Co-authors and co-PIs: relational hyperevent models for the coevolution of scientific publications and grants

Jürgen Lerner[1]*, Amin Gino Fabbrucci Barbagli[2], and Moses Boudourides[3]

[1] University of Konstanz, Germany juergen.lerner@uni-konstanz.de
[2] University of Trieste, Italy amingino.fabbruccibarbagli@phd.units.it
[3] Northwestern University, USA moses.boudourides@northwestern.edu

Abstract. This study investigates how being co-PIs of scientific grants influences co-authorship, and vice versa, within three Italian Academic Communities (IAC): sociologists (14/c and 14/d), statisticians (SECS S-01/05), and Management (P08) from 2014 to 2023. By collecting data from the Italian Ministry of Education and dimensions.ai, the study employs the Relational Hyperevent Model (RHEM) to analyze their collaboration networks over time. The study demonstrates the complex factors affecting scientific collaboration and expands the application of RHEM to new contexts. Furthermore, we will introduce a new hyperedge covariate, the geometrically-weighted subset repetition (GWSR), as a smoothed version of the formerly defined subset repetition.

Keywords: Hypergraph, Relational Hyperevent Model, Coauthorship networks, Grants

1 Introduction

In recent times, there has been a significant increase in the emphasis placed on research and development (R&D) as a critical factor linked to the competitiveness and success of both institutions and nations. Within this framework, funding may be correlated with the inclination of scholars to engage in co-authorship[12,10], which is widely recognized as a process that facilitates the exchange of ideas, knowledge, and methodologies among researchers, thereby enhancing productivity. In addition, the extensive accessibility of bibliometric data on platforms such as SciVal, Scopus, Google Scholar (GS), Web of Science (WOS), and Dimensions.ai allows for the efficient retrieval of a considerable amount of information related to academic productivity. In this context, several works have proposed network analysis as a convenient tool to study co-authorship [7,5,11], considered as a one-mode network, where two nodes representing the authors are connected if they have published a paper together or it can also be defined as a two-mode network, with two different sets of nodes, papers, and authors. Finally, it can also be represented as hypergraphs, a generalization of graphs[1], where nodes represent authors and hyperedges are the sets of authors of scientific papers. This last setting can be analyzed using the Relational Hyperevent Model (RHEM), a new class of statistical models recently proposed [2,3].

* Corresponding author

In this contribution, building on [4], we propose new specifications for modeling the co-evolution of co-authorship networks and networks linking (co-)PIs to their grants and apply these models to three Italian Academic Communities (IAC): sociologists (14/c and 14/d), statisticians (SECS S-01/05), and Management (P08) from 2014 to 2023. The data was collected from the Italian Ministry of Education and Dimensions.ai and analyzed using the Relational Hyperevent Model. Furthermore, we will introduce the geometrically-weighted subset repetition (GWSR) and its first application.

2 General Framework

The Relational Hyperevent Model (RHEM) is a recent family of statistical models useful to assess the propensity of actors to interact over time, where interaction may be polyadic, that is, more than just two actors may interact in a given event [2,6]. In particular, RHEMs can deal with time-stamped events as in co-authorship networks considering papers published in different years [3]. For this work, we applied RHEM to assess the propensity of our scholars to continue their collaboration and how the presence of grants influences their collaborations.

Formally, a hypergraph $G = (V,H)$ is defined by a set of nodes V and a set of hyperedges $H \subseteq P(V)$. Each hyperedge $h \in H$ is a subset of nodes $h \subseteq V$ of any size, where $|h|$ represents its cardinality. Given a set of nodes V, an undirected hyperevent is defined as a tuple $e = (I_e,t_e,x_e)$. $Ie \subseteq V$ represents an undirected hyperedge, denoting the participants of the event; t_e is the time of the event (i.e., the publication date); x_e is the event type and/or event weight. This allows for the categorization and differentiation of events, in our case, publication events of scientific papers, grant start events (representing the starting date of the grant), and grant end events (representing the ending date of the grant).

The ***event rate*** (*hazard rate or intensity*) [2] on h at time t (given the network with the past events) is defined as:

$$\lambda(t;I;G[E;t]) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(t \leq T \leq t+\Delta t | I_e = I \wedge t \leq t_e < t+\Delta t\})}{\Delta t} \tag{1}$$

We model the likelihood of a sequence of relational hyperevents, $E = (e_1,...,e_N)$, using a Cox proportional hazard model [14]. For a given time point t, the network of past events, $G[E;t]$, comprises all events in E occurring before t [8]. The event rate, $\lambda(t;I)$, is decomposed into a time-dependent baseline rate, $\lambda_0(t)$ typically left unspecified, which is constant for all hyperedges, and a relative event rate, $\lambda_1(t;I;\theta;G[E;t])$. This relative rate is conditional on hyperedge statistics, $s(t;I;G[E;t]) \in \mathbb{R}^k$, derived from the past event network. that indicates how the hyperedge h is embedded into the network of past events, and a parameter vector, $\theta \in \mathbb{R}^k$ [14] describing which of these statistics increase or decrease the relative event rate λ_1.

Based on the observed event sequence E, the *partial likelihood* becomes:

$$L(\theta) = \prod_{e \in E} \frac{\lambda_1(I_e;t_e,\theta;G[E;t])}{\sum_{I \in R_{te}} \lambda 1(I,t_e,\theta,G[E,t_e])} \tag{2}$$

Given the values of the statistics, $s_i(t,I,G[E;t])$ for all elements of the risk sets R_{te} at the event times t_e, the maximum likelihood estimates for Eq. (4) can be computed using standard statistical software [2].

We defined the network effects: *closure* [4][2] given two authors u and v who have previously worked with at least one-third common author, closure assesses the propensity that these two actors may interact directly in the future [6]. A positive closure in co-authorship networks indicates a higher probability of future collaboration between authors who have not previously co-authored, but share common collaborators. This suggests a tendency for overlapping hyperedges (representing co-authorship groups) to merge over time, as authors with shared connections are more likely to form new collaborations [3]. On the other hand, a negative closure indicates that the two authors, who share a common co-author, will not start working together, meaning that overlapping hyperedges may stay stable without merging [6].

The *subset repetition* of order p [2] considers an exact number p of actors that participated in the previous hyperevents and it returns their propensity to participate again in a joint event in the future. For example, $p = 1$ indicates the propensity of an individual author to continue her/his activity in the future (general attitude to publish), $p = 2$ suggests the tendency of a dyad of authors to participate again in a joint event, $p = 3$ of a triad [2] [3] and so on.

We introduce a new hyperedge covariate, the geometrically-weighted subset repetition (GWSR), as a smoothed version of the formerly defined subset repetition [6], whose scaling function is similar to those employed for geometrically-weighted statistics in exponential random graph models [13]. Let a sequence of relational hyperevents be given by:

$$E = (t_1,I_1),\ldots,(t_n,I_n) \ ,$$

where t_m is the time of the m'th hyperevent and $I_m \subseteq \mathscr{I}_{t_m}$ are the participating nodes of the m'th hyperevent. For a point in time t and a set of nodes $I \subseteq \mathscr{I}_t$, subset repetition of order p

$$sub.rep^{(p)}(t,I) = \frac{1}{\binom{|I|}{p}} \cdot \sum_{I' \in \binom{I}{p}} hy.deg(t,I') \ ,$$

where the "hyperedge degree", ignoring any decay in time, is defined by

$$hy.deg(t,I') = \sum_{t_m<t} \mathbf{1}(I' \subseteq I_m) \ ,$$

that is, we count the number of previous events (t_m,I_m), such that I' is contained in I_m. In other words: all nodes in I' co-participate in the m'th event. If we have weighted events, and/or a decay over time, this can all be incorporated in the definition of the hyperedge degree. We note that an event hyperedge I_m increases $sub.rep^{(p)}(t,I)$ by $\binom{|I \cap I_m|}{p}$. (Note that $\binom{k}{p}$ is zero if $k < p$). Thus, subset repetition of order p can be equivalently be defined by:

$$sub.rep^{(p)}(t,I) = \frac{1}{\binom{|I|}{p}} \cdot \sum_{t_m<t} \binom{|I \cap I_m|}{p} \ , \tag{3}$$

We define *geometrically-weighted subset repetition* with (fixed) real scaling parameter $\kappa \geq 0$ as follows.

$$gwsr^{(\kappa)}(t,I) = \frac{\exp(\kappa)}{|I|} \sum_{t_m<t} \left\{ 1 - \left(1 - \frac{1}{\exp(\kappa)} \right)^{|I \cap I_m|} \right\} \cdot |I \cap I_m| \ . \tag{4}$$

The higher the value of $\kappa \geq 0$, the higher the relative factor scaling the contribution of hyperedges I_m that have large overlap $|I_m \cap I|$. If $\kappa = 0$, then the behavior of the geometrically weighted subset repetition, with respect to the source nodes, is identical to the behavior of the subset repetition of order $p = 1$. If κ increases, the statistic assigns greater weight to hyperedges with a larger overlap, allowing to test whether there is a tendency for repeated co-participation in events. To investigate how grants

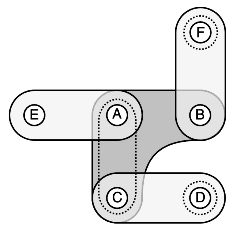

Fig. 1: Co-authors and co-PIs hypergraph. representing the co-authorship network. Dashed lines represent grants, and solid lines represent publications. A and C are co-PIs, and they wrote a joint publication with B.

influence collaboration among authors, we defined three events, two corresponding to a specific grant-related event as *grant.start* that represents the initiation of a new grant and *grant.end* that indicates the end of a grant. In both events, the participants are the researchers associated with the grant. Furthermore, we associate each hyperedge with two attributes that track the funding status over time: *prior.grants* and *ongoing.grants*. The first attribute records the total number of grants an author has received in the time period, including active and completed grants. It increments only upon the occurrence of a *grant.start* event. The second attribute, conversely, represents the number of currently active grants for an author. It increments upon a *grant.start* event and decrements upon a grant.end event. By utilizing these two attributes, we can conduct more in-depth analyses of the impact of grants on co-authorship. We can determine whether grants, in general, influence authors' propensity to collaborate. In other words, we can explore whether having current funding influences collaboration differently than all prior grants. We defined the event *author* to represent the publication of a paper. This allows us to treat each publication as a distinct event, with the associated author(s) being the source responsible for that publication (target). The interactions between hyperedges, representing authors and papers, are multifaceted and involve several concurrent counting processes. A key example of this is the number of co-authored papers. For any given pair of authors, we can track the number of papers they have co-authored. This count increments when a new paper is published where both authors are listed. Each counting process provides a different perspective on the relationships and activity within the network. By analyzing these parallel processes, we can gain a deeper understanding of collaboration patterns, research trends, and the overall dynamics of the community represented by our co-authorship network.

3 Data and results

	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
# publications	123	146	157	142	166	184	247	229	203	255	1863
#grant.start	14	12	18	18	10	17	16	4	2	1	112
#grant.end	3	7	6	6	14	18	19	14	9	16	112

Table 1: Number of granted publications, starting grant and ending grant for each year

This contribution aims to study the co-authorship networks of 3 Italian Academic Communities (IAC): sociologists (14/c and 14/d), statisticians (SECS S-01/05), and Management (P08) from 2014 to 2023. Data was collected using a rigorous and multi-step methodology using various data sources. The initial dataset was created by gathering information on the target group from the Italian Ministry of Education [9]. We retrieved the name, surname, and field of our target group. Using the Dimensions.ai API, we gathered the paper production of our communities and the grants associated with them from Dimensions.ai. The number of granted publications and ”grant.start” and ”grant.end” events are reported in Table 1.

	Explain papers	Explain grants
publication activity	$-2.759\ (0.129)^{***}$	$-22.301\ (3.101)^{***}$
closure by coauthor	$-0.234\ (0.025)^{***}$	$-3.733\ (0.979)^{***}$
grant activity	$-1.391\ (0.386)^{***}$	$-0.143\ (0.658)$
ongoing grant activity	$-1.564\ (0.189)^{***}$	$-0.898\ (0.241)^{***}$
co-authors	$3.563\ (0.130)^{***}$	$20.345\ (2.992)^{***}$
co-PIs	$0.892\ (0.359)^{*}$	$-0.845\ (0.647)$
AIC	21250.299	1369.604
Num. events	1837	149
Num. obs.	185537	14889

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

Table 2: RHEM explaining the set of authors of scientific publications (*left*) and the PIs of scientific grants (*right*), respectively.

Estimated model parameters are reported in Table 2. The model explaining papers indicates that individuals who have published more in the past will publish at a lower rate in the future, as suggested by the negative effect of “publication activity. The negative coefficient for “closure by coauthor” indicates a tendency against collaboration among co-authors of the same third author; it has been discussed [3] that this may point to sub-communities that are overlapping but that do not merge over time. The negative signs for “grant activity” and “ongoing grant activity” suggest a tendency for scientists who are PIs of more (ongoing) grants to actually publish less, all other things being equal. The positive effect for ”co-PI” and ”co-authors” suggest that scientists who have been co-PIs before are more likely to be co-PIs of future grants, and authors who have collaborated in the past are more likely to do so again in the future.

The model explaining grants suggests a tendency for scientists who have published more in the past to acquire grants at a lower rate (negative parameter of "publication activity"). We also find a negative tendency of scientists to become co-PIs of their coauthors' coauthors. The number of all past grants has no significant effect on the rate to acquire grants. Still, the number of ongoing grants has a tendency to reduce the current rate of funding (negative effect of "ongoing grant activity"), which may point to a saturation effect limiting the number of concurrent ongoing grants that scientists may have. Finally, scientists who have been co-PIs before are less likely to be co-PIs of future grants, but authors who have been co-authors are more likely to repeat the collaboration in the future.

References

1. Bretto, A. Hypergraph theory. *An Introduction. Mathematical Engineering. Cham: Springer*. **1** (2013).
2. Lerner, J., Tranmer, M., Mowbray, J. & Hancean, M. REM beyond dyads: relational hyperevent models for multi-actor interaction networks. *ArXiv Preprint ArXiv:1912.07403*. (2019)
3. Lerner, J. & Hâncean, M. Micro-level network dynamics of scientific collaboration and impact: relational hyperevent models for the analysis of coauthor networks. *Network Science*. **11**, 5-35 (2023)
4. Lerner, J., Hâncean, M. & Lomi, A. Relational hyperevent models for the coevolution of coauthoring and citation networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae068 (2024)
5. De Stefano, D., Kronegger, L., Sciabolazza, V., Vitale, M. & Zaccarin, S. Social network tools for the evaluation of individual and group scientific performance. *Teaching, Research And Academic Careers*. pp. 165 (2022)
6. Lerner, J. & Lomi, A. A dynamic model for the mutual constitution of individuals and events. *Journal Of Complex Networks*. **10**, cnac004 (2022)
7. De Stefano, D., Fuccella, V., Vitale, M. & Zaccarin, S. Quality issues in co-authorship data of a national scientific community. *Network Science*. **11**, 98-112 (2023)
8. Brandes, U., Lerner, J., Snijders, T. AB. Networks evolving step by step: Statistical analysis of dyadic event data. In: *2009 international conference on advances in social network analysis and mining*. IEEE, 2009. p. 200-205.
9. Official Ministerial Italian Higher Education Database (MUR), https://cercauniversita.cineca.it/php5/docenti/cerca.php, note = Accessed: 2025-02-27
10. Andrade, H. B., de Los Reyes Lopez, E., & Martín, T. B. Dimensions of scientific collaboration and its contribution to the academic research groups' scientific quality. *Research Evaluation*, **18**(4), 301–311 (2009).
11. Kronegger, L., Mali, F., Ferligoj, A., & Doreian, P. Collaboration structures in Slovenian scientific communities. *Scientometrics*, **90**(2), 631–647 (2012).
12. Mali, F., Pustovrh, T., Platinovšek, R., Kronegger, L., & Ferligoj, A. The effects of funding and co-authorship on research performance in a small scientific community. *Science and Public Policy*, **44**(4), 486–496 (2017).
13. Hunter, D. R., & Handcock, M. S. (2006). Inference in Curved Exponential Family Models for Networks. *Journal of Computational and Graphical Statistics*, **15**(3), 565–583.
14. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202 (1972).