Patterns of bibliographic diversity in post-2022 publications on "LLMs" and "ChatGPT"

Amin Gino Fabbrucci Barbagli[1] Evan Piepho[2] Moses Boudourides[3]

University of Trieste, Italy
amingino.fabbruccibarbagli@phd.units.it

Arizona State University,USA

Northwestern University, USA
moses.boudourides@northwestern.edu

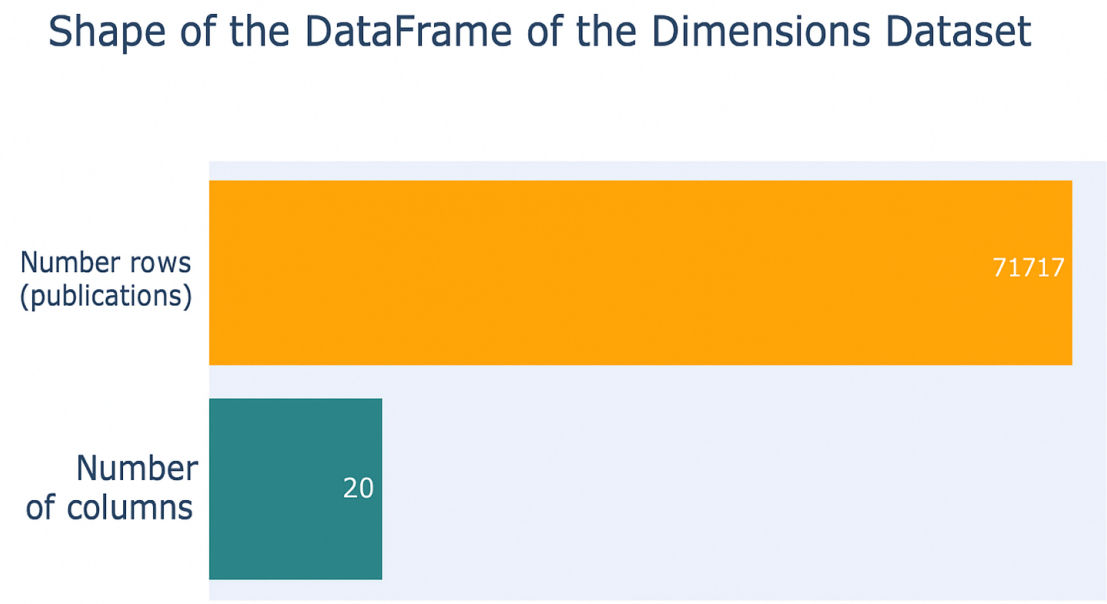
June 25, 2025

Why Choose Dimensions.ai?

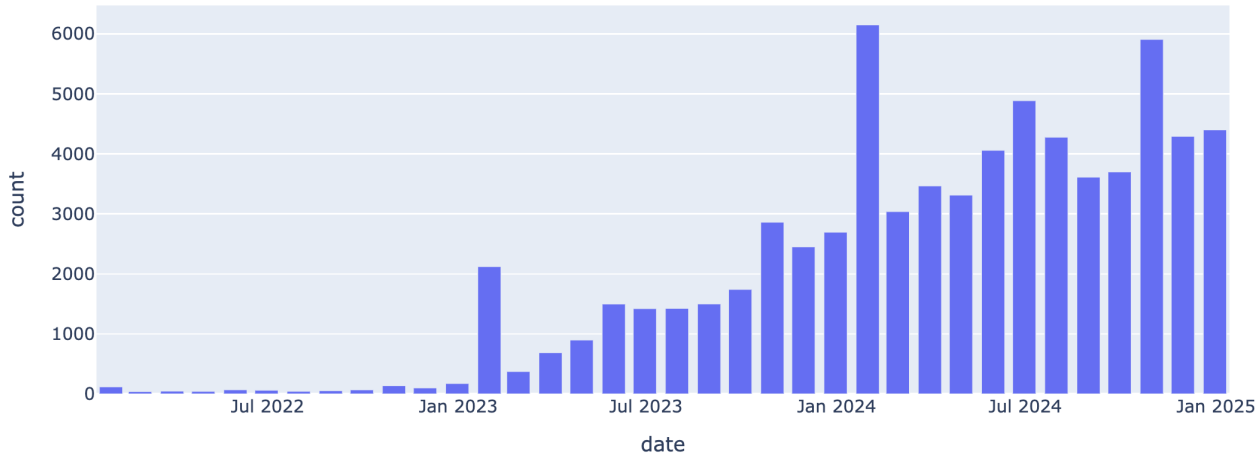
- **Why Dimensions over Scopus or Web of Science?**
  - (i) High completeness and quality of publication metadata (Delgado-Quir´os, 2024; Nguyen et al.,2022).
  - (ii) Availability of a Python client `dimcli` for efficient, programmatic querying.
  - `https://api-lab.dimensions.ai/cookbooks/1-getting-started/5-Deep-dive-DSL-language.html`

- **Query Used:** *%dslloopdf* `search publications in title, abstract only for "chatgpt" or "large language model" or "LLM" where year = 2022 - end of 2024 return publications`
- **Publication Fields (subset of 20 analyzed):**
  - `id`, `authors`, `title`, `date`, `doi`, `type`, `category_for`, `grants`, `funding_usd`
  - Full list and definitions: `https://docs.dimensions.ai/dsl/2.0.0/datasource-publications.html`

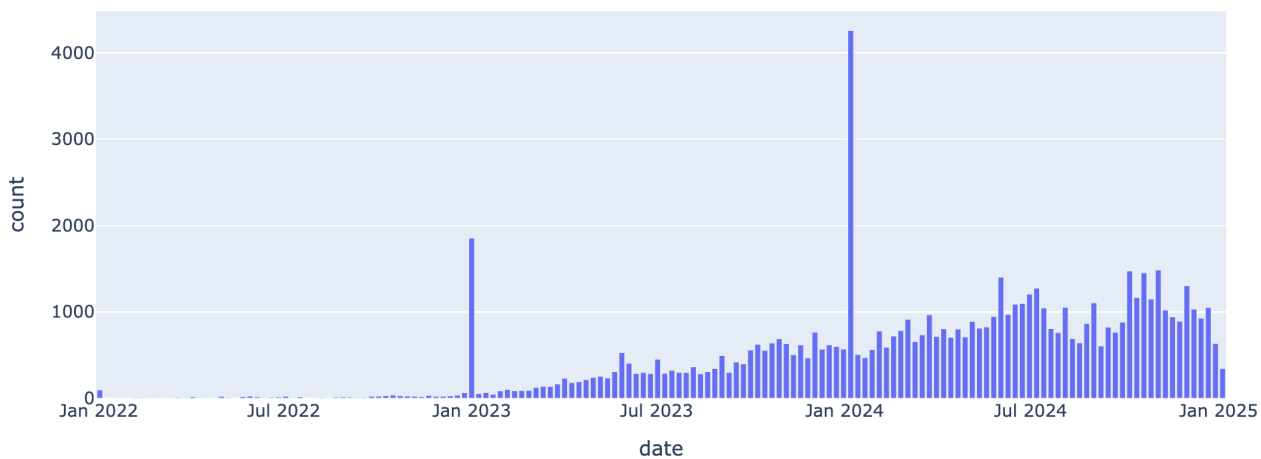
After removing duplicate publications, the unique count of each publication in the collected Dimensions dataset was established by enumerating the distinct id field associated with each publication. It is important to note that if a publication was retrieved under two or more different types—for example, as both an article (or proceeding or chapter) and a preprint—the preprint version was excluded.



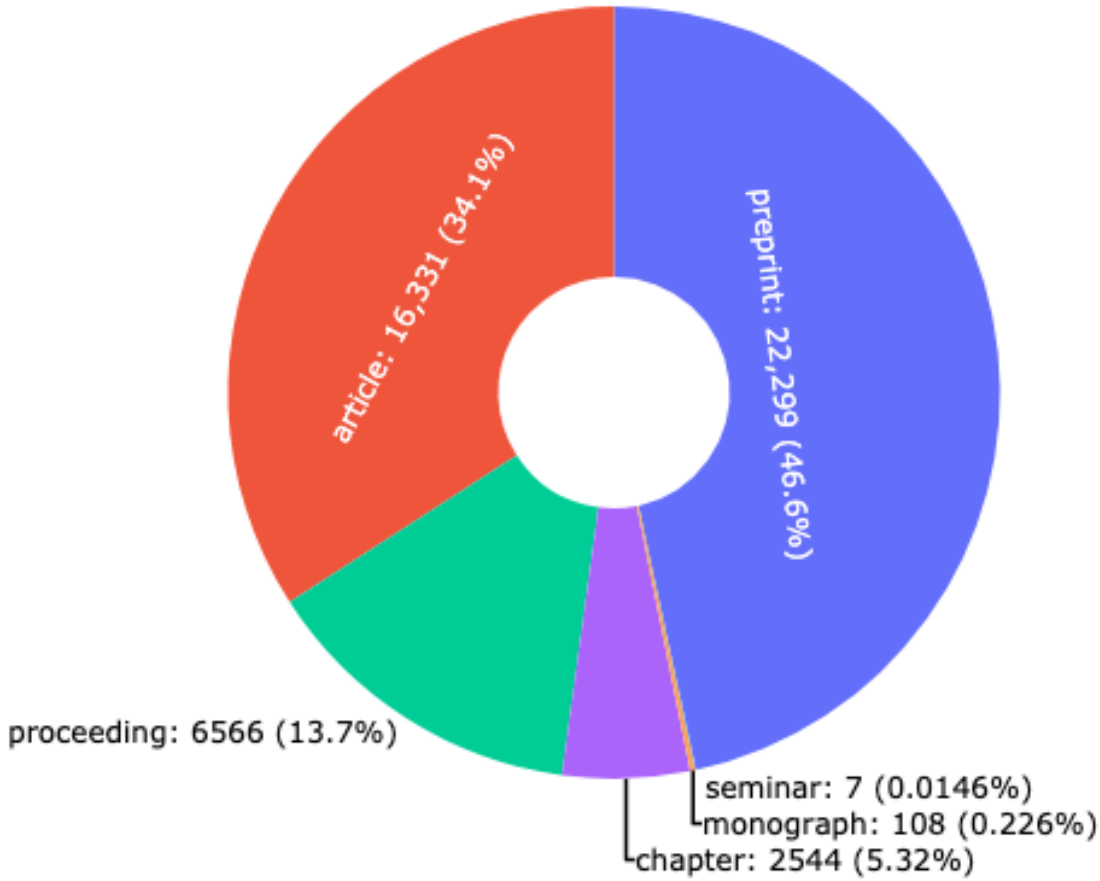
Publications per month



Publications per week



Distribution of Types of Publications (Total Number of Publications = 47,855)



To identify authors' gender, we used Namsor, an algorithmic model for the classification of names that contains a repository of 7.5 billion names, including those from 142 ethnicities, 249 countries, and 22 alphabets (https://namsor.app/about-us/). Namsor's model recognizes morphemes—the smallest units of construction within languages that help comprise words—to incorporate patterns in naming conventions when assigning a name's gender, ethnic origin, and other elements offered through their service. The accuracy of Namsor's model has been verified by multiple studies and audits, including a 2018 Science-Metrix publication that found it correctly classified the gender of Olympic medalists' names from 25 countries to within 98-99% accuracy.

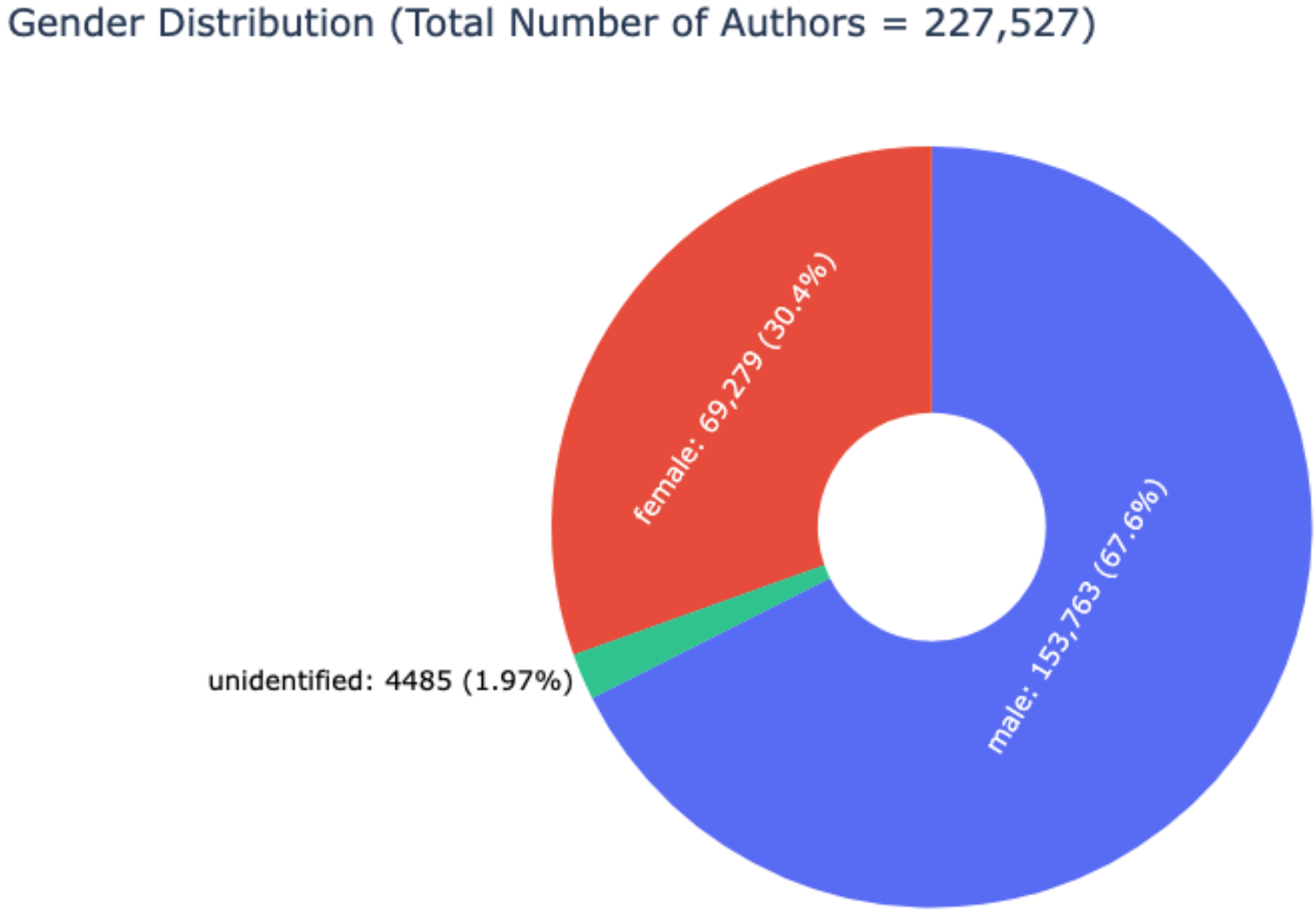
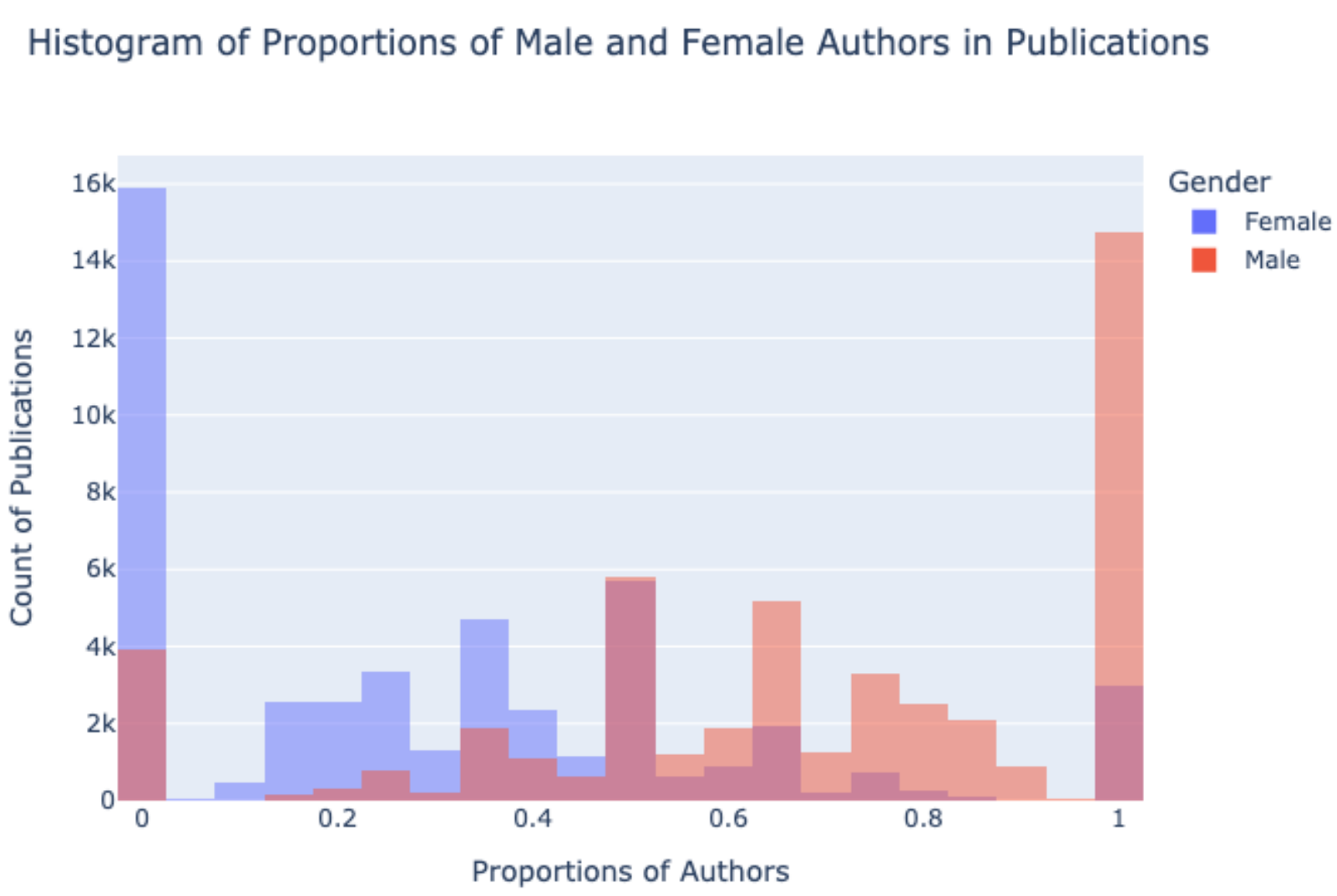
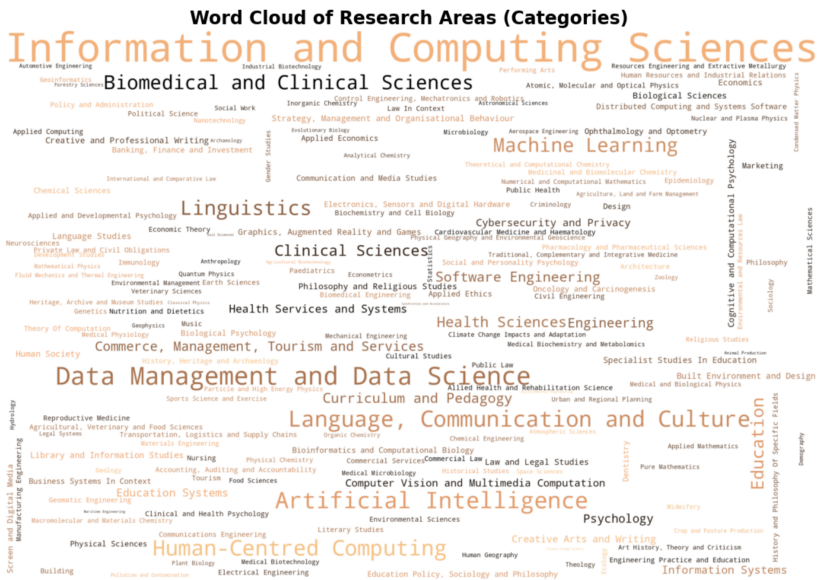


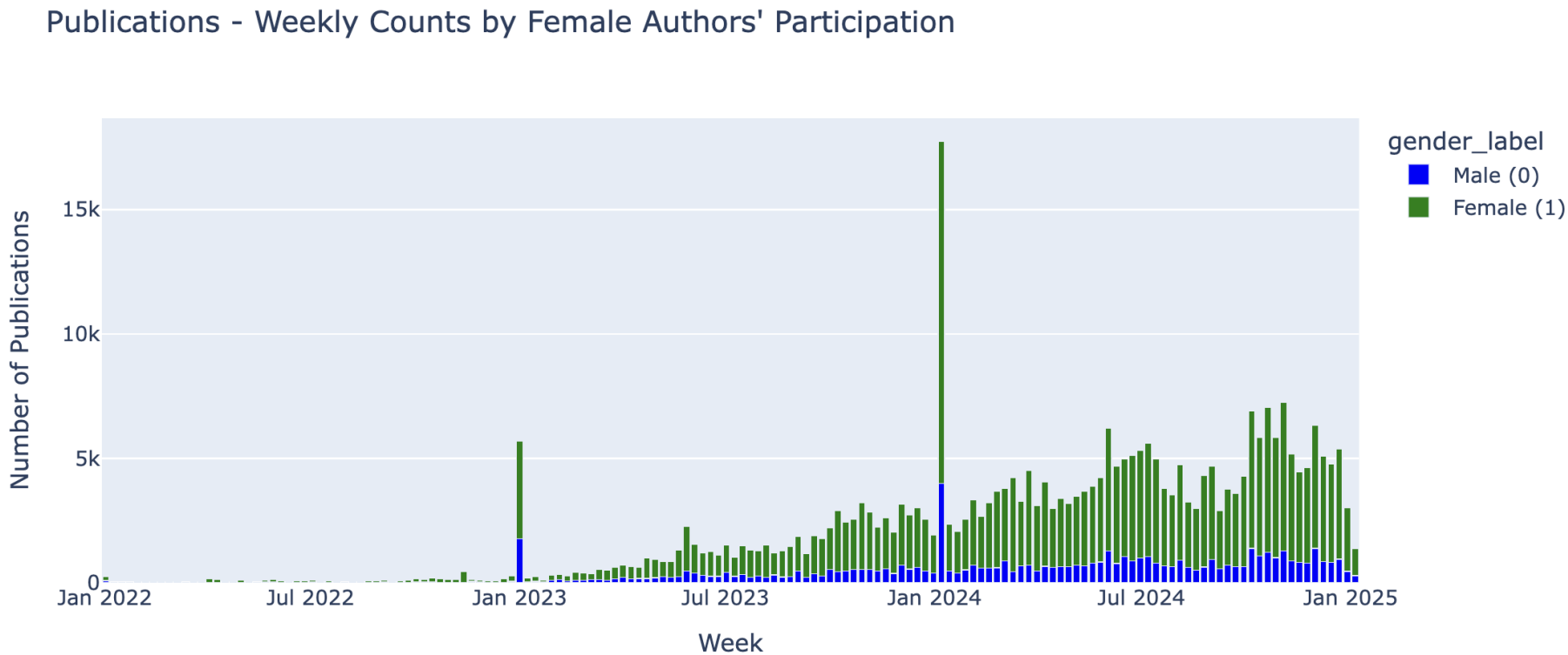
Figure: Gender Distribution


Histogram of Proportions of Male and Female Authors in Publications
Count of Publications
16k
14k
12k
10k
8k
6k
4k
2k
0
0
0.2
0.4
0.6
0.8
1
Proportions of Authors
Gender
Female
Male

The Dimensions field category for pertains to a classification of the Field of Research (FoR) of publications, which is a classification that aligns with the Australian and New Zealand Standard Research Classification (ANZSRC), which arranges research outputs into a hierarchical structure, where major fields are subdivided into more specific minor fields (https://plus.dimensions.ai/support/solutions/articles/23000018826-what-is-the-background-behind-the-fields-of-research-for- protect@classification-system-).



Dimensions.ai maintains a record of grants (in USD) awarded to publication authors (https://docs.dimensions.ai/dsl/2.0.0/datasource-grants.html).



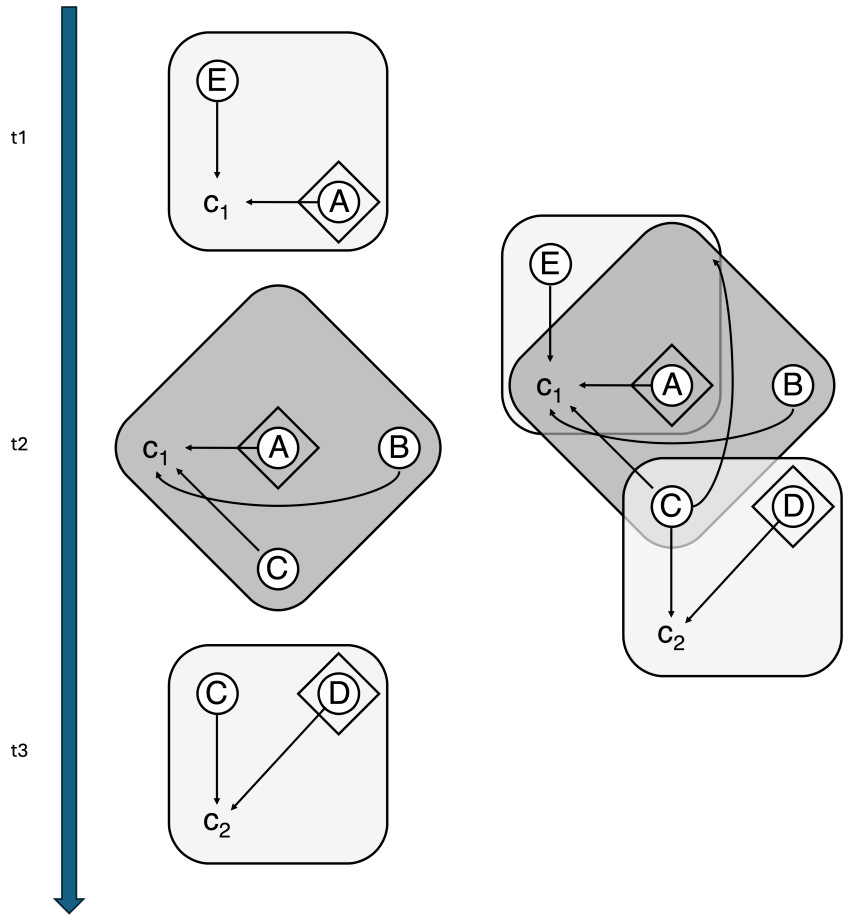
Research areas vs Number of Citations, coloured by index of female partecipation

RHEM introduction

- We focus on the use of the **Relational Hyperevent Model (RHEM)**.
- RHEMs assess the likelihood of continued interactions among actors over time.
- Aim: Analyze how collaborations are influenced by:
  - coauthorship networks
  - Presence of grants
  - Common research fields

Model Framework

- Ideal for modeling co-authorship over time.
- Each interaction (event) can involve multiple participants.



Hypergraph Definition

A hypergraph $G = (V, H)$ is defined by:

- V: Set of nodes (e.g., authors)
- $H \subset V$: Set of hyperedges (e.g., co-authored papers)
- Each hyperedge $h \subseteq V$ can involve any number of nodes

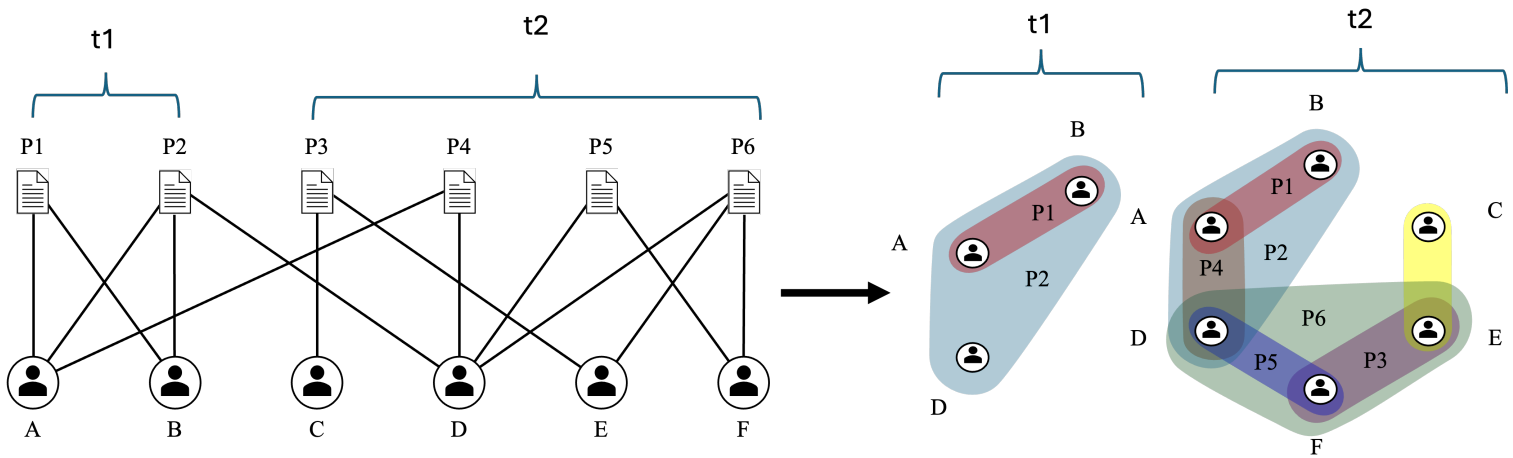
Relational Hyperevent Model (RHEM)

RHEM (Lerner et al., 2019, 2021) is a general framework for modeling networks of time-stamped hyperevents.

A relational hyperevent in an undirected hypergraph is defined as a tuple:

$$e = (h_e, t_e, x_e)$$

- h_e represents the undirected hyperedge;
- t_e is the time of the event (e.g., publication date)
- x_e is the event type and/or event weight



Event Representation

Each undirected hyperevent is a tuple:

$$(t, j, \{i_1, \ldots, i_K\}, \{c_1, \ldots, c_K\}, \{g_1, \ldots, g_K\})$$

- t: Publication time
- j: Published paper
- $\{i_k\}$: Authors
- $\{c_k\}$: Categories
- $\{g_k\}$: Grants

Rich Semantic Modeling

- Each publication = multi-actor event with metadata
- Metadata includes concepts, fields, and grants
- Captures complex academic collaboration patterns
- Enables analysis of temporal, topical, and financial influence

Relational Hyperevent Models (RHEM)

- RHEM supports modeling multi-actor, time-stamped events.
- Enables differentiation of event types:
  1. Publication events
  2. Grant start
  3. Grant end
  4. Categories attribution

Modeling Grant Influence on Collaboration

To examine the role of funding in scientific collaboration, we introduce:

Event Types:

- `grant.start`: Initiation of a new grant.
- `grant.end`: Conclusion of a grant.
- `author`: Authors associated with a publication.
- `categories`: Authors associated with a categories.
- `gender`: Authors are associated to a gender

Funding-Related Attributes:

- `prior.grants` — total number of grants an author has received up to time t.
  - Increases only on `grant.start` events.
  - Includes both active and completed grants.
- `ongoing.grants` — number of currently active grants.
  - Increases with `grant.start`, decreases with `grant.end`.

Cox Proportional Hazard Model

Event rate is decomposed into:

$$\lambda(t, h, \theta, G[E; t]) = \lambda_0(t) \cdot \lambda_1(t, h, \theta, G[E; t])$$

where

$$\lambda_1(t, h, \theta, G[E; t]) = \exp\left(\sum_{i=1}^{k} \theta_i \cdot s_i(t, h, G[E; t])\right)$$

- $\lambda_0(t)$: Baseline rate
- s_i: Network statistics
- θ_i: Parameters to estimate

Effect: Closure

Closure:

$$closure(t, h, G[E, t]) = \sum_{u,v \in \binom{h}{2}} \sum_{w \neq u,v} \frac{\min[\deg(u, w), \deg(v, w)]}{\binom{|h|}{2}}$$

- Captures indirect ties via shared collaborators
- Positive effect $\rightarrow$ more collaboration convergence
- Negative effect $\rightarrow$ structural separation

Effect: Subset Repetition

Subset Repetition of order p:

$$sub.rep^{(p)}(t, h, G[E, t]) = \sum_{h' \in \binom{h}{p}} \deg(t, h', G[E, t]) \cdot \frac{1}{\binom{|h|}{p}}$$

- $p = 1$: author productivity
- $p = 2$: dyadic repetition
- $p = 3$: triadic cohesion

Results

	Explain papers	Explain grants	Explains Concepts
publication activity	−0.689 (0.012)***	−0.192 (0.808)	−0.672 (0.013)***
author.closure	0.818 (0.009)***	0.114 (0.030)***	0.850 (0.010)***
grant.activity	0.090 (0.009)***	0.033 (0.032)	0.065 (0.010)***
ongoing.grants	−0.261 (0.008)***	−0.158 (0.032)***	−0.264 (0.009)***
co-authors	0.092 (0.019)***	−0.017 (0.815)	0.198 (0.020)***
co-PI	−0.101 (0.012)***	−0.001 (0.038)	−0.069 (0.012)***
heterophily_female	−0.459 (0.005)***	−1.648 (0.024)***	−0.454 (0.005)***
cat.gw_6.0			−0.172 (0.009)***
AIC	960596.213	21919.353	662528.634
Num. events	71062	3856	61738
Num. obs.	426372	23136	370428

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Summary

- RHEM is suitable for fine-grained, multi-actor event modeling.
- Models structured interactions such as co-authorship, grants, and concepts.
- Explored the effects of closure and GWSR.
- Enables empirical testing of collaboration dynamics in scientific networks.

Thank you for your attention!