Untangling the Academic Network: Analyzing Scientific Trajectories With Graphs and Natural Language Processing

Rodrigo Molina[1,2,*], Roxana Chiappa[3] and Sebastián Ferrada[1,2,4]

[1]Universidad de Chile, Santiago, Chile
[2]Millennium Institute for Foundational Research on Data (IMFD), Chile
[3]Universidad de Tarapacá, Chile
[4]National Center for Artificial Intelligence Research, Santiago, Chile

Abstract

We analyze 3,143 Chilean doctoral fellows from the 2014–2016 ANID cohorts through a Knowledge Graph that federates eight heterogeneous sources into 16 node types, 15 relationship types, and over 3 million edges. BERTopic yields 237 topics over 26,071 Scopus abstracts. In the intra-cohort co-authorship network, 84.2% of fellows share no publication with another fellow, and its largest component splits into 21 communities that are thematically coherent but heterogeneous in gender and socioeconomic composition.

Keywords

Knowledge Graphs, Natural Language Processing, Academic Networks

1. Introduction

Doctoral scholarship programs rank among the largest public investments in scientific human capital. In Chile, the Becas Chile and Becas CONICYT/ANID programs have awarded doctoral scholarships annually since 2008. Yet whether state funding translates into equitable integration within scientific production networks, and how gender and social origin shape that integration, remains an open question. Administrative records, bibliometric profiles, and institutional rankings exist, but they are siloed across disconnected databases that lack common identifiers and relational structure.

We address this integration gap by constructing a Knowledge Graph that fuses eight heterogeneous sources into a unified structure with 16 node types, 15 relationship types, and over 3 million edges, modeling the trajectories of 3,143 fellows from the 2014–2016 Becas ANID cohorts. The graph models a fellow's trajectory as a directed sequence of institutional positions: secondary school (whose MINEDUC socioeconomic group proxies the fellow's social origin), undergraduate institution, and doctoral program, connected to their Scopus-indexed scientific output (publications, journals, co-authors) and postdoctoral funding (Fondecyt projects). On this substrate, we layer (i) BERTopic [1] for topic modeling over 26,071 publication abstracts, yielding 237 topics that capture the semantic content of scientific production, and (ii) graph analytics to quantify structural positions within the co-authorship network.

We contribute an end-to-end pipeline from raw heterogeneous records to a queryable property graph with precomputed analytics. A single Cypher traversal over the graph recovers a complete trajectory (educational path, co-authorship structure, topic profile, postdoctoral funding). The graph and its interactive visualizations are public at https://interrogandolaacademia.cl/ and https://rodrigomolinaavila.github.io/gds-becarios/.

AMW 2026: 17th Alberto Mendelzon International Workshop on Foundations of Data Management, November 9th–November 13th, 2026, Arequipa, Peru
*Corresponding author.
rodmolina@ug.uchile.cl (R. Molina); sebastian.ferrada@uchile.cl (S. Ferrada)
0009-0006-0648-0799 (R. Molina); 0000-0002-3523-1197 (R. Chiappa); 0000-0002-9421-8566 (S. Ferrada)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

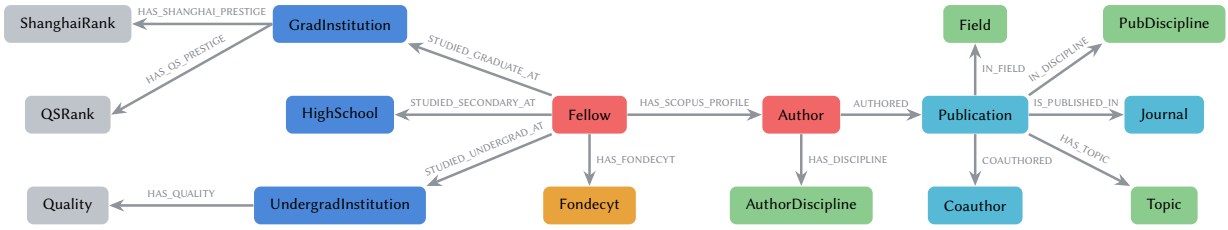

Figure 1: Node color: fellow and author (red), educational (blue), scientific (light blue), thematic (green), funding (orange), external (grey). Edge labels show relationship type.

Table 1
Graph inventory: all node types and relationships.

Nodes – Total: 155,391					
Label	**Count**	**Label**	**Count**	**Label**	**Count**
Coauthor	113,221	GradInstitution	1,288	Topic	237
Publication	26,071	Fondecyt	1,031	Field	27
Journal	5,488	HighSchool	797	AuthorDiscipline	233
Fellow	3,143	PubDiscipline	325	Quality	9
Author	2,703	ShanghaiRank	301		
UndergradInstitution	253	QSRank	264		
Relationships – Total: 3,087,304					
Type	**Count**	**Type**	**Count**	**Type**	**Count**
COAUTHORED	2,884,376	AUTHORED	28,811	HAS_FONDECYT	1,031
IN_DISCIPLINE	62,938	HAS_TOPIC	25,183	HAS_SHANGHAI_PRESTIGE	629
IN_FIELD	46,706	IS_PUBLISHED_IN	22,091	HAS_QS_PRESTIGE	597
STUDIED_SECONDARY_AT	3,143	STUDIED_GRADUATE_AT	3,143	HAS_QUALITY	253
STUDIED_UNDERGRAD_AT	3,073	HAS_SCOPUS_PROFILE	2,675	HAS_DISCIPLINE	2,655

2. Trajectory Knowledge Graph Creation

We federated eight heterogeneous sources to model the academic trajectories of 3,143 doctoral fellows (cohorts 2014–2016). ANID is Chile's national research agency; MINEDUC, its Ministry of Education; and CNA, its National Accreditation Commission. **ANID administrative records** (Transparency Portal) provide data on scholarship type, doctoral program, discipline, and funding dates. **ANID's GitHub repository** contributes postdoctoral Fondecyt projects: 1,031 projects from 941 fellows across the Postdoctorado, Iniciación, and Regular programs. **ANID Portal del Investigador** adds curricular data for 1,749 fellows (55.6%). **Scopus APIs** (Author, Abstract Retrieval, Journal) link 2,675 fellows (85.1%) to 26,071 publications, 5,488 journals, and 113,221 co-authors. **MINEDUC** provides the school-level socioeconomic group (GSE). **QS** (2025) and **Shanghai** (2024) rankings match 597 and 629 graduate institutions, respectively. **CNA** contributes 9 undergraduate accreditation tiers.

Entity resolution. Author disambiguation combined name matching, affiliation filtering, and verification. Gender was inferred in two stages: first with genderizeR [2], then with the guaguas dataset of names registered with the Chilean Civil Registry [3], reaching 99.1% effectiveness. Institutional names were normalized across sources via unification dictionaries.

Graph schema. The Knowledge Graph centers on Fellow and organizes the remaining fifteen node types into five groups (Figure 1). Educational trajectories link HighSchool, UndergradInstitution, and GradInstitution, the latter connected to ShanghaiRank and QSRank. Scientific production follows Author → Publication → Journal, with co-authorship captured via Coauthor (including an is_fellow_author flag); the COAUTHORED relationship accounts for 93.4% of all edges. Publications are thematically classified through Topic (BERTopic), Field (Scopus ASJC), and PubDiscipline. Postdoctoral funding is represented by Fondecyt. Additional nodes include AuthorDiscipline and Quality (CNA accreditation tier). Fellow nodes also store denormalized metrics (e.g., productivity over time, SJR quartile distributions, estimated rankings) to support direct Cypher aggregation. Figure 1 and Table 1 summarize the schema and graph size.

Table 2
Largest topics by publication volume and by gender dominance.

Largest (by #pubs)		**Lowest % female (>0% F)**	
Topic	n	**Topic**	**%F**
tev, detector, atlas, collisions	1,228	entropy, information, shannon	2.1
reaction, properties, synthesis	724	bubble, foam, bubbles, flow	2.9
species, south, distribution	654	exoplanets, techniques, imaging	2.9
education, school, students	586	numerical, equations, finite	3.1
health, clinical, mental	542	scheduling, problem, jobs	3.3
Most male-authored (% male)		**Most female-authored (% female)**	
Topic	**%M**	**Topic**	**%F**
magnetic, simulations, field	100.0	gall, galls, plant, host	95.7
power, system, frequency, grid	100.0	cancer, ovarian, epithelial	87.5
low, neural, energy, neuron	100.0	atomic, layer, deposition	80.0
control, systems, stability	100.0	genetic, population, latin	78.6

3. Topic Modeling

We apply BERTopic [1] to the 26,071 publication abstracts. Each document combines its title, abstract, and keywords. Documents are embedded via Sentence-BERT (`paraphrase-multilingual-MiniLM-L12-v2`) [4] into $\mathbb{R}^{384}$. UMAP [5] reduces to $\mathbb{R}^5$, and HDBSCAN [6] detects variable-density clusters. Noise points are reassigned via cosine proximity to the nearest topic centroid in the original embedding space. Each cluster is characterized by c-TF-IDF, extracting the top-10 most discriminative terms.

BERTopic hyperparameters were optimized via Optuna with a TPE sampler [7], using two-level pruning: an HDBSCAN internal validity index as a fast filter, then full c_v coherence [8] for surviving configurations.[1] A semantic fusion step merged topic pairs on embedding similarity and entailment, yielding 237 final topics. The moderate AMI indicates that topics track disciplinary boundaries without being reducible to them.

The 237 topics emerge directly from the corpus and capture the semantic structure of scientific production at several levels of granularity: broad fields such as education and public health coexist with narrow research niches such as salmon genomics, particle detector instrumentation, and species distribution modeling. No topic is exclusively female-authored, while eight topics (all in physics, engineering, and neuroscience) have zero female authors (Table 2). Topic assignments are stored as `HAS_TOPIC` relationships, enabling Cypher queries that cross-tabulate gender and semantic content.

4. Community Detection

The intra-cohort co-authorship network links fellows who share publications, weighted by the number of joint works. A Cypher projection yields 3,143 nodes; 84.2% of fellows are isolated, and the intra-cohort component holds 496 fellows and 700 edges.

We apply Louvain [10] and Leiden [11] to that component, obtaining 21 and 23 communities. Both partitions are stable across implementations and algorithms, and strongly separated from Erdős–Rényi

[1] c_v measures how consistently a topic's top terms co-occur in the corpus. The winning configuration ($n_neighbors$=17, min_dist=0.525, $min_cluster_size$=24, $min_samples$=15) achieved c_v=0.692. Against permuted null models it validates on two axes: silhouette 0.0344 vs. null -0.198 ± 0.016 and adjusted mutual information (AMI) against OECD doctoral discipline 0.329 vs. 0.0001 ± 0.0005, both p<0.0001. Topic pairs were fused when centroid cosine similarity exceeded 0.55 and bidirectional entailment reached 0.60, with entailment scored by mDeBERTa-v3 fine-tuned on MNLI and XNLI [9].

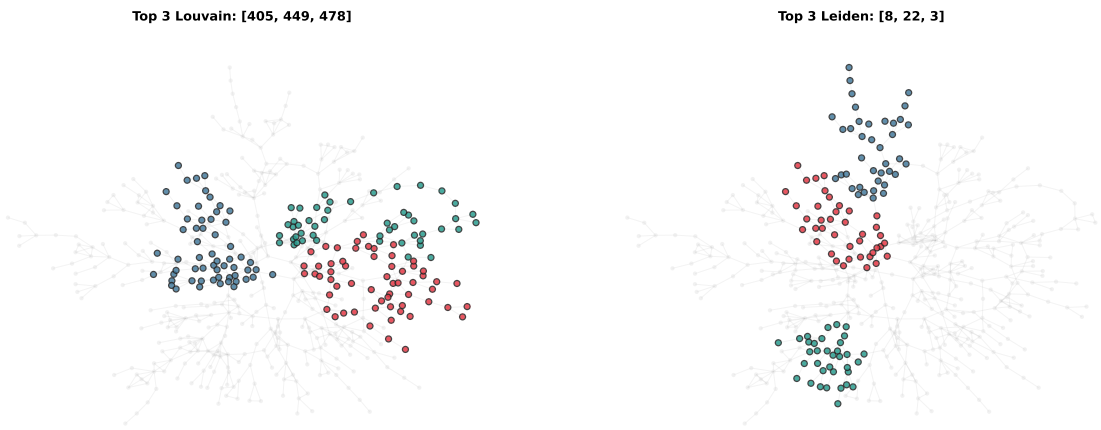

Figure 2: Louvain (left) and Leiden (right). Top 3 communities colored; grey: rest. Size $\propto$ PageRank.

nulls.[2]

The communities reflect distinct research foci. Among Louvain partitions, C478 (salmon genomics, $n = 49$) concentrates Chile's aquaculture research in high-impact international venues (Q1: 80.9%), with near gender parity (49.0% F) and the highest proportion of fellows with graduate degrees abroad (22.4%). C483 (clinical and patient-centered research, $n = 31$) is the most feminized (54.8% F) yet publishes in systematically lower-impact journals (Q1: 58.9%). C452 (microbiology and infection, $n = 33$) achieves high Q1 (78.0%) despite the lowest proportion of fellows from high-socioeconomic backgrounds (GSE Alto: 6.1%). C176 (neuroscience) and C405 (cell biology) are intermediate: 37–48% female authors, Q1 above 74%.

Leiden partitions corroborate these patterns: C3 ($n = 35$) mirrors C452's profile, with only 5.7% GSE Alto yet 65.9% Q1, while C7 ($n = 30$) achieves the highest Q1 among the top Leiden communities (82.7%) with balanced gender composition (40.0% F). Across both algorithms, seven of the twelve largest communities are organized around a single dominant topic that accounts for over 10% of their publications, indicating that the detected groups are thematically coherent rather than artifacts of the method.

We compute PageRank [12], betweenness [13], and closeness [14] on the intra-cohort component via Neo4j GDS. Figure 3 shows the centrality heatmap of each metric; Table 4 reports the 90th percentile. P90 captures the top 10% most central fellows ($n = 49$). The remaining 447 fellows serve as a baseline.

Women are underrepresented at the top: 32.7% of betweenness P90 and 36.7% of closeness P90 are female, versus 40.7% in the rest of the component. By socioeconomic origin (GSE, the school-level indicator used as a social-class proxy), the highest stratum (Alto) never exceeds its baseline share (18.4%–22.4% vs. 19.0%), while the middle stratum (Medio) matches or exceeds it (28.6%–34.7% vs. 28.8%). These patterns suggest that structural brokerage and proximity, the "bridge" and "hub" roles captured by betweenness and closeness, are more gender-skewed than recursive influence (PageRank, 42.9% female at P90). Betweenness P90 fellows also publish more Q1 papers than the baseline (+5.5 pp.), whereas PageRank and closeness P90 are indistinguishable from it (61.8% and 65.5% vs. 65.4%).

[2]NetworkX Louvain vs. GDS: NMI=0.835 (27 communities); Leiden vs. Louvain: NMI=0.939, ARI=0.847; Erdős–Rényi null: $z{>}35\sigma$ for both; stochastic block model: 3–4 coarser macro-blocks, NMI=0.467, indicating hierarchical rather than artifactual structure.

Table 3
Top 6 Louvain and Leiden communities by size, with dominant topic. Alto/Medio: high and middle school-level socioeconomic group (GSE).

Louvain (Q=0.895, 21 communities)						
ID	**Size**	**%F**	**%Alto**	**%Medio**	**%Q1**	**Dominant topic**
C405	59	37.3	27.1	15.3	78.5	cells, induced, expression
C449	58	34.5	27.6	13.8	66.4	reaction, properties, synthesis
C478	49	49.0	26.5	20.4	80.9	genes, salmon, fish
C452	33	48.5	6.1	15.2	78.0	difficile, infection, spores
C483	31	54.8	29.0	25.8	58.9	treatment, methods, patients
C176	31	48.4	22.6	6.5	74.2	neurons, brain, expression
Leiden (Q=0.886, 23 communities)						
ID	**Size**	**%F**	**%Alto**	**%Medio**	**%Q1**	**Dominant topic**
C8	40	25.0	30.0	10.0	56.3	neurons, brain, expression
C22	39	46.2	25.6	20.5	65.8	treatment, methods, patients
C3	35	40.0	5.7	31.4	65.9	properties, materials, polymer
C14	31	41.9	12.9	19.4	74.7	reaction, properties, synthesis
C7	30	40.0	16.7	13.3	82.7	cells, cell, protein
C20	27	51.9	22.2	7.4	80.3	genes, salmon, fish

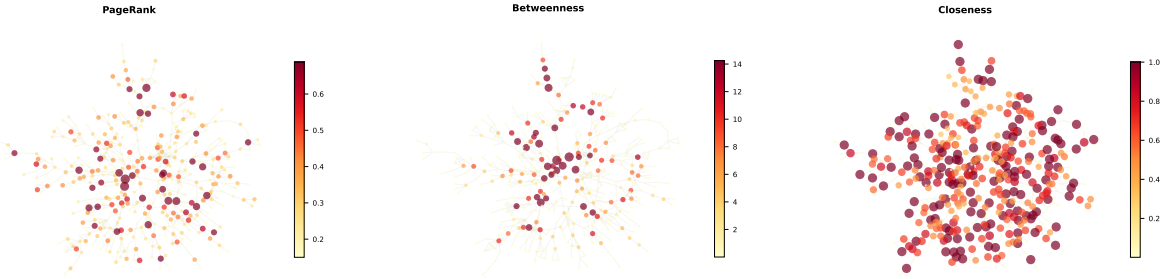

Figure 3: Centrality Heatmap

Table 4
Centrality P90: composition and quality.

Metric	n	**%F**	**%Alto**	**%Medio**	**%Q1**
PageRank P90	49	42.9	18.4	30.6	61.8
Betweenness P90	49	32.7	22.4	28.6	70.9
Closeness P90	49	36.7	18.4	34.7	65.5
Rest (447)	–	40.7	19.0	28.8	65.4

5. Conclusions and Future Work

We present a Knowledge Graph of eight heterogeneous sources with 16 node types, 15 relationship types, and over 3 million edges, modeling 3,143 Chilean doctoral trajectories and enabling single-query reconstruction. The graph and its visualizations are public at https://interrogandolaacademia.cl/ and https://rodrigomolinaavila.github.io/gds-becarios/.

On the intra-cohort co-authorship network it reveals a fragmented topology: 84.2% of fellows share no publication with a peer, and the intra-cohort component splits into 21 Louvain and 23 Leiden communities. Centrality at P90 associates structural position more with gender than with socioeconomic origin. Because current instruments reward individual output, these results argue for evaluating programmes on network integration instead. Centrality gaps remain associations, not programme effects, and Scopus coverage and a single cross-section of the 2014–2016 cohorts bound the claims.

For funders we propose tracking co-authorship and postdoctoral insertion as programme indicators; for data policy, publishing linked, anonymised trajectory records as evaluation infrastructure. Future work should separate funded from non-funded applicants to move from association to effect, analyze temporal homophily in co-authorship, add regional databases (SciELO, Redalyc) against Scopus bias,

train graph networks on initial positions to predict career outcomes, and explain qualitatively why within-cohort collaboration is rare.

Acknowledgments

Molina and Ferrada were funded by ANID - Millennium Science Initiative Program - Code ICN17_00; by CENIA FB210017, Basal ANID; and by FONDECYT Grant 11251322. Molina and Chiappa were funded by FONDECYT Grant 11241304.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Opus 5.5 (Anthropic) in order to: improve writing style, check grammar and spelling, and paraphrase and reword. The author(s) also used Claude Code (Anthropic), OMP (Oh My Phi agent harness), and DeepSeek v4.1 to support code development for the analysis pipeline and the revision of this manuscript. After using these tools, the author(s) reviewed and edited all generated content as needed and take full responsibility for the publication's content.

References

[1] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022). URL: https://arxiv.org/abs/2203.05794.

[2] K. Wais, Gender prediction methods based on first names with genderizer, The R Journal 8 (2016) 17–37. doi:10.32614/RJ-2016-002.

[3] R. Quiroga, guaguas: Nombres Inscritos en Chile (1920–2021), 2022. URL: https://CRAN.R-project.org/package=guaguas. doi:10.32614/CRAN.package.guaguas, R package version 0.3.0, data from the Chilean Civil Registry Service.

[4] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[5] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, Journal of Open Source Software 3 (2018) 861. doi:10.21105/joss.00861.

[6] R. J. G. B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: Advances in Knowledge Discovery and Data Mining (PAKDD), Springer, 2013, pp. 160–172. doi:10.1007/978-3-642-37456-2_14.

[7] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 2019, pp. 2623–2631. doi:10.1145/3292500.3330701.

[8] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM), 2015, pp. 399–408. doi:10.1145/2684822.2685324.

[9] M. Laurer, W. van Atteveldt, A. Casas, K. Welbers, Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli, arXiv preprint arXiv:2305.13703 (2023). URL: https://arxiv.org/abs/2305.13703.

[10] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (2008) P10008. doi:10.1088/1742-5468/2008/10/P10008.

[11] V. A. Traag, L. Waltman, N. J. van Eck, From louvain to leiden: guaranteeing well-connected communities, Scientific Reports 9 (2019) 5233. doi:10.1038/s41598-019-41695-z.

[12] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report 1999-66, Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/.

[13] L. C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41. doi:10.2307/3033543.

[14] A. Bavelas, Communication patterns in task-oriented groups, The Journal of the Acoustical Society of America 22 (1950) 725–730. doi:10.1121/1.1906678.