framework to easily substitute different algorithms and datasets/autorship tasks

Researching Algorithmic Bias

Work in Progress

- Manually annotating *gender* and *native language* for authors in SMAuC
- Prototype using *Generalized Unmasking* [Koppel and Schler 2004] [Bevendorff et al. 2019]

Future Plans

- Creating experiment framework to easily substitute different algorithms and datasets/autorship tasks

Thank you for your attention!