Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detecting Scientific and Technological Trajectories in Cumulative Citation Networks: The birddog Method and Evidence from Biogas Research

Roney Fraga Souza

roney.souza@ufmt.br

Universidade Federal de Mato Grosso

Luis Felipe de Souza Rodrigues

Núcleo Interdisciplinar de Planejamento Energético

Júlio Cesar de Carvalho Miranda

Universidade Federal de Mato Grosso

José Maria Ferreira Jardim da Silveira

Universidade Estadual de Campinas (UNICAMP)

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11364586/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Detecting Scientific and Technological Trajectories in Cumulative Citation Networks: The birddog Method and Evidence from Biogas Research

Roney Fraga Souza[a,*], Luis Felipe de Souza Rodrigues[b], Júlio Cesar de Carvalho Miranda[c], José Maria Ferreira Jardim da Silveira[d]

[a] *Federal University of Mato Grosso, Faculty of Economics, Cuiabá, Mato Grosso, Brazil*
[b] *University of Campinas, Interdisciplinary Center for Energy Planning (NIPE), Campinas, São Paulo, Brazil*
[c] *Federal University of Mato Grosso, Bioenergy Research Center, Cuiabá, Mato Grosso, Brazil*
[d] *University of Campinas, Institute of Economics, Campinas, São Paulo, Brazil*

Abstract

Technological trajectories are central to evolutionary economics, yet few empirical instruments observe how variants are generated, selected, and retained or abandoned within populations of research lines. Citation-network studies typically trace trajectories as main paths, a field's dominant line of descent; we extend this program from the route to the population with an unsupervised method. Re-clustering a cumulative citation network annually yields a temporal graph of cluster–year nodes linked by document migration. A deterministic trunk decomposition partitions this graph into lineages: every node belongs to exactly one trajectory. Each trajectory's state names an evolutionary mechanism. Lineages reaching the present take a life-cycle phase from growth and novelty: emergence (variety generation), maturity (retention), or dormancy (exhaustion). Tributaries, which end earlier, take a terminal fate from how their documents disperse: convergence (selection and retention), divergence (variety retained under unresolved selection), or, when

*Corresponding author

Email addresses: roneyfraga@gmail.com (Roney Fraga Souza), luisfsrodrigues.lfsr@gmail.com (Luis Felipe de Souza Rodrigues), julio.miranda@ufmt.br (Júlio Cesar de Carvalho Miranda), jmsilv@unicamp.br (José Maria Ferreira Jardim da Silveira)

extinct, dormancy. Applied to 77,347 biogas documents from OpenAlex, the method finds 19 communities and 98 trajectories. Variety and selection coexist, an open selection environment rather than lock-in: the fastest-growing living lineages are young, and 60 of 79 tributaries fold into surviving lineages, most scattering their documents across several communities. Extinct lineages lose their names but not their papers; among the largest communities, biogas upgrading grew almost entirely on its own, pretreatment largely by absorption. The field moves from rural experimentation (1990s) through industrialization and standardization (2010s) to methods intensification and systems integration (2020s). The open-source R package `birddog` implements the method.

Keywords: Technological trajectories, Emergence, Citation networks, Innovation dynamics, Path dependence, Biogas

JEL classification: B52, O33, C38, Q55

---

1. Introduction

Detecting technological trajectories matters because innovation policy and firm strategy allocate resources under uncertainty: decisions about missions, R&D portfolios, standards, and infrastructure are made while knowledge bases are still forming and competing alternatives coexist. In evolutionary settings, small initial differences can be amplified through cumulative causation, producing path dependence (David 1985) and potential technological lock-in (Arthur 1990). Empirically tracing trajectories, therefore, helps (i) identify promising lines of search, (ii) monitor convergence that may signal premature closure, and (iii) detect divergence that can open transition pathways. Citation networks provide a scalable empirical window into these processes because citations record actors' choices about which knowledge to build on, and changes in network communities reveal the ability of ideas to attract new researchers and firms.

Empirically tracing these trajectories is central to the evolutionary account of how technologies change. Technological change is the engine of capitalism (Schumpeter 1942, Chapter 7), an evolving system shaped by complex and dynamic relations rather than one converging to equilibrium (Metcalfe 1998). In the evolutionary theory of economic change, this system runs on a small set of population mechanisms: heterogeneous agents and routines generate *variety*, a selection environment retains some variants and discards others, and successful variants are inherited and reproduced (Nelson and Winter 1982; Metcalfe 1998). Innovation is therefore better understood as a continuum unfolding over time (Witt 1998; Dopfer 2001), characterized by coevolution among multiple domains within socioeconomic systems (Almudi and Fatas-Villafranca 2021). Within this view, technical change unfolds through paradigms and trajectories (Dosi 1982): paradigms define problem framings and solution principles, while trajectories represent feasible paths within a given techno-scientific pattern. Paradigms are uncertain *ex ante*, and trajectories exhibit directionality and trade-offs without guaranteeing predictability.

Despite the centrality of trajectories in evolutionary economics, few empirical instruments observe these population mechanisms directly in the evolving structure of knowledge. Verspagen (2007) established the program of mapping technological trajectories as structures within citation networks, extracting the backbone of a field as a main path through the citation graph. This route-based approach, *Main Path Analysis* (MPA), identifies influential

lines of descent by following citation chains (Hummon and Dereian 1989; Batagelj 2003; Verspagen 2007) and has since been extended in many directions (Liu and Lu 2012; Kim and Shin 2018; Masago 2022). Route extraction answers a specific question, which citations form the dominant line of descent, but it gives limited leverage on the population as a whole: how variety is generated when trajectories split, how selection resolves when alternatives compete, and how retention operates when communities merge, stall, or decline. In parallel, the literature on mapping emerging research has developed computational tools to detect novelty and growth (Rotolo et al 2015; Carley et al 2017; Xu et al 2019), yet usually without an explicit link to the evolutionary concepts of paradigms, coevolution, and path dependence that explain why those signals arise.

This paper proposes an unsupervised computational model that integrates trajectory analysis with emerging-technology mapping to address this gap. The method applies a deterministic trunk decomposition to cumulative clusterings of temporal citation networks of scientific articles or patents to detect system-level trajectories and to classify their dynamics into four states that name evolutionary mechanisms: emergence (the generation of variety as new trajectories branch from a paradigm), convergence (selection and retention as competing lines resolve into a surviving lineage), divergence (variety retained under unresolved selection), and dormancy (selective exhaustion or extinction). Reading these states off one decomposition turns abstract population mechanisms into observable, comparable quantities.

We illustrate the approach with biogas, a domain particularly relevant to innovation audiences because it sits at the intersection of the energy transition, waste management, and circular bioeconomy strategies (Manikandan et al 2023; Martín-Espejo et al 2022). Biogas innovation also exhibits multiple technological routes (e.g., anaerobic digestion configurations, biomethane upgrading and purification, and new complementary combinations), and its diffusion depends on institutional contexts, infrastructure, and environmental regulation.

What we find. In the biogas application, we analyze 77,347 scientific documents and identify a strongly core–periphery structure: a citation network of 40,822 documents (about 53% of the corpus) whose giant connected component, 40,300 documents, is organized into nineteen research communities. We document rapid growth in biogas research since 2010, followed by a recent post-pandemic stabilization, consistent with a field moving toward paradigmatic maturity. The thematic and trajectory evidence also indicates

a long-run shift from a 1990s agricultural and rural experimentation phase toward a 2010s stage of industrialization and standardization (pretreatment protocols, upgrading to biomethane, and digestate valorization) and, more recently, a 2020s inflection toward methods intensification and systems integration. Importantly, stable cores coexist with higher-variation niches where new combinations are forming, suggesting continuing scope for reconfiguration rather than a fully locked-in regime.

This paper contributes to evolutionary economics in three ways. First, it extends the citation-network approach to technological trajectories (Hummon and Dereian 1989; Verspagen 2007) from the dominant route to a system-level decomposition in which every cluster–year node belongs to exactly one lineage, so the detected trajectories partition a field rather than tracing one path through it. Second, it operationalizes the core mechanisms of evolutionary change—variety, selection, retention, and path dependence—as measurable trajectory states, bridging the structural vocabulary of community evolution with the indicator-based diagnostics of the emergence literature. Third, it provides an open implementation in the R (R Core Team 2025) package `birddog` (Souza 2026), enabling replication and extension of the proposed methodology.

2. Theoretical framework

2.1. Technological change as an evolutionary process

Economic systems divide between those that tend toward equilibrium and those governed by perpetual evolution (Metcalfe 1998). The Schumpeterian view places technological change in the second category, as the force that animates capitalism (Schumpeter 1942, Chapter 7): innovations are not discrete events but parts of temporally continuous processes (Witt 1998; Dopfer 2001), marked by coevolution, in which innovative practices, reproductive structures, and patterns of choice transform in reciprocal interdependence (Volberda and Lewin 2003; Almudi and Fatas-Villafranca 2021). Against growth theories that treat technical progress as exogenous (Solow 1956; Romer 1986), the neo-Schumpeterian school (Nelson and Winter 1982; Dosi 1988; Rosenberg 1982) runs the economy on a small set of population mechanisms: heterogeneous agents and routines generate *variety*, a selection environment retains some variants and discards others, and successful variants are inherited and reproduced. The methodological consequence is

direct. The relevant empirical object is a population of competing and recombining knowledge communities observed as it changes, not an equilibrium configuration and not a single representative lineage.

2.2. Paradigms, trajectories, and path dependence

The conceptual apparatus for reading that population comes from Dosi (1982). A *technological paradigm* is a pattern of solutions to selected technological problems, grounded in principles drawn from the natural sciences and in specific material technologies; it defines which problems count and which directions of search are admissible. A *technological trajectory* is the movement of multidimensional trade-offs among the variables the paradigm treats as relevant: the feasible path within a techno-scientific pattern. The distinction separates radical innovation, which tends to break an established paradigm, from incremental innovation, which is continuity along a trajectory. A future paradigm is uncertain *ex ante*; a trajectory has direction without predictability. Time imprints irreversibility on the process: small early differences are amplified through cumulative causation into path dependence (David 1985) and, under increasing returns to adoption, into technological lock-in (Arthur 1990). The temporal shape of the process is the familiar succession of S-curves (Andersen 1999; Ehrnberg 1995): a gestation that can last years, rapid growth on entry into a wider market, a proliferation of products and entrants, and maturity as performance limits and saturation approach (Priestley et al 2020). These phases are what a trajectory detector must be able to see.

2.3. Micro-foundations: priority, routines, and the citation record

The population dynamics rest on identifiable micro-level incentives. In research settings, the reward structure of science channels effort through priority races: the first to communicate an advance obtains recognition through eponymy, prizes, and publication, while close seconds receive little, a winner-take-all contest that steers search toward some lines rather than others (Stephan 1996). At the organizational level these incentives materialize through routines, coordinated patterns of action that are self-reinforcing in stable environments and become barriers when a paradigm shift demands reconfiguration; search is the effort to modify or replace them (Nelson and Winter 1982; Becker 2008). These micro commitments leave a public trace. A citation records an author's choice about which knowledge to build on; aggregated over a field, citation choices form communities at the meso level;

followed over time, the growth of a community measures the *power of attraction* of a line of research, its capacity to recruit new researchers and resources. The method of Section 4 operates exactly on this trace, which is what lets it connect micro-level choices, meso-level community formation, and macro-level paradigm consolidation within one representation.

2.4. Consolidation: dominant designs, lock-in, and sectoral regimes

Consolidation has an observable signature. A *dominant design* marks the moment a particular technical solution becomes the sectoral standard; once established, it shifts competition away from architectural variation toward cost, scale, and incremental performance, and it raises barriers to subsequent radical innovation (Utterback 1996). Positive feedback intensifies the process: rising adoption increases a technology's relative value, attracting further adoption in a self-reinforcing cycle that can lock in a solution even when superior alternatives exist (Arthur 1990). Detecting the early stages of such consolidation gives policy a temporal window for intervention before lock-in closes it. The pace and shape of these dynamics differ by sector, from supplier-dominated to science-based patterns (Pavitt 1984), and by technological regime, from the turbulent, high-entry Schumpeter Mark I pattern to the concentrated, stable Mark II (Malerba and Orsenigo 1996). Indicators intended for comparative analysis must therefore read on signals that remain meaningful across regimes.

2.5. The emergence pillars and the states as observables

A separate literature, reviewed in Section 3, developed computational detectors of emerging science and technology. Its conceptual consolidation supplies the measurement footing this paper builds on: emerging technologies are characterized by radical novelty, relatively fast growth, coherence, prominent impact, and uncertainty (Rotolo et al 2015), refined into four operational pillars, *novelty*, *growth*, *community*, and *persistence* (Carley et al 2017), the basis of the FUSE emergence-scoring program (Garner et al 2017; Carley et al 2018). Table 1 fixes the mapping that the rest of the paper operationalizes: each of the four states is a property of a trajectory, never of a static group; it is read on one of two axes according to trajectory type; and it names an evolutionary mechanism while resting on measurable pillars.

Table 1: Trajectory states as evolutionary observables. States are properties of trajectories (flow), never of groups (stock); central trajectories read on a growth axis, absorbed trajectories on a destination axis. Dormancy appears on both axes with distinct meanings.

State	Type	Evolutionary mechanism	Primary signal
Emergence	central	variety generation: a young line gains power of attraction	high growth and novelty
Maturity	central	retention: an established line on its S-curve plateau	slowed spine growth
Dormancy	central	selective exhaustion of a living line	stalling or shrinking spine
Convergence	absorbed	selection and retention: competing lines resolve into a surviving lineage	low destination entropy
Divergence	absorbed	variety retained under unresolved selection	high destination entropy
Dormancy	absorbed	extinction: the line ends with no successor, or its cohort drops out of every living community	no successor above the Jaccard floor, or high dormant share

3. Literature review

Empirical work on trajectories and emergence divides into four families: route-based reconstruction in citation networks, structure-based tracking of evolving communities, content- and indicator-based proxies, and emergence-oriented detection for foresight. Table 2 summarizes them; the subsections discuss them in turn, treating content-based and indicator-based proxies separately.

Table 2: Four families of trajectory and emergence detection, their unit of analysis, and the blind spot each leaves for the systemic question.

Family	Exemplars	Unit of analysis	Blind spot
Route-based (MPA and variants)	Hummon and Dereian (1989); Verspagen (2007); Liu and Lu (2012); Kim and Shin (2018); Masago (2022)	the citation route	the population around the route: branching appears as path divergence and fusion, not as communities recruiting, growing, merging, and declining
Structure-based tracking	Shibata et al (2008); Shibata et al (2011); Chen et al (2012); Ding et al (2024)	communities re-clustered per snapshot, or document-level curves	community identity rests on heuristic matching across snapshots; transition labels rather than lineage life-course states
Content- and indicator-based proxies	Lee et al (2020); Liu et al (2020); Zhang et al (2023); Corrocher et al (2021); Françoso et al (2024); Kayal and Waters (1999)	topics, regimes, field aggregates	the trajectory as an explicit traceable object; citations absent or collapsed into aggregates and symmetric similarities
Emergence mapping	Small et al (2014); Carley et al (2018); Porter et al (2020)	emergent topics and terms (ranked snapshots)	evolutionary concepts motivate definitions but are not operationalized; dynamics enter only as inputs to a score

3.1. Citation-network approaches

A central family of approaches operationalizes trajectories through citation networks, based on the premise that citations record choices about which knowledge to build on. Verspagen (2007) established this program by mapping technological trajectories as main paths through a patent citation network, tracing the backbone of fuel-cell research. *Main Path Analysis* (MPA) extracts these influential routes by weighting edges with traversal counts and following citation chains (Hummon and Dereian 1989; Batagelj 2003; Verspagen 2007). Subsequent work extends MPA in multiple directions: identifying junctures and derivative paths (Kim and Shin 2018), periodizing the evolution of a field using global main paths (Kim et al 2020), recover-

ing dominant knowledge through genetic backward–forward logic (Park and Magee 2017), incorporating persistence-based lineage indicators (Martinelli and Nomaler 2014), scaling route extraction via topological skeletonization (Masago 2022), and refining route extraction by integrating node importance and semantic citation information (Xu et al 2022; Jiang and Liu 2023; Yeo et al 2013). Other variants combine route extraction with complementary indicators to identify strategic positions and interpret technological change, such as junctions and bottlenecks (Fontana et al 2009), clustering and small-world structure (Guan and Shi 2012), and combinations with centrality, specialization, and influence measures (Epicoco 2013; Linares et al 2019; Ferrari et al 2019; Qi et al 2023; Watanabe and Takagi 2021).

MPA-based approaches reconstruct canonical lines of influence and identify the key patents or documents along a route. Their unit of analysis is the route itself, which answers where the dominant line of descent runs. Variants do register branching along the route, as junctures, derivative paths, and the splitting and fusion of main paths (Verspagen 2007; Liu and Lu 2012; Kim and Shin 2018); what branches there is the path itself, not the communities that recruit, grow, merge, and decline around it. A complementary question concerns the population: how the full set of trajectories in a field generates variety, selects among competing lines, and retains or abandons them. Addressing it calls for a system-level object in which every document belongs to a lineage, rather than a single route through the network.

3.2. Structure-based tracking of evolving communities

A second family focuses on the evolution of field structure rather than single routes, tracking communities across successive snapshots of the network. Shibata et al (2008) detect emerging research fronts by repeatedly clustering a growing cumulative citation network and linking clusters across years through the papers they share, an approach applied to regenerative medicine in Shibata et al (2011). Community transition analysis and timeline visualizations extend the tracking to sliding windows, revealing births, deaths, branching, and merging patterns in technological domains (Chen et al 2012). Other approaches shift the target from retrospective reconstruction to forecasting, treating trajectories as predictable citation dynamics at the patent level (Ding et al 2024). Predictive models serve technology intelligence, but they typically operationalize "trajectory" as micro-level citation curves, rather than as meso-level reconfiguration of knowledge communities.

3.3. Content-based approaches

A related family of alternatives is content-led. Topic-modeling approaches map shifts in themes across time and geography (Lee et al 2020), hybrid models combine citation information with topic structure to track dynamic research fronts (Liu et al 2020), and the community detection and topological measures of the structure-based family have been paired with structural topic modelling to delineate a field's groups and emerging themes (Maria et al 2023). Other approaches use link prediction or neural network models to quantify novel combinations or infer evolutionary patterns, often without representing trajectories as explicit evolving communities (Sun et al 2022; Zhang et al 2023). These tools capture changes in vocabulary and thematic emphasis, but citations are either absent or collapsed into symmetric similarities, which limits the leverage for diagnosing structural events such as convergence, divergence, and dormancy.

3.4. Indicators and econometric models

Beyond networks, trajectory dynamics are also proxied through indicators and econometric models that characterize technological regimes and opportunities. For example, econometric analyses examine relationships between collaboration intensity, funding, and publication novelty (Shin et al 2022). Regime measurement using opportunity, cumulativeness, originality, and complexity indicators (Corrocher et al 2021) and multilevel models of embeddedness in knowledge spaces (e.g., coreness and betweenness) (Françoso et al 2024) offer further views on cumulativeness and variety. Time-to-impact indicators such as technology cycle time (TCT) measure the pace of knowledge turnover (Kayal and Waters 1999), and growth-curve models diagnose diffusion stages and potential saturation (Li and Xu 2022). These approaches are often interpretable for policy purposes, but they typically do not treat trajectories as explicit objects that can be traced through the evolving structure of knowledge communities.

3.5. Emergence mapping and science–technology intelligence

A parallel body of literature detects emerging topics and technologies for science and technology intelligence and foresight. This work is motivated by the need to identify early signals of novelty under uncertainty and to support strategic investment and policy prioritization. Conceptually, emerging technologies are commonly characterized by attributes such as novelty, relatively rapid growth, coherence, prominent impact, and uncertainty (Rotolo

et al 2015), later refined into operational dimensions emphasizing novelty, persistence, community, and growth (Carley et al 2017). Methodologically, emergence studies span large-scale topic detection and "research front" discovery (Small et al 2014), as well as the community-tracking detectors of the structure-based family (Shibata et al 2008, 2011). Quantitative indicators have been proposed to score emergence and track rapidly growing areas (Carley et al 2018) and comparatively evaluated across technologies (Ranaei et al 2019); the field's development has itself been traced through main-path analysis of its citation network (Xu et al 2019) and benchmarked in community-wide prediction contests (Porter et al 2020).

The technological-trajectory and emergence literatures are rarely connected, despite their methodological similarities; both read the structure of a field from its citation record, much as innovation studies has mapped its own structure from practitioner surveys (Fagerberg and Verspagen 2009). Emergence research has typically been driven by computational opportunity and science–technology intelligence needs, rather than by a theoretical commitment to evolutionary economics. A related divide separates the data domains: indicator and econometric studies such as Corrocher et al (2021), Françoso et al (2024), Kayal and Waters (1999), and Li and Xu (2022) work on patents, while emergence mapping has concentrated on scientific fields and publication-based signals. This separation limits cumulative theorizing about how novelty, growth, and community formation relate to directionality and structural reconfiguration in innovation systems.

Taken together, the literature offers four partial views: route-based reconstruction (MPA and variants), structure-based tracking of evolving communities, content- and indicator-based proxies, and emergence-oriented detection for foresight. Our framework builds on these by treating trajectories as evolving communities in temporal citation networks and by applying one procedure to both scientific and patent documents. It classifies each trajectory into a state that names an evolutionary mechanism: emergence (variety generation), convergence (selection and retention), divergence (retained variety), and dormancy (selective exhaustion). This renders comparable dynamics across science and technology fronts within a single evolutionary vocabulary.

4. Methodology

Our methodological goal is to detect and characterize technological and scientific trajectories as system-level objects that evolve over time, using

temporal citation networks of either scientific publications or patents. The method rests on one organizing distinction. A group at a given year is a static *stock*: a cluster of documents in a single cumulative snapshot, with no past or future inside itself. A trajectory is a temporal *flow*: a path of cluster–year nodes across the whole period. Emergence, convergence, divergence, and dormancy are properties of flow, never of a stock, so asking for the "emergence of a group" is a category error; the operative question is which trajectories emerge, and into which final group they flow.

Building on prior work on time-sliced clustering and trajectory mapping in citation networks (Shibata et al 2008, 2011; Chen et al 2012), we implement a reproducible pipeline (Figure 1) that combines (i) network construction and community detection, (ii) reverse cumulative temporal clustering, (iii) a deterministic decomposition of the resulting temporal directed acyclic graph (DAG) into node-disjoint trajectories, and (iv) classification of each trajectory's dynamics by trajectory type. Unlike the route-based extraction of Main Path Analysis, the decomposition assigns every cluster–year node to exactly one lineage, so the detected trajectories partition the field rather than tracing a single dominant route. Elements of this pipeline have been previously applied by the authors in related empirical studies (Souza 2016; Souza et al 2021, 2022; Maria et al 2023; Biazatti et al 2024; Miranda et al 2025; Felizardo et al 2025).

We call this procedure the birddog method, after the open-source R package that implements it (Souza 2026). The method described here is the one implemented from version 2.0.0, whose trunk decomposition replaced the per-group detector of earlier versions; the package also bundles the optional extension modules of Steps 7 and 10, which are not part of the method.

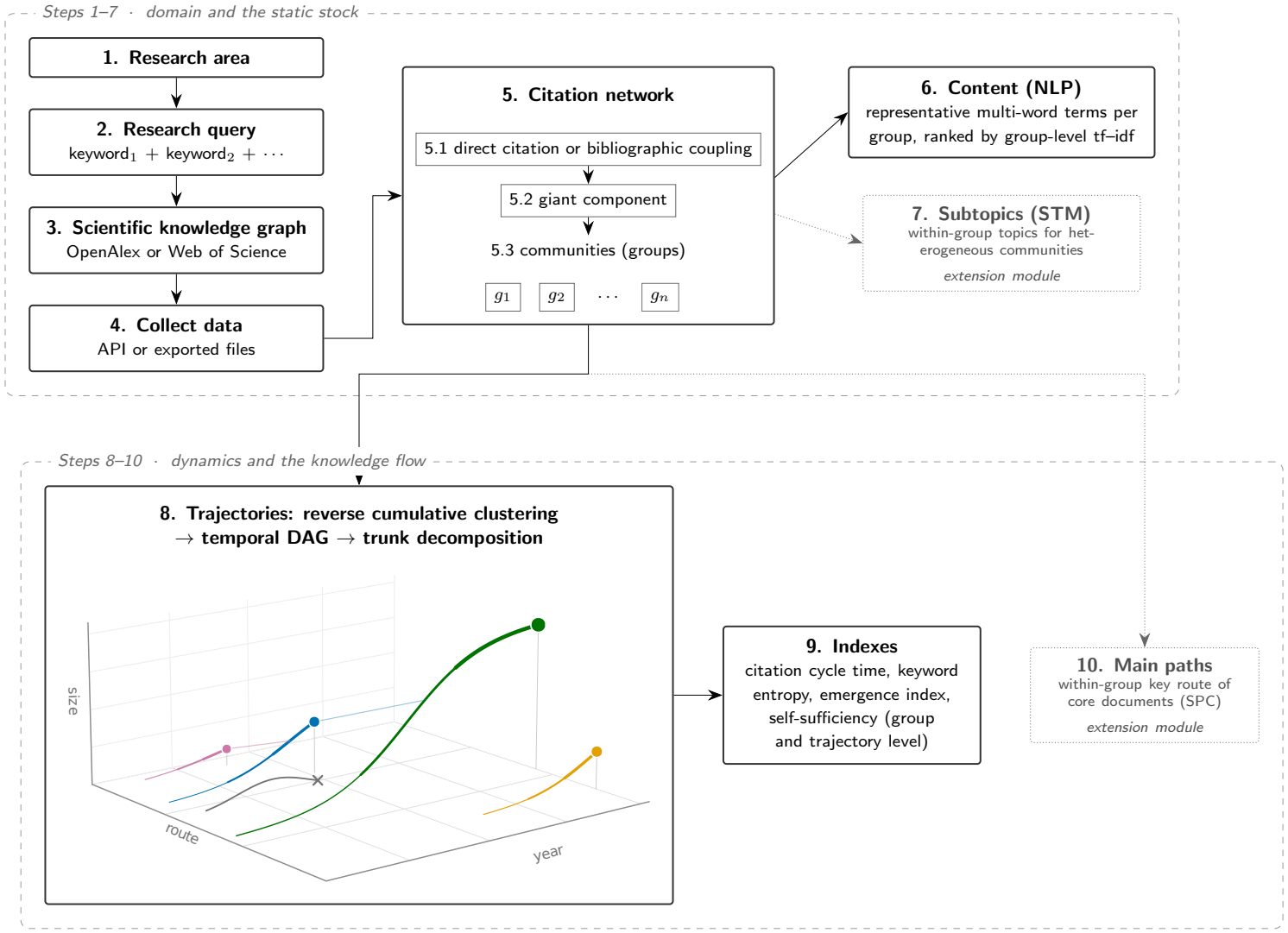


Figure 1: Workflow of the `birddog` trajectory-detection pipeline; the procedure applies to scientific publications or patents. Step 8 sketches the flow layer in the package's three-dimensional trajectory view (publication year × route × cumulative size): lineages grow as competing S-curves, a dominant central matures, a young central emerges, tributaries end and hand their papers to the dominant spine (absorption), and an extinct lineage dries up. Structural Topic Modeling (Step 7) and within-group main paths (Step 10) are optional extension modules, not parts of the core method.

The workflow in Figure 1 is implemented as ten steps. Steps 1–7 define and map the research domain (the static stock), while Steps 8–10 quantify change and summarize the knowledge flow (the dynamic trajectories). Two of the ten are optional extension modules rather than parts of the core method: Structural Topic Modeling (Step 7) and within-group main paths (Step 10).

1. **Research area**: select the domain to be investigated.
2. **Research query**: construct the search query for the domain.
3. **Scientific knowledge graph**: choose the data source (e.g., Web of Science versus OpenAlex).
4. **Collect data**: import via API or via exported files (e.g., BibTeX/CSV).
5. **Citation network**: choose direct citation or bibliographic coupling; extract the giant component; detect groups.
6. **Content NLP**: extract representative terms for each group.
7. **STM** (extension module): when groups are heterogeneous, estimate within-group subtopics.
8. **Trajectory**: build the temporal DAG from the reverse cumulative clusterings and decompose it into central, absorbed, and extinct trajectories; classify each by type.
9. **Indexes**: compute renewal and heterogeneity (citation cycle time and entropy) at the group and trajectory levels, and the trajectory-level emergence and self-sufficiency indicators.
10. **Main paths** (extension module): extract within-group main paths to summarize core citation routes.

Table 3 fixes the vocabulary used throughout the paper; each term is defined operationally in the subsections below.

Table 3: Canonical vocabulary of the method.

Term	Definition
stock	a group at a year (one cumulative clustering snapshot); static, carries no dynamic state
flow	a trajectory: a temporal path of cluster–year nodes; carries all dynamics
trajectory / lineage	a node-disjoint chain on the temporal DAG (trunk decomposition)
central trajectory	the spine of a final community, reaching the last year; one per final group
absorbed trajectory	a tributary that merges into another trajectory before the final year
extinct lineage	a chain whose head has no successor above the Jaccard floor; its documents can still reach final communities
absorption / handoff	a tributary merges into its absorber at the absorption year; transitive
confluence tree	per final community, the central trunk plus its tributaries at every depth
terminal cohort	the documents in the last node of a trajectory
cumulative clustering	cumulative networks re-clustered every year; a document never leaves the network
emergence	a young, fast-growing central (novelty and growth)
maturity	the growth plateau of a central (S-curve), a trajectory-level state
dormancy	a central whose spine stalls or shrinks; a tributary that goes extinct or whose cohort mostly vanishes
convergence	an absorbed cohort concentrated in one destination (low destination entropy)
divergence	an absorbed cohort scattered across destinations (high destination entropy)
diversification	rising keyword entropy within a group; a content property, not a state
formation	the tributaries that fed a target trajectory (inflow), distinct from its destination (outflow)
self-sufficiency	the endogenous share of a central's spine (Equation 5)

4.1. Data, scope, and unit of analysis

The first step of the methodology is to define the focal research domain to be investigated. The second step is to design a search strategy (keywords and Boolean expressions) that best represents that domain. In both steps, domain expertise helps improve precision.

For scientific publications, we rely on scientific knowledge graphs (Priem

et al 2022), large bibliographic databases such as OpenAlex and Web of Science. These platforms provide standardized metadata and normalized references, which enable the construction of citation-based networks (direct citations and bibliographic coupling) (Batagelj and Cerinšek 2013).

The pipeline is agnostic to the document type: it requires only documents carrying a publication year, a title, an abstract, and normalized citation relations among them, so patent corpora satisfy its requirements as naturally as scientific ones. Patent data can be collected from sources such as Derwent Innovation and Google Patents, under the same discipline of a query-based, transparent corpus definition. Patent-specific retrieval and parsing algorithms are under development in `birddog`: the version current at the time of writing implements OpenAlex and Web of Science for scientific documents, and patent support is planned for a subsequent release.

4.2. Static mapping of the knowledge structure

Network construction. We first construct a citation network $G = (V, E)$ for the full document collection, where nodes are documents and edges represent knowledge relations derived from citations. Three representations are commonly used: direct citation, bibliographic coupling, and co-citation. Because our focus is on identifying what is new (trends) rather than emphasizing established foundations, we rely on the first two options. In direct citation, two documents are connected when one cites the other, the technique at the core of Garfield's historiography (Garfield 2004); in bibliographic coupling, two documents are connected when they share at least one reference (Kessler 1963). There is no consensus in the literature on which network representation is best for detecting trajectories or emergence (Shibata et al 2009; Boyack and Klavans 2010, 2020). Based on the proposing author's prior applications, in fields with strong heredity—where knowledge accumulation is concise and progressive over time—direct citation captures most relevant relationships. Bibliographic coupling is particularly useful for incorporating the most recent scientific articles and patents and for aggregating documents in more dispersed knowledge fields.

Component filtering and community detection. As in prior network-based mapping, we retain the giant connected component to focus on the dominant knowledge structure and to exclude weakly connected residues that are unlikely to belong to the focal domain (Kitsak et al 2018). When the network contains multiple sizeable components that are substantively related to the

domain, the fragmentation can itself be informative, suggesting a domain that is still consolidating and exhibiting emergence-related variety rather than a single stabilized core. In a bibliometric evaluation of the Brazilian bioenergy program BIOEN, the second-largest component, disconnected by citation from the giant component, concentrated research on ethanol-fed fuel cells, one of the most innovative lines of the corpus (Souza et al 2022); secondary components therefore deserve inspection before being discarded. We then detect communities through citation similarity using modularity-based algorithms (Louvain or Leiden) to identify densely connected clusters (Blondel et al 2008; Traag et al 2019).

We assess cluster coherence by inspecting hub roles (Z_iP_i metrics (Guimera and Amaral 2005)), prestige measures for documents within the network, together with the consistency of their research topics; cohesive groups have hubs with compatible research agendas.

Group interpretation (terms, topics, hubs). Given the partition of the publications into groups $(g_1, \ldots, g_n)$, we treat these communities as fields of the main domain. We characterize the content of each group by extracting representative multi-word terms (Wijffels 2025) from titles and abstracts and ranking them using tf–idf computed at the group level. This weighting favors terms that are salient within a group while remaining relatively distinctive across groups (Robertson 2004).

When a group is large and internally heterogeneous, as in Maria et al (2023), term lists can be complemented with within-group subtopic discovery through Structural Topic Modeling (STM), the optional extension module of Step 7 (Roberts et al 2019): a soft, text-based decomposition, with documents loading on multiple subtopics, laid over the hard citation-based partition. STM reads the stock, never the flow: it refines the interpretation of a group at a snapshot, and no trajectory-level result depends on it. The biogas application does not exercise this module.

To support interpretation and comparison across groups, we report both descriptive and positional attributes: group size, growth rate, average publication year, country and author predominance, and hub roles based on within-group connectivity and cross-groups bridging (prestige Z_iP_i) (Guimera and Amaral 2005).

Inter-group influence. Influence between communities is directional and asymmetric, so the symmetric similarity used to build the network is the wrong

tool for it. We lift every document-to-document citation onto its pair of (citing group, cited group) to obtain the cross-citation matrix C_{ij} ("group i cites group j"). The diagonal C_{ii} records intra-group cohesion and is dropped before any normalization, so the out-strength $o_i = \sum_k C_{ik}$, the in-strength $\iota_j = \sum_k C_{kj}$, and the total m count inter-group flows only; the balances below are unchanged by this choice, since the diagonal cancels. Each channel $i \to j$ is normalized four ways:

$$\underbrace{\delta_{ij} = \frac{C_{ij}}{o_i}}_{\text{debt}}, \qquad \underbrace{\alpha_{ij} = \frac{C_{ij}}{\iota_j}}_{\text{audience}}, \qquad \underbrace{s_{ij} = \frac{C_{ij}}{\sqrt{o_i\,\iota_j}}}_{\text{Salton}}, \qquad \underbrace{\rho_{ij} = \frac{C_{ij}}{o_i\,\iota_j/m}}_{\text{surprise}}. \tag{1}$$

Debt is the share of i's citations owed to j; audience is the share of j's recognition supplied by i; the Salton index (Salton and McGill 1983) is a size-free strength of the directed channel; surprise compares the observed flow with a directed configuration null (Fortunato 2010), with values above one marking over-represented channels. Netting reciprocal flows gives the net channels $\nu_{ij} = C_{ij} - C_{ji}$, and each group's balance $\beta_i = \iota_i - o_i$ assigns its role: *source* ($\beta_i > 0$, foundational, a net knowledge exporter), *sink* ($\beta_i < 0$, a frontier consumer), or *broker* ($\beta_i \approx 0$). The net-influence edges form the directional spine of who leads whom in the field; the application reads the matrix through the surprise channels and this net spine, with debt, audience, and Salton available as further readings of the same flows.

The partition obtained at the final year is the *stock*: a static photograph of the field in which each document belongs to exactly one group. Because a snapshot carries no temporal information, the dynamic states (emergence, convergence, divergence, and dormancy) cannot be read from it; they are properties of the flow layer developed next.

4.3. Dynamic tracking of trajectories

Reverse cumulative clustering. To represent how communities evolve, we generate a sequence of cumulative networks and cluster each one in reverse temporal order. For each year $t_k \in [t_1, t_T]$, we build a cumulative subnetwork G_k that contains every document published up to t_k and compute a partition with the same algorithm (Louvain or Leiden) and fixed settings across years. The design is cumulative rather than time-sliced, with one decisive consequence: once published, a document is present in every later cumulative network, so no document is ever removed and, by the final year, every

document belongs to some group. Two cautions follow. First, the group labels ($c1g1, c1g2, \dots$) are size ranks within a year, not stable identities: the same community drifts across labels as it grows or shrinks relative to the others, so tracking is always by document identity, never by label. Second, this membership over time, recorded as one row per node–document, is the single input to the trajectory detector.

The temporal DAG. At each year, a group is a stock: a set of documents densely connected by citation similarity in that year's cumulative network, with no history of its own. History enters when the stocks are connected. We model group evolution as a temporal DAG $H = (\mathcal{V}, \mathcal{E})$ whose nodes are group–year pairs, admitted only when large enough ($|D(v)| \geq n_{\min}$, a node size floor), and whose edges join nodes in consecutive years when documents migrate from one group to the other: since clustering is cumulative, the same documents are re-partitioned every year, and an edge records how much of a group's membership reappears in a group of the following year. Edge weight is the Jaccard overlap of the two nodes' document sets $D(\cdot)$:

$$J(u, v) = \frac{|D(u) \cap D(v)|}{|D(u) \cup D(v)|}, \qquad \mathrm{yr}(v) = \mathrm{yr}(u) + 1. \tag{2}$$

An edge is retained when its weight clears a floor τ_J and the source keeps only its strongest few successors; because $\tau_J > 0$, every retained edge carries at least one shared document. Figure 2 shows the object on a small conceptual example; in the biogas application, the same construction yields 522 group–year nodes and 682 migration edges spanning 1989–2025 (Section 5.4). The DAG is the substrate of trajectory detection: on it, a trajectory is a path, and every question about trajectories becomes a question about paths. The annual clustering-and-linking design is related to prior approaches in the emergence literature (Shibata et al 2008, 2011), but here it feeds a system-level decomposition rather than only a visualization.

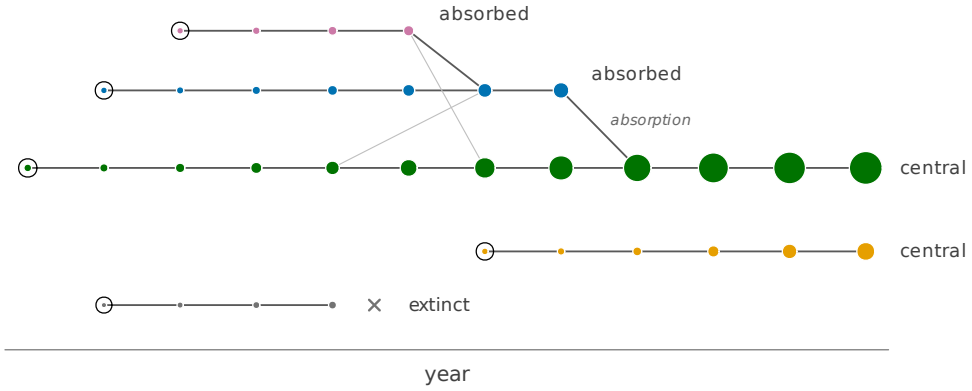


Figure 2: The substrate of trajectory detection, drawn as a conceptual example. Each ball is a group–year node, one community observed in one cumulative snapshot, sized by its document count; each edge joins consecutive years and records document migration, weighted by the Jaccard overlap of the two memberships; faint edges are the admitted alternatives that rank below the heaviest successor, the forks where detectors can disagree. A trajectory is a path on this graph, and the five lineages exhibit every fate the classification admits: the dominant central (green) and a late-emerging central (orange) reach the final year; the blue tributary ends and hands its documents to the central (absorption), after itself receiving an earlier pink tributary, making the absorption transitive; the gray lineage keeps no admitted successor and goes extinct (×). Black rings mark births, nodes with no admitted predecessor. The same cast, in the same colours, grows as competing S-curves in the flow panel of Figure 1 (Step 8).

The decomposition problem. The detection task is stated on the DAG alone, before any algorithm: partition the group–year nodes into node-disjoint temporal chains, each read as one lineage. Given any such partition, the classification of a chain is mechanical:

- a chain whose latest node reaches the final year is **central**: the spine of a final group, one per group;
- a chain that stops earlier is **absorbed**: its head still has a heaviest successor outside its own chain, and the chain owning that successor is its absorber, with the merge dated to the following year;
- a chain whose head has no successor above the threshold is **extinct**: it ends before reaching any final group and has no absorber, although the documents of its last node can still be tracked to the final year.

Absorption is transitive: a tributary can be absorbed by another tributary that is itself later absorbed. The absorber relation therefore organizes the absorbed chains into a forest rooted at the centrals, one *confluence tree* per final group, with the central as the trunk and tributaries entering at every depth. A many-way confluence into a final group is thus a tree of pairwise merges resolving onto one spine, not a star of direct tributaries.

Braid: trunk decomposition by mutual dominance. The canonical detector, *braid*, solves the problem with strictly local information. For each node we take its single heaviest successor $\sigma(\cdot)$ and its single heaviest predecessor $\pi(\cdot)$ (weight first, shared-document count as tie-break). An edge (u, v) is a *trunk edge* under mutual dominance, that is, when v is the heaviest successor of u and u is the heaviest predecessor of v:

$$T = \big\{(u, v) \in \mathcal{E} : \sigma(u) = v \ \wedge \ \pi(v) = u\big\}. \tag{3}$$

Mutual dominance gives each node at most one trunk-in and one trunk-out edge, so the trunk edges partition the DAG into node-disjoint directed chains, each a lineage. The criterion is itself classical, appearing as reciprocal nearest neighbours in clustering and as the locally dominant edges of approximate weighted matching (Preis 1999); in community tracking, a bidirectional inclusion measure underlies Group Evolution Discovery (Bródka et al 2013), which labels the transition between two successive groups, whereas braid uses the mutual-heaviest rule to partition the whole DAG into node-disjoint chains.

Channel: global cheapest-path routing. A global sibling, *channel*, routes each node along its cheapest path to a final-year community. Each Jaccard weight becomes an additive cost $c(u, v) = -\log w(u, v) \geq 0$, so the minimum-cost route maximizes the product of the Jaccard weights, the chain of highest geometric-mean coherence. A single backward sweep in descending year computes, for every node, the cost of its cheapest route to a final node and the successor that route takes,

$$\varphi(v) = \min_{(v,u)\in\mathcal{E}} \big[c(v, u) + \varphi(u)\big], \qquad \beta(v) = \arg\min_{(v,u)\in\mathcal{E}} \big[c(v, u) + \varphi(u)\big], \tag{4}$$

with $\varphi(v) = 0$ on final-year nodes; a dead end keeps $\varphi(v) = \infty$ and its forward pointer falls back to the locally heaviest successor $\sigma(v)$. Each

final-year node roots one in-tree of routed pointers, the watershed of its community. A second sweep, in ascending year, cuts the forest into chains: where several routed predecessors converge on a node, the chain continues through the one whose accumulated cost from a birth is cheapest, and the others end there; each central is therefore the cheapest birth-to-final chain of its watershed. Classification follows the same rules as braid, with the routed successor in place of σ. Channel also reports each chain's latest-departure year, the most recent birth that reaches the chain's tail by a time-respecting path, adopting the latest-departure notion of temporal graph theory (Wu et al 2014). The two detectors agree on most networks and part only at forks, where a node's locally heaviest edge points to one destination while its strongest downstream path leads to another; braid is the default because it commits to nothing beyond one hop and is robust to long-range noise, and the content-coherence validation described below adjudicates between them. Algorithms 1 and 2 summarize the two detectors.

Algorithm 1 Braid: trunk decomposition by mutual dominance

Require: temporal DAG $H = (\mathcal{V}, \mathcal{E})$ with Jaccard weights w; final year Y

for all $v \in \mathcal{V}$ **do**
  $\sigma(v) \leftarrow$ heaviest successor of v; $\pi(v) \leftarrow$ heaviest predecessor of v (ties by shared documents)
end for
$T \leftarrow \{(u, v) \in \mathcal{E} : \sigma(u) = v \ \wedge \ \pi(v) = u\}$ (trunk edges, mutual dominance)
decompose $(\mathcal{V}, T)$ into maximal chains (each node has at most one trunk-in and one trunk-out)
for all chains C with head h (latest node) **do**
  if $\mathrm{yr}(h) = Y$ **then**
    C is **central**
  else if h has a successor in $\mathcal{E}$ **then**
    C is **absorbed** into the chain owning $\sigma(h)$, at year $\mathrm{yr}(h) + 1$
  else
    C is **extinct**
  end if
end for
return chains and the confluence forest induced by the absorber relation

Algorithm 2 Channel: global cheapest-path routing

Require: temporal DAG $H = (\mathcal{V}, \mathcal{E})$ with costs $c(u, v) = -\log w(u, v)$; final year Y

$\varphi(v) \leftarrow 0$ for final-year nodes; $\varphi(v) \leftarrow \infty$ otherwise
for $t = Y - 1$ **down to** t_1 **do**
  for all v with $\mathrm{yr}(v) = t$ **do**
    $\varphi(v) \leftarrow \min_{(v,u)} [c(v, u) + \varphi(u)]$; $\beta(v) \leftarrow \arg\min$ (ties by heavier edge, then shared documents)
  end for
end for
$\beta(v) \leftarrow \sigma(v)$ wherever $\varphi(v) = \infty$ (dead ends: myopic fallback)
cut the β-forest into chains by an ascending-year sweep: at each convergence, the chain continues through the predecessor whose accumulated cost from a birth is cheapest; the other predecessors end their chains there (each central = the cheapest birth→final chain)
classify chains as in Algorithm 1, with β in place of σ
return chains, confluence forest, and each chain's latest-departure year (the most recent birth reaching its tail)

An open family of detectors. Braid and channel are two members of a family that the substrate defines. Any algorithm that partitions the temporal DAG into node-disjoint chains is a trajectory detector in the sense of this paper, and it inherits the classification rules, the confluence forest, and the content-coherence validation of Section 5.9 unchanged. Natural candidates include a maximum-weight path cover, solvable as one bipartite assignment per year pair (Kuhn 1955), event-driven matching in the community-tracking tradition (Greene et al 2010; Dakiche et al 2019), and partial optimal transport between consecutive cohorts (Li et al 2025). Framing detection this way, on a network whose nodes are the groups of each year and whose edges are document migrations to the next, turns trajectory detection into a well-posed graph problem on which new detectors can compete under a fixed evaluation criterion; we regard this substrate, as much as the two detectors shipped with `birddog`, as the methodological contribution.

What "death" means. Because clustering is cumulative, the disappearance of a trajectory is not the disappearance of its papers. A lineage that stops being its own line has not lost its documents; they have been re-clustered and, in nearly every case, absorbed by a lineage still flowing toward the present. A

river that vanishes from the map did not evaporate; it flowed into a larger river, losing its name but not its water. True document death, papers dropping out of every final group, is measured separately as a dormant share; even an extinct chain, which no lineage inherits, can hand its documents on to surviving communities. This qualifies the inheritance from time-sliced designs (Shibata et al 2008, 2011): there a cluster can genuinely vanish between windows, but in a cumulative design the same reading would mislead.

Trajectory states by type. Because a central reaches the present while an absorbed lineage stops, the two answer different questions and take states on two different axes. A central takes a life-cycle phase on a growth-and-novelty axis: **emergence** (young and growing fast), **dormancy** (the spine stalls or shrinks), or **maturity** (established, growth slowed, the plateau of the S-curve). An absorbed lineage takes a terminal fate on the entropy of its terminal cohort's destination. Tracking the documents of the lineage's last node to the final year gives their distribution over the k surviving destination groups; writing p_g for the share of the cohort landing in group g, the normalized Shannon entropy $\mathcal{H} = -\sum_g p_g \log p_g / \log k \in [0, 1]$ measures how concentrated that destination is, with $\mathcal{H} = 0$ when a single destination survives. The fate is, in priority order, **dormancy** (the cohort mostly dropped out, or the lineage is extinct), **convergence** (low entropy: the cohort concentrates coherently in one destination, a merge seen from the absorbed side), or **divergence** (high entropy: the cohort scatters across groups). Dormancy is tested first, so entropy is read only once the cohort has somewhere to go. The term dormancy thus appears on both axes, a living core that lost momentum or a cohort that vanished, which is why a central is never labelled convergence or divergence.

Thresholds are data-driven by default, computed from the field's own distributions: the growth cuts sit one median absolute deviation above and below the centrals' median growth, novelty and destination entropy are split at their medians, and the dormant share keeps a fixed majority rule of 0.5. The resulting states are therefore relative to the field. A median split divides its population in half by construction, so the balance between convergence and divergence says where each tributary sits against the others, not how concentrated destinations are in absolute terms. Fixed cuts can be supplied instead, and the thresholds actually used are stored with the result, so the classification is reproducible and auditable rather than a qualitative judgement.

Reading the flow: destination, formation, self-sufficiency. The states just defined are categorical: they name what happened to a lineage, not how much. Two quantitative layers complete the reading: the three flow readers of this paragraph measure where a lineage's documents come from and where they go, and the indicators of the next score how the lineage behaves along the way, with the emergence index reading as intensity what the emergence state reads as category. *Destination* takes the documents in a trajectory's terminal node (its terminal cohort) and tracks them by identity to the final year, returning the cohort's distribution across final groups, the dormant share (the fraction that genuinely dropped out), and the immediate absorber; the shape of this distribution is what the fate axis classifies. *Formation* is the inverse: the tributaries that fed a target trajectory, its direct children in the confluence tree, with papers partitioned so that no document is double-counted; a central's formation is its community's whole river. *Self-sufficiency* scores how much of a central's spine it grew on its own rather than carried in by tributaries,

$$\text{self_sufficiency}(C) = 1 - \frac{|\,\text{spine papers also carried by any tributary}\,|}{|\,\text{spine papers}\,|} \in [0,1], \tag{5}$$

near 1 for an endogenous community and near 0 for a confluence of other lineages; the numerator counts unique papers shared with the whole tributary subtree, not a sum of handoff weights, which would double-count.

Trajectory indicators. We characterize trajectories with indicators grounded in the emergence literature's operational dimensions of novelty, growth, community, and persistence (Rotolo et al 2015; Carley et al 2017). Knowledge renewal is the citation cycle time (CCT), adapting technology cycle time (Kayal and Waters 1999) to citations: for a node with papers P, each paper's median reference age is taken first, and the node's CCT is the median of those medians,

$$\text{CCT} = \underset{i \in P}{\text{median}}\Big(\underset{r \in R_i}{\text{median}}\,(T_i - T_r)\Big), \tag{6}$$

where R_i are paper i's references and T denotes publication years; the median-of-medians keeps reference-rich papers from dominating, and a low or falling value marks a fast-renewing front. Thematic heterogeneity is Pielou's evenness, the normalized Shannon entropy of a node's keyword distribution

(Shannon 1948; Pielou 1966), rising with diversification and falling with consolidation. Over the living centrals, an outlier-resistant emergence index ranks trajectories by summing winsorized robust z-scores of three pillars: growth (the annualized rate of an exponential fit to the lineage's recent size curve), novelty (the share of the lineage's documents that arrived within a recent window of its timeline), and recruitment (the net documents gained over the lifespan, the power of attraction of Section 2 made countable),

$$\text{emergence}(C) = \sum_{x \in \{\text{growth, novelty, recruitment}\}} \text{clip}_{\pm 3}\left(\frac{x_C - \text{med}(x)}{\text{MAD}(x)}\right), \qquad (7)$$

with the cap at three near the standard robust outlier boundary. Three companions refine the reading: an intensive variant drops the size-carrying recruitment pillar to rank on rate alone; emergence density, the persistence pillar, measures staying power as the share of a lineage's year-steps still growing at the emergence rate (Carley et al 2017); and the reach ratio measures downstream propagation as the fraction of later nodes reachable forward from the lineage's birth in the DAG. The community pillar is completed by counting distinct authors per trajectory and by crediting emergence ownership to the authors of living fronts (Garner et al 2017). Influential documents and their bridging roles inside each trajectory are identified with within-module Z_i-scores and participation coefficients (Guimera and Amaral 2005).

Validating the partition. A partition should gather nodes about the same thing and keep different topics apart. We score this with the silhouette of Rousseeuw (1987) over a content distance, under one rule: the signal must be independent of what the detector routed on. The detector routes on document overlap, so coherence is scored on the papers' references (bibliographic coupling) or on their keywords. Each node is profiled from the papers new to its year, as the count vector of their features; every paper is new exactly once, so the profiles partition the corpus and cross-node similarity is genuine. Distance is one minus the Salton cosine between profiles, and the score is the mean silhouette weighted by each node's incremental document count, with singleton chains scoring zero. It is read at two resolutions, per trajectory and per final group; the latter is the verdict, since it tests whether each tributary reached a content-appropriate destination, exactly where braid and channel can disagree. The protocol therefore runs head-to-head across detectors, reporting which leads, by how much, and on which contested nodes.

Main paths. The optional extension module of Step 10 computes a within-group key-route main path (Liu and Lu 2012; Batagelj 2003): the highest-traffic citation route inside a final community, labelled by the documents and authors that carry it. It is an auxiliary, document-level view, not the trajectory detector: the route names the dominant line of descent inside one community, while the trunk decomposition tracks the population of lineages around it.

Implementation and reproducibility. All steps are implemented in the open `birddog` R package. Upstream community detection is stochastic, so the seed is reported with the results; the detector is deterministic given the DAG. The parameters that matter (the community and DAG node floors, the Jaccard floor τ_J, the number of successors kept per node, and the state thresholds) are reported to enable sensitivity analysis and replication.

5. Results

We illustrate the proposed methodology with biogas, a domain that sits at the intersection of multiple innovation-policy challenges: the energy transition, waste management, and circular bioeconomy strategies (Manikandan et al 2023; Martín-Espejo et al 2022). Biogas innovation features multiple competing and coexisting technological routes, making it a suitable test case for trajectory detection.

5.1. Data and corpus

We retrieved 77,347 documents (published through 2025) from OpenAlex (Priem et al 2022) using the query `title_and_abstract.search = ‘‘( biogas )’’`. Articles form the majority of the corpus (62,248 documents), followed by book chapters, dissertations, preprints, and reviews.

The direct-citation network is built on the corpus's peer-reviewed research documents, articles and reviews (63,155 documents after deduplication), and restricted to internal citations (edges where both citing and cited documents belong to the dataset): 40,822 documents, 65% of that universe and about 53% of the full corpus. Its giant connected component retains 40,300 documents (98.7% of the network), indicating a well-connected knowledge structure; the remaining documents form small, disconnected clusters loosely related to the focal domain.

We detected communities on the giant component with the Louvain algorithm (Blondel et al 2008) (seed 888, with a 30-document community floor), obtaining nineteen research communities, summarized in Table 4.

5.2. The static stock: nineteen communities

The nineteen groups partition the giant component into communities of varying size, age, and structural role (Table 4). A group is a stock, a single cumulative snapshot, so it carries no dynamic state; the table reads a short label, size, average publication year, the recent growth of the community's central spine, the dominant hub role on the Z_iP_i plane (Guimera and Amaral 2005), the inter-group influence role (net citation exporter or importer), and the leading country. The labels distil the content analysis of Step 6, each community's representative multi-word terms ranked by group-level tf–idf, read against its most-cited internal documents; Table C.7 profiles every community with its label, top terms, and a one-line description. Figure 3 shows the resulting map: the giant component laid out by citation structure, with every document coloured by its final community.

Table 4: The static stock: nineteen biogas communities on the giant component (40,300 documents; the 168 documents in communities below the 30-document floor are not shown). Labels are shared with Figure 3 and profiled in Table C.7. Avg. yr is the mean publication year, in years and months (2016+6m = 2016 plus six months). Growth is the recent annual rate of the community's central spine; "n/a" marks five communities whose distinct spine is only one year old at the final snapshot (a size-rank individuation), for which growth is not estimated. Hub is the dominant $Z_i P_i$ role (R5 provincial, R6 connector; Appendix Appendix G); Influence is the sign of the net inter-group citation balance (Source = net exporter, Sink = net importer); Top country is the community's modal country, counting every distinct author country on each document. c1g18 is a geotechnical homonym captured by the query, not a community of the biogas field (see text).

Group	Label	Docs	Avg. yr	Growth	Hub	Influence	Top country
c1g1	Foundational Backbone	7,841	2016+7m	+25.0%	R6	Sink	China
c1g2	Pretreatment	7,769	2016+6m	+20.5%	R6	Source	China
c1g3	Industrial Upgrading	4,438	2018+7m	+25.2%	R6	Sink	China
c1g4	Agriculture	3,013	2017+10m	+23.1%	R6	Sink	China
c1g5	Urban Waste	2,769	2016+8m	−12.1%	R6	Source	China
c1g6	Microbiology	2,417	2016+8m	+1.0%	R6	Source	China
c1g7	Industrial Effluents	2,123	2014+10m	+5.7%	R6	Sink	Indonesia
c1g8	Advanced Bio	1,978	2017+11m	−7.3%	R5	Sink	China
c1g9	Optimization	1,707	2016+9m	+14.3%	R6	Sink	Germany
c1g10	Combustion	1,510	2016+7m	+4.6%	R5	Sink	India
c1g11	Algae	1,432	2017+10m	+7.2%	R6	Sink	China
c1g12	Additives	1,320	2019+2m	+54.5%	R6	Sink	China
c1g13	Energy Crops	714	2015+7m	n/a	R6	Source	Germany
c1g14	Landfills	473	2014+11m	n/a	R6	Source	Brazil
c1g15	Biotechnology	247	2017+7m	n/a	R5	Sink	United States
c1g16	Local Residues	193	2016+5m	n/a	R6	Sink	Indonesia
c1g17	Polymers	73	2019+10m	+37.7%	R6	Sink	Italy
c1g18	Geotechnics	59	2019+1m	+40.2%	R5	Sink	Mexico
c1g19	Microgrids	56	2021+9m	n/a	R5	Sink	India

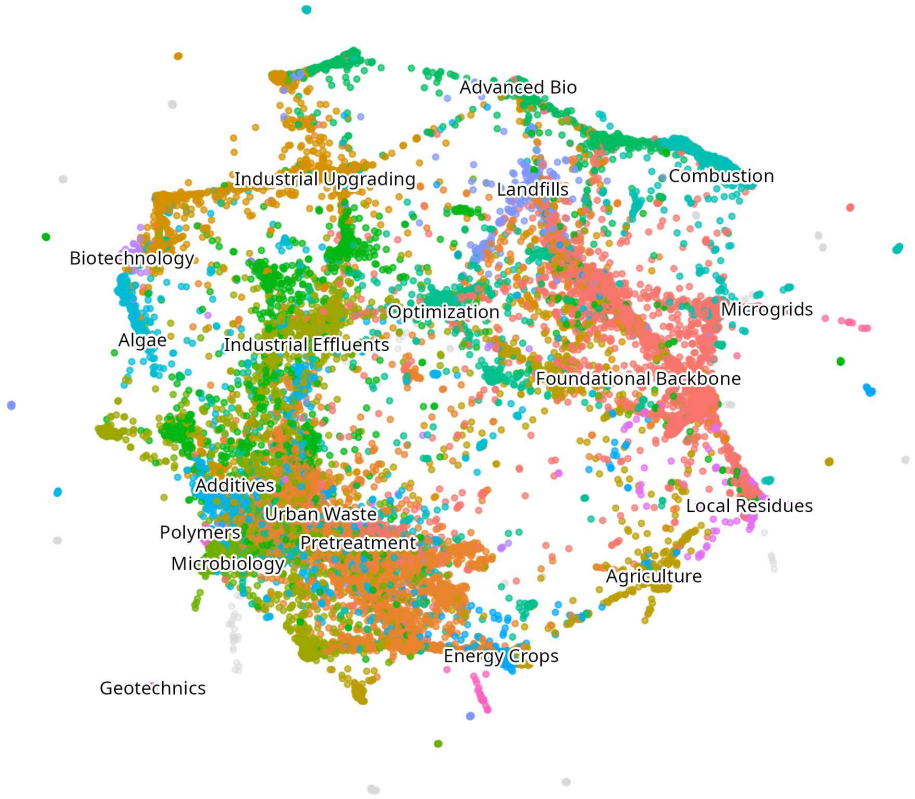


Figure 3: The static stock as a map: the giant component (40,300 documents) laid out by citation structure, with each document coloured by its final community and labels at the community centroids. The communities are the nineteen groups of Table 4.

Two large communities anchor the field: c1g1 (7,841 documents, the foundational backbone of biogas production and utilization, from household digesters to plant-scale energy systems) and c1g2 (7,769 documents), the strongest net knowledge exporter with a citation balance of +5,100, built on pretreatment of lignocellulosic biomass and biomethane-potential (BMP) standardization, the methods on which the rest of the field builds. The microbiology and metagenomics community c1g6 is a second strong source (+2,927), the field's engine room for the microbial pathways of methanogenesis. The steepest growth belongs to c1g12 (nanoparticle and trace-element enhancement of anaerobic digestion, including zero-valent iron nanoparticles, cobalt and nickel supplementation, and biochar): it is the strongest

net importer in the field (−3,489), consistent with an applied front drawing heavily on materials science. Two young niches are just crystallizing: c1g17 (anaerobic degradation of bioplastics, average year 2019+10m) and c1g19 (biogas-powered isolated microgrids and load-frequency control, 2021+9m, the youngest community). A third young community, c1g18 (2019+1m), belongs to the corpus but not to the field. In geotechnical engineering, "biogas" names the nitrogen bubbles that denitrifying bacteria produce to desaturate sands and mitigate liquefaction, and the keyword query captures this literature through its vocabulary alone. We keep c1g18 in the decomposition, which isolates it (Section 5.6), but leave it out of the substantive reading of the field.

Geography tracks these roles. China leads the large cores, with India a strong second in the pretreatment community c1g2; India leads the combustion community c1g10 and the microgrids niche c1g19; Indonesia leads the agro-industrial-effluent community c1g7 (palm-oil-mill effluent and sugarcane vinasse, average year 2014+10m); Brazil leads the landfill-gas recovery community c1g14; Italy leads the bioplastics niche c1g17; and Mexico leads the geotechnical homonym c1g18. The structure is core–periphery: large sources and near-balanced cores coexist with smaller, fast-growing importers where new combinations form. The next sections turn these stocks into flow, which is where emergence, convergence, divergence, and dormancy live.

5.3. The influence structure: who leads whom

Netting the cross-citation matrix C_{ij} underlying Equation 1 gives the field a directional structure (Figure 4). Only five communities are net knowledge exporters: the pretreatment and BMP community c1g2 (+5,100 citations), the microbiology engine room c1g6 (+2,927), the sludge community c1g5 (+1,565), and the small energy-crops and landfill communities c1g13 (+513) and c1g14 (+91). Every other community imports more recognition than it supplies, with the heaviest importers at the applied frontier: the additives front c1g12 (−3,489), the modeling community c1g9 (−2,189), and the reforming and fuel-cell community c1g8 (−1,683). The largest net channels all leave the upstream core: c1g2 feeds the additives front (1,633 net citations), the foundational core (1,334), and the modeling community (1,160), while c1g6 feeds upgrading (638) and additives (523). The one broker is c1g7, essentially at balance (−4 citations across more than eleven thousand intergroup flows); the smallest communities (c1g16, c1g18, c1g19) trade too few citations for their sign to carry weight. The reading is structural: protocols

and microbial fundamentals are the field's upstream sources, and the fast-growing fronts are downstream consumers that convert that knowledge into applications. The surprise normalization adds a finer signature: the young microgrids niche c1g19 cites the reforming and combustion communities far above the configuration null ($\rho = 13.5$ and 4.8), the profile of a downstream integrator assembling an application from established component knowledge.

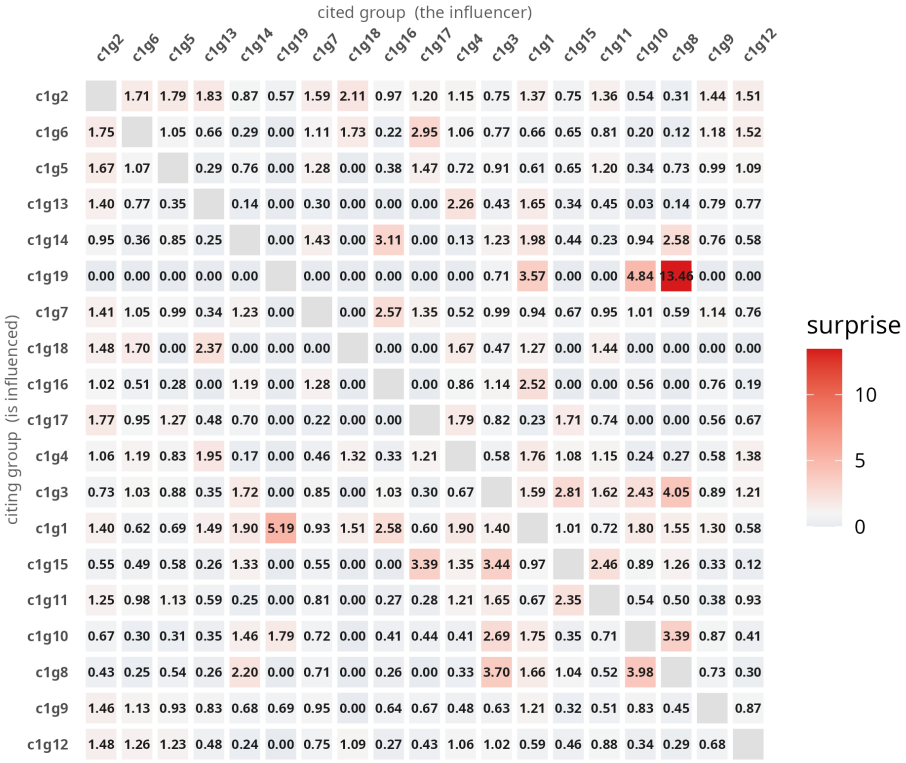


Figure 4: Inter-group influence: rows cite, columns are cited. Cell values are the surprise normalization ρ_{ij} of Equation 1; values above one mark channels over-represented relative to a directed configuration null. Groups are ordered by net citation balance, sources first.

5.4. From stock to flow: the trajectory decomposition

We turn the nineteen stocks into flow with the trunk decomposition of Section 4. Reverse cumulative clustering used the Louvain algorithm with fixed settings across years; the temporal DAG kept consecutive-year Jaccard edges above the default floor, retaining the two strongest successors per node.

Braid decomposed the DAG into node-disjoint lineages. The coherence validation confirmed that choice: braid attained the higher silhouette on both independent signals (references and keywords) at the final-group resolution, which judges whether each tributary reaches a content-appropriate destination.

The decomposition returns 98 trajectories: 19 central spines, one per final group, and 79 absorbed tributaries. Of the tributaries, 60 merge into a living lineage and 19 are extinct: their last node has no successor above the Jaccard floor, so no lineage inherits them. Figure 5 shows the result at full scale, the empirical counterpart of the conceptual Figure 2: every group–year node of the biogas DAG, with the three systems the rest of the paper follows picked out in colour, the endogenous upgrading line c1g3, the confluent urban-waste river c1g5, and the emergent additives front c1g12. A long single-colour run is a central spine; a colour change along a path marks a handoff; ringed nodes are the births along the highlighted systems; the grey mass holds the sixteen other communities and the nineteen extinct lineages. Who absorbs whom, and how much each river carries, is quantified in Section 5.6.

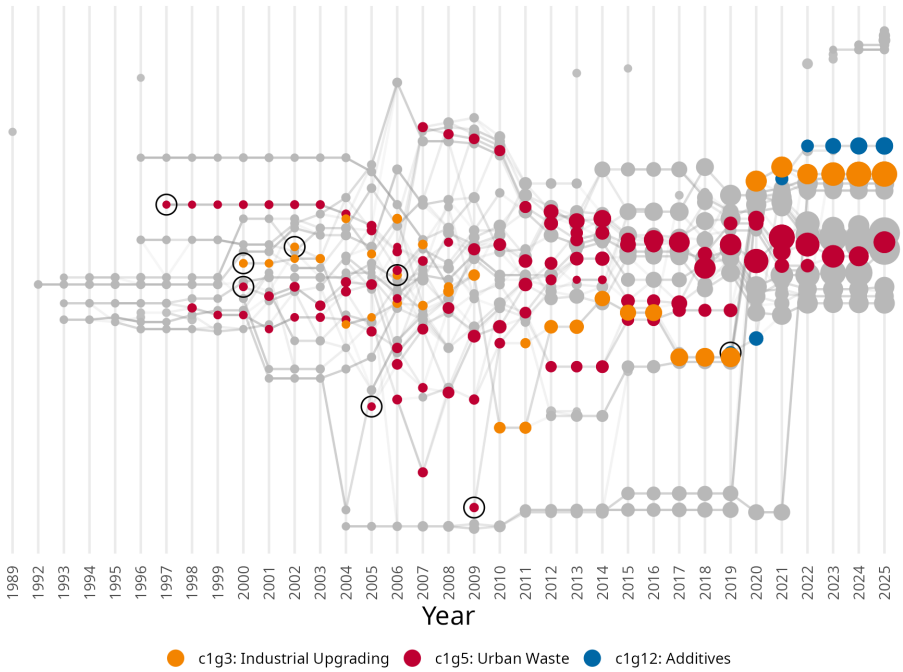


Figure 5: The trajectory decomposition at full scale: the biogas temporal DAG (522 group–year nodes, 682 migration edges, 1989–2025). Node size is the document count; edges join consecutive years where document migration clears the Jaccard floor. Three systems are coloured by the final community each braid chain drains into, resolved transitively through absorptions: c1g3 (Industrial Upgrading, endogenous), c1g5 (Urban Waste, confluent), and c1g12 (Additives, emergent); black rings mark the births of their chains, nodes with no admitted predecessor: the first year at least $n_{\min}$ documents cohere into a community without inheriting membership from any group of the previous year, dating the moment a lineage individuates in the citation record. All other lineages, the nineteen extinct ones included, are grey.

5.5. Trajectory dynamics: four states by type

States are read by trajectory type on two axes (Table 5), under the default data-driven cuts of `birddog`.[1] Central trajectories, which reach the present,

[1]The default cuts adapt to the field (Section 4): here, emergence requires at least 35.6% annual growth and a novelty of 0.65; dormancy, a growth of −0.8% or less; and convergence, a destination entropy below 0.73. With the fixed cuts that `birddog` also offers (15%, 0.30, −5%, and 0.5, respectively), the four largest cores (c1g1–c1g4), growing 20–25% a year, become emergent too, and the 60 tributaries with a living destination split into 12 convergences and 48 divergences instead of 30 and 30. Every other trajectory keeps

take a life-cycle phase from growth and novelty. Absorbed tributaries, which stopped, take a terminal fate: convergence or divergence, read from the entropy of their destination cohort, or dormancy, when the lineage ends with no successor.

Table 5: Trajectory states by type, under the default data-driven cuts. Central spines take a life-cycle phase (emergence, maturity, dormancy) on growth and novelty; absorbed tributaries take a terminal fate: convergence or divergence on the entropy of their terminal cohort's destination, dormancy by extinction.

Type	Emergence	Maturity	Dormancy	Convergence	Divergence	Total
Central	3	14	2	–	–	19
Absorbed	–	–	19	30	30	79

Fourteen of the nineteen centrals are mature, established communities whose spine still grows but has left its fast phase. Three are emergent: c1g12 (nanoparticle and trace-element enhancement, +55% recent growth at age seven), c1g18 (+40%), and c1g17 (bioplastics degradation, +38%). The community c1g18 is emergent by the numbers but lies outside the field: it is the geotechnical homonym of Section 5.2. Two are dormant, their spines shrinking rather than growing: the sewage-sludge community c1g5 (−12%) and the long-established biogas-reforming and fuel-cell community c1g8 (−7%, age twenty-two). The emergence index, which ranks living fronts on winsorized growth, novelty, and recruitment, places the two large cores c1g1 and c1g2 at the top by absolute magnitude and c1g12 as the strongest genuinely young front (the volume-free intensive variant puts the homonym c1g18 first by a hair). Five centrals (c1g13–c1g16 and c1g19) carry a maturity label by default rather than by an S-curve reading: their distinct spine is only one year old at the final snapshot, a size-rank individuation, so their growth is not estimated.

Among the tributaries, the 60 that reach a living destination split into 30 convergences and 30 divergences, an even split by construction, since the default cut is their median destination entropy. A converging tributary hands a concentrated cohort to a single destination, a merge seen from the absorbed

its state, including the three emergent centrals (the novelty of c1g12 sits exactly at the default cut); the two dormant centrals, which shrink faster than either dormancy cut; and the 19 extinct tributaries, whose fate depends on the Jaccard floor rather than on these cuts.

side; a diverging one scatters its cohort across several groups. In absolute terms, most cohorts scatter: the median normalized destination entropy is 0.73. The 19 extinct lineages carry the dormancy fate by extinction; none qualifies by drop-out, since no terminal cohort loses more than 10.4% of its documents. The author-level reading, crediting emergence ownership of the living fronts, is tabulated in Appendix F.

5.6. Confluence and self-sufficiency

Absorption is transitive, so tributaries do not enter a final group as a flat set but through a confluence tree, with merges as deep as five levels. The largest rivers gather many tributaries: the sewage-sludge and municipal-wastewater community c1g5 draws 15, and the foundational core c1g1 and the agro-industrial-effluent community c1g7 draw 12 each. The tributary-by-community map behind these counts is in Appendix E (Figures E.12 and E.13).

Self-sufficiency measures how much of a central's final community the lineage grew itself rather than received from the terminal cohorts of absorbed tributaries, its own or others'. The endogenous pole is the microgrids community c1g19 (1.0, no inflow at all), with the geotechnical homonym c1g18 (0.97), the reforming and fuel-cell community c1g8 (0.94), and the microalgae community c1g11 (0.93) close behind: communities that formed largely on their own. For c1g18 the near-endogeneity is what a homonym looks like in the flow: a literature that shares the field's word but not its citations grows on its own. The confluent pole is the energy-crops community c1g13 (0.09): no tributary is assigned to it, yet 648 of its 714 final documents arrive from the terminal cohorts of tributaries absorbed elsewhere, so its spine individuates in 2025 almost entirely from carried-in material. It is followed by the large source community c1g2 (0.25) and the landfill community c1g14 (0.26); among the large cores, the sludge community c1g5 sits at 0.44. Self-sufficiency is not size: among the three largest communities, the pretreatment core c1g2 is strongly confluent while the upgrading core c1g3 is near-endogenous (0.92), a contrast the static size ranking cannot reveal. Figure 6 shows the full ranking.

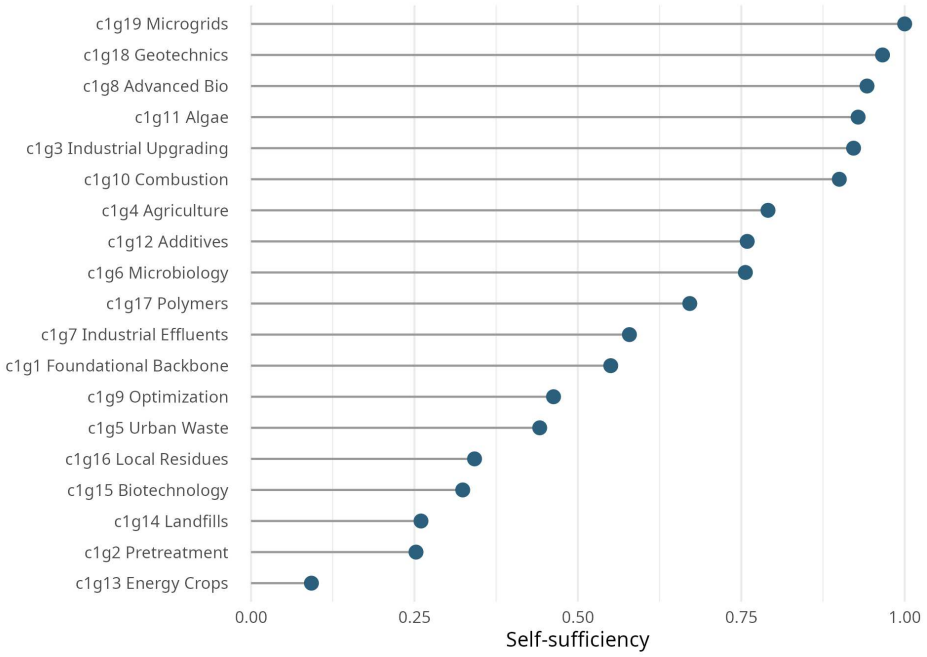


Figure 6: Self-sufficiency of the nineteen central trajectories: the share of each central's spine grown endogenously rather than carried in by tributaries. The endogenous pole (c1g19, c1g18, c1g8, c1g11, c1g3) contrasts with the confluent pole (c1g13, c1g2, c1g14).

5.7. Reading the flow: four trajectories

Four lineages make the states and the structural fates concrete.

A clean absorption (convergence). The tributary tr5, a 2,170-document lineage running 2009–2022 on wastewater and membrane-bioreactor methane recovery, merges into the sewage-sludge central c1g5 in 2023. Its terminal cohort is highly concentrated (Figure 7): 88% of its last-node documents sit in c1g5 at the final year, with a normalized destination entropy of 0.27 and almost no drop-out (about 1%). The lineage lost its name, not its papers, the concrete meaning of "death" as absorption.

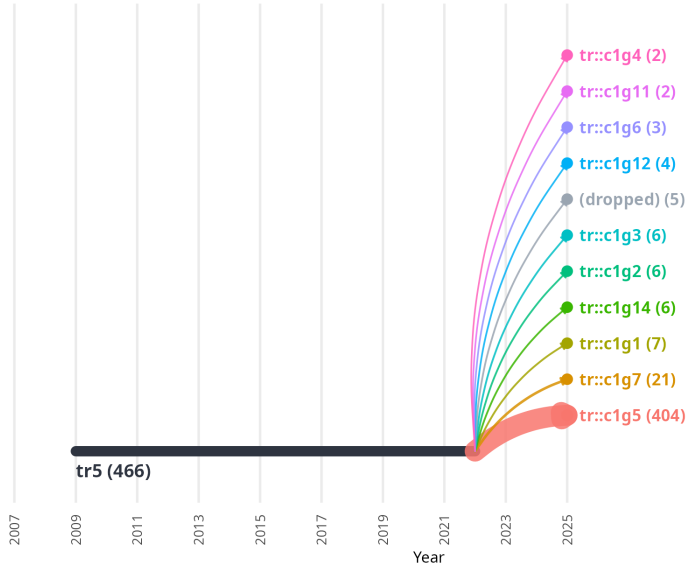


Figure 7: A clean absorption: the tributary tr5 ends in 2022 and its terminal cohort of 466 documents is tracked by identity to the final year. 404 documents (88% of the surviving cohort) sit in the sewage-sludge central c1g5, five drop out, and the remainder scatters thinly across ten other communities.

A scattering (divergence) inside the same river. The tributary tr7, a 1,218-document lineage running 2000–2013 on foundational cattle-manure experiments and the technical limits of mesophilic reactors, is absorbed in 2014 not by the central but by tr4, itself absorbed by the trunk tributary tr1, which merges into the sewage-sludge central c1g5 in 2022. tr7 is a divergence: its terminal cohort of 301 documents scatters with a destination entropy of 0.77, the largest single destination holding 25%. Structure and documents split here: the lineage's name resolves three levels down the c1g5 confluence tree, while its cohort spreads across the field, landing more heavily in the pretreatment core c1g2 (25%) and the agro-industrial-effluent community c1g7 (20%) than in c1g5 itself (5%). In all, c1g5 gathers fifteen tributaries across five depths, the deepest confluence tree of the decomposition; Figure 8 draws its main feeders, with tr5, the clean absorption above, one of the two direct ones.

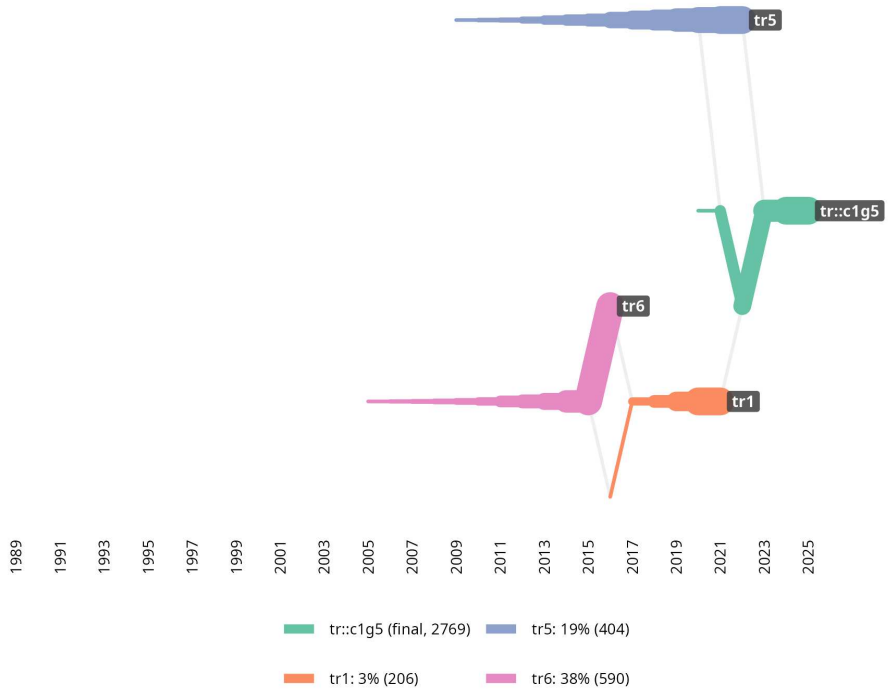


Figure 8: The confluence tree of the sewage-sludge central tr::c1g5, drawn for the feeders that deliver at least 100 documents to the final community; the full tree gathers fifteen tributaries across five depths, the deepest confluence of the decomposition. Band width is community size per cumulative year; the legend gives each tributary's delivery to the final community, in documents and as a share of its cumulative size. The heaviest deliveries come from tr6 (590 documents) and tr5 (404); the trunk tributary tr1 delivers only 206 of its 4,736-document terminal cohort, most of which re-partitions into the pretreatment core c1g2.

An emergent front. The central c1g12 is the field's clearest living emergence. Its main citation route runs from early high-rate fermentation work through cobalt and nickel supplementation studies to zero-valent-iron and cobalt/nickel nanoparticles and biochar around 2015–2016, a lineage that turns trace-element chemistry into a distinct, fast-growing branch of anaerobic-digestion research. The structural signature matches: it is the strongest net knowledge importer in the field (citation balance −3,489), carries high novelty, and holds the top emergence index among genuinely young centrals.

An extinction (structural death). The tributary tr19, a 2015–2017 lineage of 476 documents on the methanotrophic conversion of methane into

products, ends in 2017: no successor clears the Jaccard floor, so no chain inherits its name, which is the strict meaning of extinction and the reason it carries the dormancy fate in Table 5. Its papers, however, do not die. Of its 212-document terminal cohort none drops out (dormant share 0); the cohort scatters, with 22% landing in the biotechnology community c1g15, 21% in the foundational core c1g1, and 13% in c1g2. Extinction here is structural rather than documentary: the lineage loses its edge continuity while its documents go on to seed other communities, including the methanotroph-biotechnology community where its largest share settles. The dormancy fate marks the death of the lineage; the dormant share measures the death of papers, and the two can part company. The overlap measure explains why they part. When a small cohort moves into a much larger community, the Jaccard overlap falls toward the ratio of their sizes, so absorption by a community more than twenty times larger leaves no edge above $\tau_J = 0.05$. For the seventeen extinct lineages whose documents reappear the following year, the community receiving most of them is a median 25 times the lineage's size, against 2.5 times for the absorbed tributaries, and ten of the seventeen move 63–96% of their cohort into that single community. Extinction in this decomposition therefore mostly marks absorption into a community too large for the Jaccard floor to register, not the end of a line of research.

5.8. From structure to dynamics: synthesis

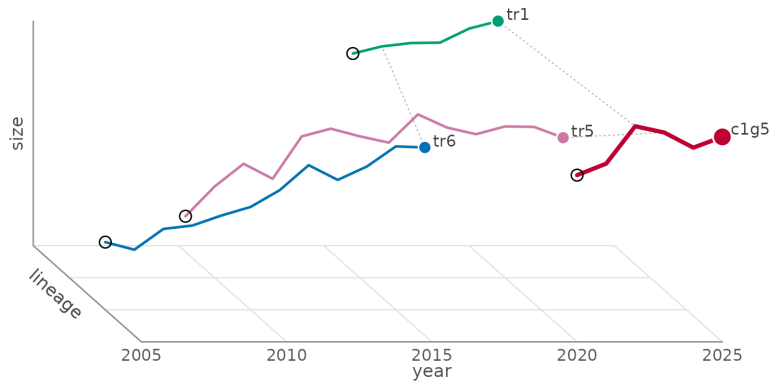
The static stock (Section 5.2) gave a core–periphery pattern of nineteen communities; the flow decomposition (Sections 5.4–5.7) showed how these communities formed and where they now sit. Together the two lenses trace a three-phase periodization. The first phase, through the 1990s and early 2000s, was agricultural and rural experimentation. It is carried less by any single final community than by the early tributaries now absorbed into the foundational anaerobic-digestion core c1g1, on household digesters, agricultural residues, and rural dissemination in India and the developing world, by the foundational cattle-manure lineage tr7 (Section 5.7), and by the agro-industrial-effluent lineages of c1g7 (palm-oil-mill effluent and sugarcane vinasse, average year 2014+10m). The second phase, accelerating through the 2010s, was industrialization and standardization: pretreatment and biomethane-potential protocols (c1g2), biogas cleaning and upgrading to biomethane (c1g3), and digestate valorization for agriculture (c1g4) maturing into large applied cores. The third phase, from the 2020s, is an inflection toward methods intensification and systems integration, carried by nanopar-

ticle and trace-element enhancement of digestion (c1g12) and biogas-powered microgrids (c1g19), with the anaerobic degradation of bioplastics (c1g17) opening a new substrate at the margin. The periodization is not imposed; it emerges from the flow, which trajectories are emergent, which are dormant, and where the tributaries drain.

The flow evidence shows a field consolidating and exploring at once rather than locked in. Consolidation shows in the two shrinking spines (c1g5 and c1g8) and in the 60 tributaries absorbed into living lineages; exploration shows in the young fronts that grow fastest (c1g12 and c1g17) and in absorbed cohorts that mostly scatter across several communities rather than concentrating in one, the signature of active search in adjacent spaces. Drawing on Arthur (1990) and Dosi (1982), this coexistence of consolidation and variety matches the emergence literature's definition, radical novelty growing fast while remaining coherent (Rotolo et al 2015; Carley et al 2017), and it points to continuing scope for reconfiguration rather than a fully locked-in regime.

Self-sufficiency and confluence sharpen the reading. The field's growth is not endogenous everywhere: young communities such as c1g19 formed on their own, while the pretreatment core c1g2 is strongly confluent (self-sufficiency 0.25) and the energy-crops community c1g13 is built almost entirely from material carried in by tributaries absorbed elsewhere. The same decomposition that dates each merge also says how much of each community was built by absorption rather than by internal growth, a distinction the static map cannot make. Figure 9 closes the reading with the two regimes, each community's chains drawn as competing growth curves: the confluent c1g5, whose young spine caps a tree of fifteen tributaries, against the endogenous c1g3, one long self-propelled curve trailed by four small early merges. The same view for all nineteen communities is in Appendix E, and c1g3 is read end to end, with the key-route extension module of Step 10, in Appendix H.

(a) c1g5 Urban Waste, confluent: self-sufficiency 0.44



(b) c1g3 Industrial Upgrading, endogenous: 0.92

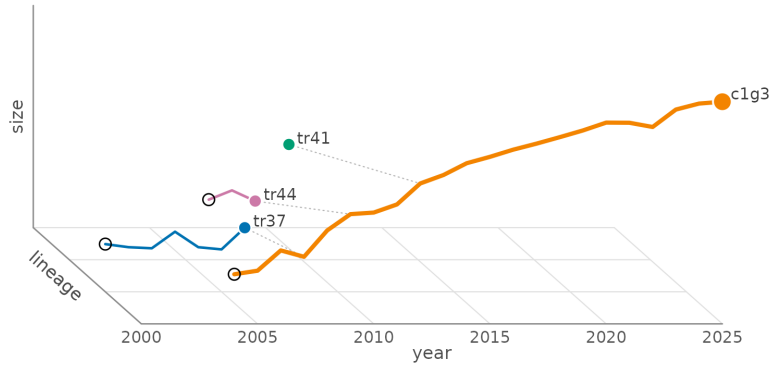


Figure 9: Two growth regimes as competing lineage curves (year × lineage × size, logarithmic size scale based at the $n_{\min}$ node floor; the visual grammar of Figure 2). (a) The confluent c1g5: the spine (front) individuates only in 2020 and inherits the mass of the trunk tributary tr1; lanes are the feeders delivering more than 1.5% of their cumulative size to the final community (tr1, tr5, tr6; the twelve smaller tributaries are omitted), and dotted connectors mark handoffs. (b) The endogenous c1g3: one spine growing from 2004 under its own recruitment, with its early tributaries (tr70, which delivers nothing to the final, is omitted). Rings mark births; balls mark chain ends; self-sufficiency 0.44 against 0.92.

Methodologically, the payoff over static community detection and route-based main-path analysis is that the four states are read off one decompo-

sition in which every group–year node sits in exactly one lineage and every lineage carries a state and a destination. Consolidation (dormant centrals, converging tributaries) and exploration (emergent centrals, diverging tributaries) are then two readings of the same object, the flow, rather than separate diagnoses.

5.9. Validating the decomposition

The trunk decomposition is validated on content signals the detector never saw. Because braid and channel route on document overlap, coherence is scored on the papers' references (bibliographic coupling) and on their keywords, with incremental node profiles, at two resolutions (Section 4); all detection parameters are fixed and reported (Appendix I). Table 6 reports the document-weighted mean silhouettes.

Table 6: Document-weighted mean silhouette of the trajectory partition on two independent content signals, at two resolutions, with incremental node profiles. Higher is better; the final-group resolution is the verdict, since it scores whether each tributary reaches a content-appropriate destination.

	References (coupling)		Keywords	
Detector	trajectory	final group	trajectory	final group
braid	0.076	**0.053**	**0.121**	**−0.091**
channel	**0.088**	0.037	0.078	−0.104

Braid leads in three of the four cells, including both signals at the decisive final-group resolution (+0.016 on references, +0.013 on keywords); channel's single win is on references at the trajectory resolution (+0.012). The detectors disagree on 314 contested nodes, and the selection rule is recorded with the result: when the two signals disagree at the chosen resolution within a minimum margin of 0.02, the default detector is kept, which here retains braid. Two honest qualifications attach to the absolute numbers. The silhouettes are modest throughout, and on keywords at the final-group resolution they are negative for both detectors: nineteen communities of one applied field overlap heavily in vocabulary, so content signals separate trajectories only weakly at that granularity. The comparison is therefore relative, and its force is that two independent signals crown the same detector at the resolution that matters.

6. Discussion

What the decomposition adds to the trajectory program. The key route of c1g3 (Appendix H) shows what route extraction does well: it names the documents and authors that carry an idea through four decades. What it cannot say is what the field did with that idea, and that is what the decomposition supplies. Around the same community, the flow layer dates four tributary absorptions, measures a self-sufficiency of 0.92, tracks renewal and consolidation year by year, and places the community among nineteen lineages whose states are read from one object. The two views are complements, not rivals: Main Path Analysis answers where the dominant line of descent runs, and the trunk decomposition answers how the population around it generates variety, selects, retains, and goes dormant. In this sense, the decomposition adds a population view to the route view of the citation-network program of trajectory mapping (Hummon and Dereian 1989; Verspagen 2007).

An open selection environment. Read through the evolutionary vocabulary, the biogas field is a selection environment that is still open. Variety is still generated: the fastest-growing spines of the field are young, led by the additives front c1g12 and the bioplastics niche c1g17. Selection visibly resolves: 60 of the 79 tributaries fold into living lineages, through confluence trees up to five levels deep. What selection retains, however, is rarely a single line: most absorbed cohorts scatter across several destinations (median normalized entropy 0.73), so the knowledge of an ended lineage is recombined across communities rather than concentrated in one. Exhaustion is limited: two centrals are dormant, and the nineteen extinct lineages lose their names, not their papers (eighteen lose no documents at all, the nineteenth about a tenth). The hub documents show the same at the document level: extinct lineages hold few hubs, while the widest-reaching hubs, mostly papers published by 2010, entered the network through tributaries that later diverged (Appendix Appendix G). Consolidation and exploration coexist in the same decomposition, which is why the field reads as paradigmatically maturing without being locked in. A route-based reading could not have made this diagnosis, because none of its elements (absorption dates, destination entropies, self-sufficiency) exists at the route level.

Policy implications. For innovation policy and technology governance, the framework offers actionable diagnostics. Emergence signals (rapid growth, high novelty, falling citation cycle time) can flag research fronts that merit

early-stage support or monitoring, such as additive-enhanced digestion (c1g12) and bioplastics degradation (c1g17) in the biogas case. The case of c1g18 adds a caution: the second fastest-growing young community in the corpus is a homonym of the field, so an emergence signal tells analysts where to look, not what to fund. Dormancy signals (stalling or shrinking spines, rising citation cycle time) can alert policymakers to fields where lock-in risk is increasing and where deliberate diversification may be warranted. Divergence patterns identify moments when a community's search space is branching, creating windows for policy intervention that can shape which directions are pursued. Convergence signals can indicate the formation of de facto standards, informing decisions about infrastructure investment and regulatory alignment.

7. Conclusions

This paper proposed an unsupervised computational model for detecting and classifying technological and scientific trajectories as system-level objects that evolve over time. The approach integrates a deterministic trunk decomposition of cumulative clusterings of temporal citation networks with indicator-based diagnosis, enabling the classification of trajectories into four interpretable states: emergence, convergence, divergence, and dormancy. We demonstrated the method on a corpus of 77,347 biogas-related documents retrieved from OpenAlex.

Contributions. The paper offers three contributions to evolutionary economics. First, it extends the citation-network approach to technological trajectories (Hummon and Dereian 1989; Verspagen 2007) from the dominant route to a system-level decomposition, connecting the structural vocabulary of community evolution (births, splits, merges, declines) with the indicator-based diagnostics of the emergence literature (novelty, growth, persistence). Second, the method operationalizes the core mechanisms of evolutionary change (variety generation, selection, retention, and path dependence) as measurable trajectory states, read from growth, novelty, and destination entropy and supported by citation cycle time and keyword entropy, indicators designed for comparison across sectors and temporal scales. Third, the open implementation in the `birddog` R package enables replication and extension, supporting the cumulative development of trajectory-based evidence in evolutionary studies of technological change.

Findings. The biogas application revealed a field with a strongly core–periphery knowledge structure: nineteen research communities span anaerobic digestion fundamentals, biomethane upgrading, microbiology, and emerging niches such as nanoparticle-enhanced digestion and biogas-powered microgrids. The trajectory analysis identified a three-phase evolution: from 1990s rural experimentation, through 2010s industrialization and standardization, to a 2020s inflection toward methods intensification and systems integration. Large established cores (e.g., c1g1, foundational anaerobic digestion) coexist with fast-growing niches (e.g., c1g12, nanoparticle-enhanced digestion; c1g17, bioplastics degradation) and with communities so young they individuated only in 2025 (e.g., c1g19, biogas-powered microgrids), suggesting that the field retains scope for technological reconfiguration despite paradigmatic maturation in its largest communities.

Limitations and future work. Several limitations should guide interpretation. The application uses scientific articles only. The pipeline accepts patents as well, since it needs only documents with a publication year, text, and citation relations, and patent data would bring the analysis closer to the technological side of the trajectories it detects. A corpus that joins both, linking papers and patents through the scientific references that patents cite, would let the decomposition follow knowledge as it moves from science into technology. The corpus is also defined by a keyword query, which captures homonyms along with the field: c1g18 is a geotechnical literature in which biogas means biogenic nitrogen bubbles. The decomposition isolated it as a near-endogenous community, but corpus definition still calls for domain reading. Scale is also untested: the biogas corpus holds 77,347 documents, and none of the earlier applications of the pipeline (Souza 2016; Souza et al 2021, 2022; Maria et al 2023; Biazatti et al 2024; Miranda et al 2025; Felizardo et al 2025) reaches a million. A broad area of knowledge with millions of documents would test the computational cost of re-clustering one cumulative network per year, whether the decomposition stays readable when the DAG holds thousands of nodes, and how far the results generalize beyond a single applied field. Finally, parameter choices (the Jaccard floor, the community and node floors, the number of successors kept per node, and the community-detection algorithm) influence the granularity of detected trajectories. The Jaccard floor matters in a specific way: because the overlap divides by the union of two document sets, it understates the absorption of small lineages by much larger communities, and most of the nineteen extinct lineages are

such absorptions (Section 5.7); an asymmetric overlap, the share of a node's documents that reappear in its successor, would register them. The state thresholds raise a related caution. The default cuts are computed from each field's own distributions, so states are relative to the field: every application splits its tributaries with a living destination evenly between convergence and divergence, and emergence is measured against the field's own distribution. Comparisons across fields should therefore rest on the underlying measures (growth, novelty, destination entropy) or on fixed cuts, whose effect on the biogas states is reported in Section 5.5. A systematic sensitivity sweep across these parameters and thresholds would strengthen robustness and is left for future work.

Four lines of future work follow from the results. The trunk decomposition is one detector among possible ones: braid and channel were compared here (Section 5.9), and other rules for extracting lineages from the temporal DAG, together with other community-detection algorithms, can be benchmarked with the same content-coherence test. The comparison with Main Path Analysis remained qualitative and limited to one community (Appendix H); a systematic comparison across all communities, measuring how far main paths and trunk lineages share documents and periods, would turn the complementarity argued in Section 6 into a measured result. The 98 trajectories were read here only as far as needed to illustrate their states; each carries a technical content, in substrates, processes, and upgrading routes, whose detail and implications for biogas technology and policy deserve a reading of their own with domain experts. Last, the `birddog` package runs in R, which restricts its users to those who program; a graphical interface would open the method to analysts and policy practitioners who do not.

Statements and Declarations

Funding. No funding was received for conducting this study.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Author contributions. Conceptualization: Roney Fraga Souza and José Maria Ferreira Jardim da Silveira; Methodology and Software: Roney Fraga Souza; Formal analysis and Investigation: Roney Fraga Souza, Luis Felipe de Souza Rodrigues, and Júlio Cesar de Carvalho Miranda; Writing – original draft: Roney Fraga Souza; Writing – review and editing: all authors; Supervision:

José Maria Ferreira Jardim da Silveira. All authors read and approved the final manuscript.

Data availability. The bibliometric data analyzed in this study were retrieved from OpenAlex (Priem et al 2022). The `birddog` R package implementing the methodology (version 2.0.0) is available at `https://roneyfraga.com/birddog`. The hub documents, with their citations by final group and by trajectory, are provided as Online Resource 1. Analysis code and processed data are available from the corresponding author upon reasonable request.

Generative AI use. The authors used a large language model for language editing and LaTeX formatting. The authors reviewed and edited all content and take full responsibility for the manuscript.

References

Almudi I, Fatas-Villafranca F (2021) Coevolution in Economic Systems. Cambridge University Press, Cambridge, https://doi.org/10.1017/9781108767798

Andersen B (1999) The hunt for s -shaped growth paths in technological innovation: a patent study ∗. Journal of Evolutionary Economics 9(4):487–526. https://doi.org/10.1007/s001910050093

Angelidaki I, Alves M, Bolzonella D, et al (2009) Defining the biomethane potential (bmp) of solid organic wastes and energy crops: a proposed protocol for batch assays. Water Science and Technology 59(5):927–934. https://doi.org/10.2166/wst.2009.040

Arthur WB (1990) Positive feedbacks in the economy. Scientific American 262(2):92–99. https://doi.org/10.1038/scientificamerican0290-92

Batagelj V (2003) Efficient algorithms for citation network analysis. arXiv preprint cs/0309023 https://doi.org/10.48550/ARXIV.CS/0309023

Batagelj V, Cerinšek M (2013) On bibliographic networks. Scientometrics 96(3):845–864. https://doi.org/10.1007/s11192-012-0940-1

Becker MC (ed) (2008) Handbook of Organizational Routines. Edward Elgar Publishing, Cheltenham, UK

Biazatti MJ, Justi ACA, Souza RF, et al (2024) Soybean biorefinery and technological forecasts based on a bibliometric analysis and network mapping. Environmental Development 52:101074. https://doi.org/10.1016/j.envdev.2024.101074

Blondel VD, Guillaume JL, Lambiotte R, et al (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Boyack KW, Klavans R (2010) Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology 61(12):2389–2404. https://doi.org/10.1002/asi.21419

Boyack KW, Klavans R (2020) A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. Quantitative Science Studies 1(4):1570–1585. https://doi.org/10.1162/qss_a_00085

Bródka P, Saganowski S, Kazienko P (2013) GED: the method for group evolution discovery in social networks. Social Network Analysis and Mining 3(1):1–14. https://doi.org/10.1007/s13278-012-0058-8

Carley SF, Newman NC, Porter AL, et al (2017) A measure of staying power: Is the persistence of emergent concepts more significantly influenced by technical domain or scale? Scientometrics 111(3):2077–2087. https://doi.org/10.1007/s11192-017-2342-x

Carley SF, Newman NC, Porter AL, et al (2018) An indicator of technical emergence. Scientometrics 115(1):35–49. https://doi.org/10.1007/s11192-018-2654-5

Chen SH, Huang MH, Chen DZ (2012) Identifying and visualizing technology evolution: A case study of smart grid technology. Technological Forecasting and Social Change 79(6):1099–1110. https://doi.org/10.1016/j.techfore.2011.12.011

Chen Y, Yang G, Sweeney S, et al (2010) Household biogas use in rural china: A study of opportunities and constraints. Renewable and Sustain-

able Energy Reviews 14(1):545–549. https://doi.org/10.1016/j.rser.2009.07.019

Corrocher N, Malerba F, Morrison A (2021) Technological regimes, patent growth, and catching-up in green technologies. Industrial and Corporate Change 30(4):1084–1107. https://doi.org/10.1093/icc/dtab025

Dakiche N, Benbouzid-Si Tayeb F, Slimani Y, et al (2019) Tracking community evolution in social networks: A survey. Information Processing & Management 56(3):1084–1102. https://doi.org/10.1016/j.ipm.2018.03.005

David PA (1985) Clio and the economics of qwerty. The American economic review 75(2):332–337

Ding M, Yu W, Zeng T, et al (2024) Ptns: patent citation trajectory prediction based on temporal network snapshots. Scientific Reports 14(1). https://doi.org/10.1038/s41598-024-75913-0

Dopfer K (2001) Evolutionary Economics:Framework for Analysis, Springer Netherlands, p 1–44. https://doi.org/10.1007/978-94-010-0648-4_1

Dosi G (1982) Technological paradigms and technological trajectories. Research Policy 11(3):147–162. https://doi.org/10.1016/0048-7333(82)90016-6

Dosi G (1988) Sources, procedures, and microeconomic effects of innovation. Journal of economic literature pp 1120–1171

Ehrnberg E (1995) On the definition and measurement of technological discontinuities. Technovation 15(7):437–452. https://doi.org/10.1016/0166-4972(95)96593-i

Epicoco M (2013) Knowledge patterns and sources of leadership: Mapping the semiconductor miniaturization trajectory. Research Policy 42(1):180–195. https://doi.org/10.1016/j.respol.2012.06.009

Fagerberg J, Verspagen B (2009) Innovation studies–the emerging structure of a new scientific field. Research Policy 38(2):218–233. https://doi.org/10.1016/j.respol.2008.12.006

Felizardo SB, Justi ACA, Souza RF, et al (2025) Transforming wastes into resources: Innovations in cotton biorefineries for a sustainable future. Waste and Biomass Valorization https://doi.org/10.1007/s12649-025-03136-z, URL https://link.springer.com/article/10.1007/s12649-025-03136-z

Ferrari VE, da Silveira JMFJ, Dal-Poz MES (2019) Patent network analysis in agriculture: a case study of the development and protection of biotechnologies. Economics of Innovation and New Technology 30(2):111–133. https://doi.org/10.1080/10438599.2019.1684645

Fontana R, Nuvolari A, Verspagen B (2009) Mapping technological trajectories as patent citation networks. an application to data communication standards. Economics of Innovation and New Technology 18(4):311–336. https://doi.org/10.1080/10438590801969073

Fortunato S (2010) Community detection in graphs. Physics Reports 486(3–5):75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Françoso MS, de Lima Avanci V, Mazoni AF (2024) Green technologies in the knowledge space: Insertion and the moderating role of industry knowledge bases. Journal of Evolutionary Economics 34(3):675–697. https://doi.org/10.1007/s00191-024-00871-9

Garfield E (2004) Historiographic mapping of knowledge domains literature. Journal of Information Science 30(2):119–145. https://doi.org/10.1177/0165551504042802

Garner J, Carley S, Porter AL, et al (2017) Technological emergence indicators using emergence scoring. In: 2017 Portland international conference on management of engineering and technology (PICMET), IEEE, pp 1–12, https://doi.org/10.23919/PICMET.2017.8125288

Greene D, Doyle D, Cunningham P (2010) Tracking the evolution of communities in dynamic social networks. In: 2010 International Conference on Advances in Social Networks Analysis and Mining. IEEE, p 176–183, https://doi.org/10.1109/asonam.2010.17

Guan J, Shi Y (2012) Transnational citation, technological diversity and small world in global nanotechnology patenting. Scientometrics 93(3):609–633. https://doi.org/10.1007/s11192-012-0706-9

Guimera R, Amaral LA (2005) Functional cartography of complex metabolic networks. Nature 433:895–900. https://doi.org/10.1038/nature03288, URL www.nature.com/nature.

Hansen KH, Angelidaki I, Ahring BK (1998) Anaerobic digestion of swine manure: inhibition by ammonia. Water Research 32(1):5–12. https://doi.org/10.1016/S0043-1354(97)00201-7

Hummon NP, Dereian P (1989) Connectivity in a citation network: The development of dna theory. Social Networks 11(1):39–63. https://doi.org/10.1016/0378-8733(89)90017-8

Jiang X, Liu J (2023) Extracting the evolutionary backbone of scientific domains: The semantic main path network analysis approach based on citation context analysis. Journal of the Association for Information Science and Technology 74(5):546–569. https://doi.org/10.1002/asi.24748

Kayal A, Waters R (1999) An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. IEEE Transactions on Engineering Management 46(2):127–131. https://doi.org/10.1109/17.759138

Kessler MM (1963) Bibliographic coupling between scientific papers. American Documentation 14(1):10–25. https://doi.org/10.1002/asi.5090140103

Kim J, Shin J (2018) Mapping extended technological trajectories: integration of main path, derivative paths, and technology junctures. Scientometrics 116(3):1439–1459. https://doi.org/10.1007/s11192-018-2834-3

Kim SH, Lee N, King PE (2020) Dimensions of religion and spirituality: A longitudinal topic modeling approach. Journal for the Scientific Study of Religion 59(1):62–83. https://doi.org/10.1111/jssr.12639

Kitsak M, Ganin AA, Eisenberg DA, et al (2018) Stability of a giant connected component in a complex network. Physical Review E 97(1). https://doi.org/10.1103/physreve.97.012309

Kuhn HW (1955) The hungarian method for the assignment problem. Naval Research Logistics Quarterly 2(1–2):83–97. https://doi.org/10.1002/nav.3800020109

Lee NK, Han Y, Xong W, et al (2020) Two layer-based trajectory analysis of the research trend in automotive fuel industry. Scientometrics 124(3):1701–1719. https://doi.org/10.1007/s11192-020-03506-5

Li M, Xu X (2022) Tracing technological evolution and trajectory of biomass power generation: a patent-based analysis. Environmental Science and Pollution Research 30(12):32814–32826. https://doi.org/10.1007/s11356-022-24339-0

Li M, Yan X, Yan L, et al (2025) Flexible and probabilistic topology tracking with partial optimal transport. IEEE Transactions on Visualization and Computer Graphics 31(10):7951–7969. https://doi.org/10.1109/tvcg.2025.3561300

Linares IMP, De Paulo AF, Porto GS (2019) Patent-based network analysis to understand technological innovation pathways and trends. Technology in Society 59:101134. https://doi.org/10.1016/j.techsoc.2019.04.010

Liu H, Chen Z, Tang J, et al (2020) Mapping the technology evolution path: a novel model for dynamic topic detection and tracking. Scientometrics 125(3):2043–2090. https://doi.org/10.1007/s11192-020-03700-5

Liu JS, Lu LY (2012) An integrated approach for main path analysis: Development of the hirsch index as an example. Journal of the American Society for Information Science and Technology 63(3):528–542. https://doi.org/10.1002/asi.21692

Malerba F, Orsenigo L (1996) Schumpeterian patterns of innovation are technology-specific. Research Policy 25(3):451–478. https://doi.org/10.1016/0048-7333(95)00840-3

Manikandan S, Krishnan RY, Vickram S, et al (2023) Emerging nanotechnology in renewable biogas production from biowastes: Impact and optimization strategies – a review. Renewable and Sustainable Energy Reviews 181:113345. https://doi.org/10.1016/j.rser.2023.113345

Maria MR, Ballini R, Souza RF (2023) Evolution of green finance: A bibliometric analysis through complex networks and machine learning. Sustainability 15(2):967. https://doi.org/10.3390/su15020967

Martinelli A, Nomaler O (2014) Measuring knowledge persistence: a genetic approach to patent citation networks. Journal of Evolutionary Economics 24(3):623–652. https://doi.org/10.1007/s00191-014-0349-5

Martín-Espejo JL, Gandara-Loe J, Odriozola JA, et al (2022) Sustainable routes for acetic acid production: Traditional processes vs a low-carbon, biogas-based strategy. Science of The Total Environment 840:156663. https://doi.org/10.1016/j.scitotenv.2022.156663

Masago FK (2022) Methodologies to aid the evaluation of bioenergy research areas through patents documents. Phd thesis, Universidade de Campinas, Campinas, São Paulo

Metcalfe JS (1998) Evolutionary Economics and Creative Destruction. No. 1 in The Graz Schumpeter Lectures, Routledge, London

Miranda JCdC, Souza RF, Justi ACA, et al (2025) The landscape of green and biohydrogen technology: A data-driven exploration using non-supervised methods. International Journal of Hydrogen Energy 105:1467–1490. https://doi.org/10.1016/j.ijhydene.2025.01.089

Nelson RR, Winter SG (1982) An Evolutionary Theory of Economic Change. Belknap Press of Harvard University Press, Cambridge, MA

Park H, Magee CL (2017) Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. PLOS ONE 12(1):e0170895. https://doi.org/10.1371/journal.pone.0170895

Pavitt K (1984) Sectoral patterns of technical change: Towards a taxonomy and a theory. Research Policy 13(6):343–373. https://doi.org/10.1016/0048-7333(84)90018-0

Pielou EC (1966) The measurement of diversity in different types of biological collections. Journal of Theoretical Biology 13:131–144. https://doi.org/10.1016/0022-5193(66)90013-0

Porter AL, Chiavetta D, Newman NC (2020) Measuring tech emergence: A contest. Technological Forecasting and Social Change 159:120176. https://doi.org/10.1016/j.techfore.2020.120176

Preis R (1999) Linear time 1/2-approximation algorithm for maximum weighted matching in general graphs. In: STACS 99: 16th Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science, vol 1563. Springer, p 259–269, https://doi.org/10.1007/3-540-49116-3_24

Priem J, Piwowar H, Orr R (2022) Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:220501833 https://doi.org/10.48550/arXiv.2205.01833

Priestley M, Sluckin TJ, Tiropanis T (2020) Innovation on the web: the end of the s-curve? Internet Histories 4(4):390–412. https://doi.org/10.1080/24701475.2020.1747261

Qi Z, Han Y, Afrane S, et al (2023) Patent mining on soil pollution remediation technology from the perspective of technological trajectory. Environmental Pollution 316:120661. https://doi.org/10.1016/j.envpol.2022.120661

R Core Team (2025) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Ranaei S, Suominen A, Porter A, et al (2019) Evaluating technological emergence using text analytics: two case technologies and three approaches. Scientometrics 122(1):215–247. https://doi.org/10.1007/s11192-019-03275-w

Roberts ME, Stewart BM, Tingley D (2019) stm: An r package for structural topic models. Journal of Statistical Software 91(2). https://doi.org/10.18637/jss.v091.i02

Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for idf. Journal of Documentation 60(5):503–520. https://doi.org/10.1108/00220410410560582

Romer PM (1986) Increasing returns and long-run growth. Journal of political economy 94(5):1002–1037. https://doi.org/10.1086/261420

Rosenberg N (1982) Inside the Black Box: Technology and Economics, 1st edn. Cambridge University Press, United Kingdom

Rotolo D, Hicks D, Martin BR (2015) What is an emerging technology? Research Policy 44(10):1827–1843. https://doi.org/10.1016/j.respol.2015.06.006

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salton G, McGill MJ (1983) Introduction to modern information retrieval. McGraw-Hill, New York

Schumpeter JA (1942) Capitalism, Socialism and Democracy. Harper & Brothers, New York, NY

Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shibata N, Kajikawa Y, Takeda Y, et al (2008) Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation 28(11):758–775. https://doi.org/10.1016/j.technovation.2008.03.009

Shibata N, Kajikawa Y, Takeda Y, et al (2009) Comparative study on methods of detecting research fronts using different types of citation. Journal of the American Society for Information Science and Technology 60(3):571–580. https://doi.org/10.1002/asi.20994, URL http://onlinelibrary.wiley.com/doi/10.1002/asi.20994/abstract

Shibata N, Kajikawa Y, Takeda Y, et al (2011) Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. Technological Forecasting and Social Change 78(2):274–282. https://doi.org/http://dx.doi.org/10.1016/j.techfore.2010.07.006, URL http://www.sciencedirect.com/science/article/pii/S004016251000154X

Shin H, Kim K, Kogler DF (2022) Scientific collaboration, research funding, and novelty in scientific knowledge. PLOS ONE 17(7):e0271678. https://doi.org/10.1371/journal.pone.0271678

Small H, Boyack KW, Klavans R (2014) Identifying emerging topics in science and technology. Research Policy 43(8):1450–1467. https://doi.org/10.1016/j.respol.2014.02.005

Solow RM (1956) A contribution to the theory of economic growth. The Quarterly Journal of Economics 70(1):65. https://doi.org/10.2307/1884513

Souza RF (2016) Detecção de linhas de pesquisas emergentes em redes de publicações científicas. PhD thesis, Universidade Estadual de Campinas

Souza RF (2026) birddog: Sniffing Emergence and Trajectories in Academic Papers and Patents. URL https://github.com/roneyfraga/birddog/releases/tag/v2.0.0, R package version 2.0.0

Souza RF, Ballini R, Silveira JMFJ, et al (2021) Is entrepreneurship an emerging area of research? a computational response. REGEPE - Revista de Empreendedorismo e Gestão de Pequenas Empresas https://doi.org/10.14211/ibjesb.e1742

Souza RF, Miranda JCdC, da Silva Júnior JJ, et al (2022) Bioenergy research in brazil: A bibliometric evaluation of the bioen research program. Biofuels, Bioproducts and Biorefining 17(1):191–206. https://doi.org/10.1002/bbb.2441

Stephan PE (1996) The economics of science. Journal of Economic literature 34(3):1199–1235

Sun X, Chen N, Ding K (2022) Measuring latent combinational novelty of technology. Expert Systems with Applications 210:118564. https://doi.org/10.1016/j.eswa.2022.118564

Traag VA, Waltman L, van Eck NJ (2019) From louvain to leiden: guaranteeing well-connected communities. Scientific Reports 9(1). https://doi.org/10.1038/s41598-019-41695-z

Utterback JM (1996) Mastering the Dynamics of Innovation. Harvard Business School Press, Boston, MA

Verspagen B (2007) Mapping technological trajectories as patent citation networks: a study on the history of fuel cell research. Advances in Complex Systems 10(01):93–115. https://doi.org/10.1142/s0219525907000945

Volberda HW, Lewin AY (2003) Co-evolutionary dynamics within and between firms: From evolution to co-evolution. Journal of management studies 40(8):2111–2136. https://doi.org/10.1046/j.1467-6486.2003.00414.x

Watanabe I, Takagi S (2021) Technological trajectory analysis of patent citation networks: Examining the technological evolution of computer graphic processing systems. The Review of Socionetwork Strategies 15(1):1–25. https://doi.org/10.1007/s12626-020-00066-1

Weiland P (2010) Biogas production: current state and perspectives. Applied Microbiology and Biotechnology 85(4):849–860. https://doi.org/10.1007/s00253-009-2246-7

Wijffels J (2025) udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. https://doi.org/10.32614/CRAN.package.udpipe, URL https://CRAN.R-project.org/package=udpipe, r package version 0.8.12

Witt U (1998) Imagination and leadership – the neglected dimension of an evolutionary theory of the firm. Journal of Economic Behavior & Organization 35(2):161–177. https://doi.org/10.1016/s0167-2681(98)00058-4

Wu H, Cheng J, Huang S, et al (2014) Path problems in temporal graphs. Proceedings of the VLDB Endowment 7(9):721–732. https://doi.org/10.14778/2732939.2732945

Xu S, Hao L, An X, et al (2019) Review on emerging research topics with key-route main path analysis. Scientometrics 122(1):607–624. https://doi.org/10.1007/s11192-019-03288-5

Xu S, Wang C, An X, et al (2022) A novel developmental trajectory discovery approach by integrating main path analysis and intermediacy. Journal of Information Science 50(3):651–672. https://doi.org/10.1177/01655515221101835

Yeo W, Kim S, Lee JM, et al (2013) Aggregative and stochastic model of main path identification: a case study on graphene. Scientometrics 98(1):633–655. https://doi.org/10.1007/s11192-013-1140-3

Zhang L, Tan R, Peng Q, et al (2023) A holistic method for radical concept generation based on technological evolution: A case application of dc charging pile. Computers & Industrial Engineering 179:109213. https://doi.org/10.1016/j.cie.2023.109213

Appendix A. Publication growth

Figure A.10 shows the annual number of biogas publications in the corpus (2000–2025; 73,885 of the 77,347 documents fall in this window, the remainder predating 2000) with an exponential fit $y(t) = b_0 e^{b_1(t-2000)}$. The estimates are $b_1 = 0.076$ (s.e. 0.008), an annual growth rate of 7.9% and a doubling time of 9.1 years. The fit summarizes the long run; the year-by-year profile is sigmoidal, with fast growth through the 2010s, a plateau around 4,800–5,100 documents per year in 2019–2024, and an uptick in 2025, consistent with a field approaching paradigmatic maturity while its emergent fronts keep growing.

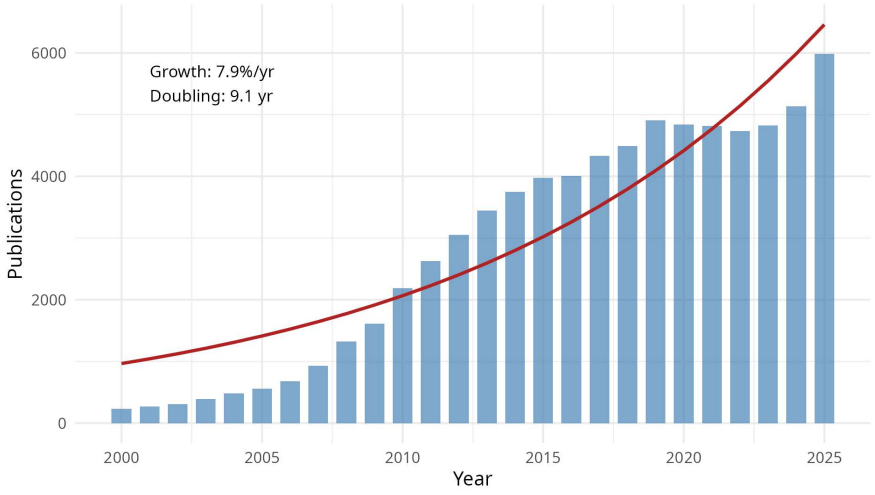


Figure A.10: Annual biogas publications (2000–2025) with an exponential growth fit: 7.9% per year, doubling time 9.1 years.

Appendix B. Geographic distribution

Figure B.11 shows the country composition of each community's central trajectory across the fifteen most productive countries. China and India

dominate the large cores, but leadership is community-specific and tracks Table 4: Indonesia leads the agro-industrial-effluent community (c1g7) and the local-residues community (c1g16), Brazil the landfill community (c1g14), Germany the energy-crops community (c1g13), Italy the bioplastics niche (c1g17), and the United States the methanotroph-biotechnology community (c1g15).

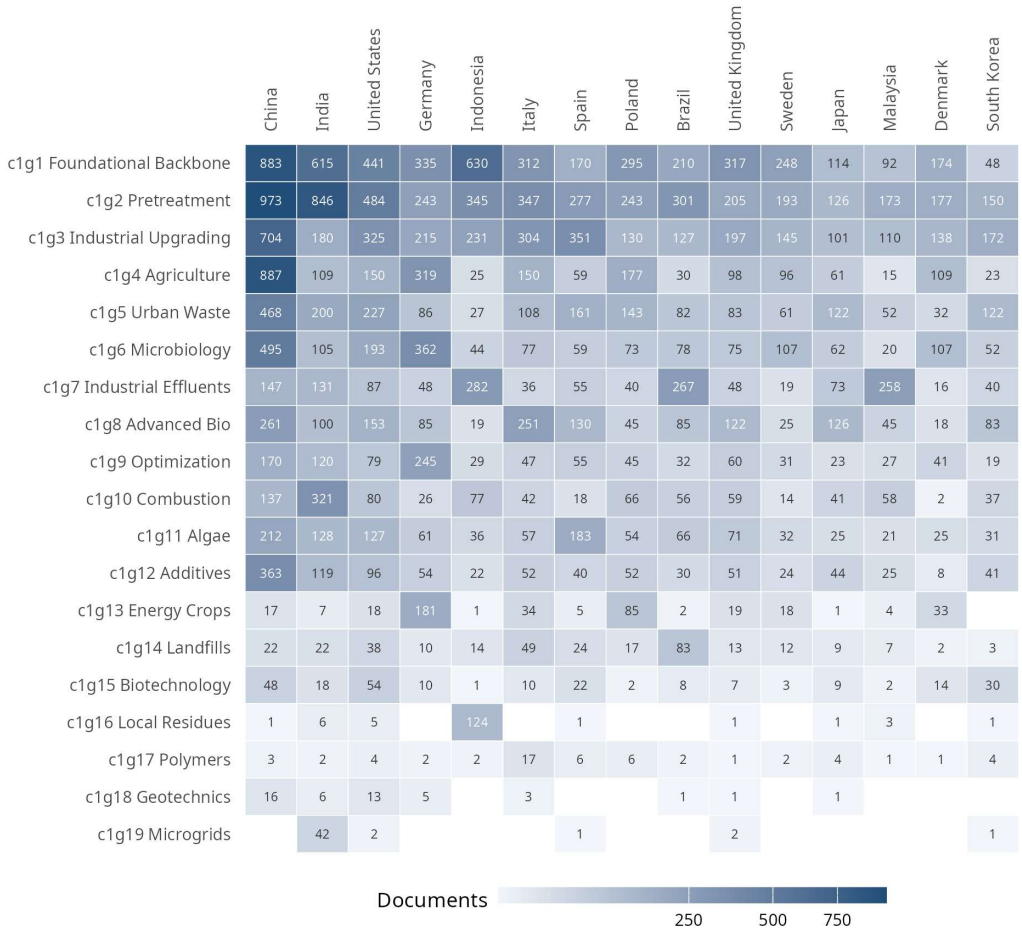


Figure B.11: Documents of each community's central trajectory by country (fifteen most productive countries; square-root colour scale).

Appendix C. Profiles of the nineteen communities

Table C.7 profiles each final community with its short label, its five most distinctive terms (group-level tf–idf over titles and abstracts), and a one-line description.

Table C.7: Profiles of the nineteen final communities

Group	Label	Top terms	Description
c1g1	Foundational Backbone	cross ref; biogas plants; household biogas; biogas plant; renewable energy	Foundational research backbone on biogas production and utilization pathways, from household digesters in the developing world to plant-scale renewable energy systems.
c1g2	Pretreatment	food waste; water hyacinth; methane yield; biogas production; rice straw	Pretreatment of lignocellulosic biomass, anaerobic co-digestion of organic wastes, and biomethane potential (BMP) protocols to enhance methane yields.
c1g3	Industrial Upgrading	gas separation; swing adsorption; biotrickling filter; biogas upgrading; separation performance	Technical processes for biogas purification, upgrading to biomethane, and removal of trace contaminants.
c1g4	Agriculture	biogas slurry; chemical fertilizer; biogas residues; biogas digestate; organic amendments	Valorization of digestate and biogas slurry as organic fertilizer and soil conditioner, assessing nutrient availability, crop response, and environmental risks.
c1g5	Urban Waste	anaerobic membrane; thermal hydrolysis; anaerobic membrane bioreactor; wastewater treatment; sewage sludge	Anaerobic digestion of sewage sludge and municipal wastewater, with sludge pretreatment and anaerobic membrane bioreactors for energy-positive treatment.
c1g6	Microbiology	microbial community; microbial communities; syntrophic acetate; acetate oxidation; syntrophic acetate oxidation	Metagenomic and microbial community analysis to understand the biological pathways of methanogenesis.

Group	Label	Top terms	Description
c1g7	Industrial Effluents	palm oil; oil mill effluent; palm oil mill effluent; mill effluent; oil mill	Anaerobic treatment of agro-industrial effluents such as palm oil mill effluent (POME), sugarcane vinasse, and distillery stillage, including fermentative hydrogen production.
c1g8	Advanced Bio	oxide fuel; solid oxide fuel; solid oxide; oxide fuel cell; solid oxide fuel cell	Biogas reforming into hydrogen and syngas, with catalyst design and integration into solid oxide fuel cells for high-efficiency power generation.
c1g9	Optimization	neural network; digestion model; anaerobic digestion model; prediction of biogas; artificial neural network	Process modeling and optimization of anaerobic digestion, from mechanistic models (ADM1) and mixing studies to artificial neural network prediction.
c1g10	Combustion	dual fuel; fuel mode; diesel engine; ignition engine; equivalence ratio	Evaluation of biogas combustion, thermal efficiency, and emission characteristics in dual-fuel and spark-ignition internal combustion engines.
c1g11	Algae	microalgal biomass; microalgae biomass; photosynthetic biogas; chlorella sp.; c. vulgaris	Microalgae as substrate and partner for biogas: digestion of algal biomass and algal-bacterial photobioreactors for digestate treatment and photosynthetic upgrading.
c1g12	Additives	interspecies electron transfer; vortex layer; direct interspecies; direct interspecies electron transfer; layer apparatus	Additive enhancement of anaerobic digestion through trace element supplementation, biochar, and iron-based conductive nanoparticles promoting direct interspecies electron transfer.
c1g13	Energy Crops	cup plant; matter yield; silphium perfoliatum; dry matter yield; sweet sorghum	Energy crops for biogas, including maize, grasses, cup plant, and sweet sorghum, evaluated through crop rotations, dry matter yields, and green biorefinery concepts.

Group	Label	Top terms	Description
c1g14	Landfills	landfill gas; landgem model; solid waste landfills; waste landfills; landfill site	Landfill gas generation and recovery, including methane estimation models, leachate recirculation, and energy recovery from municipal solid waste landfills.
c1g15	Biotechnology	single cell protein; cell protein; microbial protein; methanol production; mp production	Biotechnological conversion of methane and biogas by methanotrophic bacteria into value-added products such as single-cell protein, methanol, and ectoine.
c1g16	Local Residues	cair tahu; limbah cair; limbah cair tahu; industri tahu; limbah tahu	Small-scale digestion of local residues, notably Indonesian tofu-industry wastewater, for household and community energy recovery.
c1g17	Polymers	polylactic acid; thermoplastic starch; biodegradable bioplastics; maize starch; compostable packaging	Degradation and biodegradation of bioplastics and polymers under anaerobic conditions.
c1g18	Geotechnics	microbial denitrification; degree of saturation; liquefaction mitigation; pore pressure; nitrogen gas	Biogenic gas in geotechnical engineering: nitrogen bubbles produced by denitrifying bacteria to desaturate sands and mitigate soil liquefaction; a homonym of energy biogas captured by the query.
c1g19	Microgrids	load frequency; frequency regulation; frequency control; turbine generator; load frequency control	Frequency control and optimal load management for isolated renewable microgrids powered by biogas generators.

Appendix D. The field atlas

Table D.8 lists all 98 trajectories of the decomposition: the nineteen central spines followed by the 79 tributaries in decreasing order of cumulative size, each with its type, state, period, size, absorber (with the absorption year), and a one-line description.

Table D.8: The field atlas: all 98 trajectories of the biogas citation network

ID	Type	State	Years	Size	Into	Description
c1g2	central	maturity	2013–2025	12047		Pretreatment of lignocellulosic biomass, anaerobic co-digestion of organic wastes, and biomethane potential (BMP) protocols to enhance methane yields.
c1g1	central	maturity	2008–2025	10242		Foundational research backbone on biogas production and utilization pathways, from household digesters in the developing world to plant-scale renewable energy systems.
c1g3	central	maturity	2004–2025	6689		Technical processes for biogas purification, upgrading to biomethane, and removal of trace contaminants.
c1g5	central	dormancy	2020–2025	5636		Anaerobic digestion of sewage sludge and municipal wastewater, with sludge pretreatment and anaerobic membrane bioreactors for energy-positive treatment.
c1g6	central	maturity	2012–2025	5126		Metagenomic and microbial community analysis to understand the biological pathways of methanogenesis.
c1g4	central	maturity	2007–2025	4811		Valorization of digestate and biogas slurry as organic fertilizer and soil conditioner, assessing nutrient availability, crop response, and environmental risks.
c1g7	central	maturity	2004–2025	3959		Anaerobic treatment of agro-industrial effluents such as palm oil mill effluent (POME), sugarcane vinasse, and distillery stillage, including fermentative hydrogen production.
c1g8	central	dormancy	2004–2025	3790		Biogas reforming into hydrogen and syngas, with catalyst design and integration into solid oxide fuel cells for high-efficiency power generation.
c1g11	central	maturity	2009–2025	2272		Microalgae as substrate and partner for biogas: digestion of algal biomass and algal-bacterial photobioreactors for digestate treatment and photosynthetic upgrading.
c1g10	central	maturity	2011–2025	2243		Evaluation of biogas combustion, thermal efficiency, and emission characteristics in dual-fuel and spark-ignition internal combustion engines.
c1g9	central	maturity	2021–2025	2234		Process modeling and optimization of anaerobic digestion, from mechanistic models (ADM1) and mixing studies to artificial neural network prediction.
c1g12	central	emergence	2019–2025	2114		Additive enhancement of anaerobic digestion through trace element supplementation, biochar, and iron-based conductive nanoparticles promoting direct interspecies electron transfer.

ID	Type	State	Years	Size	Into	Description
c1g13	central	maturity	2025–2025	714		Energy crops for biogas, including maize, grasses, cup plant, and sweet sorghum, evaluated through crop rotations, dry matter yields, and green biorefinery concepts.
c1g14	central	maturity	2025–2025	473		Landfill gas generation and recovery, including methane estimation models, leachate recirculation, and energy recovery from municipal solid waste landfills.
c1g15	central	maturity	2025–2025	247		Biotechnological conversion of methane and biogas by methanotrophic bacteria into value-added products such as single-cell protein, methanol, and ectoine.
c1g16	central	maturity	2025–2025	193		Small-scale digestion of local residues, notably Indonesian tofu-industry wastewater, for household and community energy recovery.
c1g17	central	emergence	2024–2025	75		Degradation and biodegradation of bioplastics and polymers under anaerobic conditions.
c1g18	central	emergence	2023–2025	59		Biogenic gas in geotechnical engineering: nitrogen bubbles produced by denitrifying bacteria to desaturate sands and mitigate soil liquefaction; a homonym of energy biogas captured by the query.
c1g19	central	maturity	2025–2025	56		Frequency control and optimal load management for isolated renewable microgrids powered by biogas generators.
tr1	absorbed	convergence	2016–2021	7116	c1g5 (2022)	Large-scale community focusing on waste management and anaerobic digestion for fertilizer production.
tr2	absorbed	convergence	2015–2024	6170	c1g1 (2025)	Primary core lineage tracking the general evolution of foundational anaerobic digestion research.
tr3	absorbed	convergence	2024–2024	2555	c1g2 (2025)	Extensive community dedicated to agricultural substrate research and local energy production in Asia.
tr4	absorbed	convergence	2008–2015	2291	tr1 (2016)	Research on technology dissemination barriers and successes in developing regions.
tr5	absorbed	convergence	2009–2022	2170	c1g5 (2023)	Integration of wastewater treatment and membrane bioreactors for enhanced methane recovery.
tr6	absorbed	convergence	2005–2016	1569	tr1 (2017)	Evolution of sewage sludge treatment and management strategies for urban facilities.
tr7	absorbed	divergence	2000–2013	1218	tr4 (2014)	Foundational experiments on cattle manure and the technical limits of mesophilic reactors.
tr8	absorbed	divergence	2018–2018	1168	tr2 (2019)	Macro-level estimates of national bioenergy potential and regional crop management planning.
tr9	absorbed	convergence	2018–2018	900	c1g2 (2019)	Large-scale pretreatment research and methods for stabilizing industrial waste digestion.
tr10	absorbed	convergence	2012–2019	895	tr1 (2020)	Dynamic modeling of anaerobic processes and the development of biological process indicators.

ID	Type	State	Years	Size	Into	Description
tr11	absorbed	convergence	2012–2013	865	c1g4 (2014)	Techno-economic and regional feasibility studies for biogas systems in Europe.
tr12	absorbed	convergence	2021–2021	776	c1g7 (2022)	Focus on multi-waste synergies and distributed energy generation from landfills and wastewater.
tr13	absorbed	convergence	1993–2013	758	c1g2 (2014)	Early path for household digester design and rural energy poverty management in South Asia.
tr14	absorbed	convergence	2018–2018	699	c1g7 (2019)	Infrastructure and regional energy potential mapping for the Brazilian biogas agenda.
tr15	absorbed	divergence	2007–2011	623	tr11 (2012)	Technical focus on energy crop yields, seasonal variability, and silage additives.
tr16	absorbed	convergence	2018–2022	607	c1g1 (2023)	Digestion of local residues and performance monitoring of small-scale decentralized units.
tr17	absorbed	divergence	2000–2011	592	c1g7 (2012)	Foundational anaerobic treatment engineering and wastewater bioreactor succession.
tr18	absorbed	convergence	2014–2014	568	tr2 (2015)	Long-term lineage for agricultural biomass, energy crop silages, and nutrient recovery.
tr19	extinct	dormancy	2015–2017	476	—	Specialized converted gas conversion and methanotrophic bacterial biotechnology.
tr20	absorbed	divergence	2005–2012	470	c1g6 (2013)	Metagenomics and community succession in thermophilic and high-ammonia digestion.
tr21	absorbed	convergence	2019–2021	449	c1g9 (2022)	Computational optimization of reactor mixing patterns and agitator design.
tr22	extinct	dormancy	2012–2015	430	—	Early research on biological gas purification and industrial effluent treatment.
tr23	absorbed	divergence	1993–2010	327	tr17 (2011)	Foundational manure management and the kinetics of agricultural anaerobic reactors.
tr24	absorbed	convergence	1996–2011	290	c1g1 (2012)	Early core lineage of Indian biogas research and rural digester technology dissemination.
tr25	absorbed	divergence	1998–2007	208	tr7 (2008)	Focus on temperature profiles, digestion kinetics, and cold-climate biogas limits.
tr26	absorbed	divergence	2019–2020	206	tr12 (2021)	Nanoparticle integration and bioelectrochemical engineering for process stabilization.
tr27	absorbed	convergence	1992–2008	205	tr24 (2009)	Early research on diverse agricultural residues and microbial degradation pathways.
tr28	absorbed	convergence	1996–2010	188	c1g1 (2011)	Socio-economic studies on adoption factors and rural energy policy in developing nations.
tr29	absorbed	convergence	1993–2007	178	c1g1 (2008)	Foundational research on wastewater-integrated systems and basic mesophilic reactor performance.
tr30	absorbed	convergence	2007–2010	174	c1g1 (2011)	Analysis of specific industrial residues and process safety markers in large-scale plants.

ID	Type	State	Years	Size	Into	Description
tr31	absorbed	divergence	2001–2007	172	tr20 (2008)	Early microbial community analysis using spatial succession and identification methods.
tr32	extinct	dormancy	2022–2023	158	—	Integration of renewable energy sources and decentralized grid stabilization.
tr33	absorbed	divergence	2010–2010	148	tr13 (2011)	Foundational agricultural substrate research focusing on cow dung and local residues in Asia.
tr34	extinct	dormancy	2022–2022	144	—	Advanced biological conversion of methane to methanol and high-value bioproducts.
tr35	absorbed	divergence	1997–2006	117	tr6 (2007)	Early technical limits and economic feasibility of agricultural biogas plants.
tr36	absorbed	convergence	2006–2008	110	tr4 (2009)	Early path for wastewater desulfurization and chemical gas scrubbing methods.
tr37	absorbed	convergence	2000–2006	101	c1g3 (2007)	Foundational studies on algal-bacterial consortia and circular nutrient recycling.
tr38	absorbed	divergence	2006–2007	100	tr7 (2008)	Early research on syngas combustion, thermal efficiency, and engine performance.
tr39	absorbed	divergence	2007–2008	100	tr17 (2009)	Integration of biogas technology into urban wastewater treatment and nutrient recovery.
tr40	absorbed	divergence	2009–2009	99	tr5 (2010)	Foundational studies on high-pressure digestion and membrane separation technology.
tr41	absorbed	convergence	2011–2011	98	c1g3 (2012)	Early path for landfill gas modeling, methane recovery, and leachate treatment.
tr42	extinct	dormancy	2012–2013	98	—	Research on anaerobic filter performance and acidogenesis residue digestion.
tr43	extinct	dormancy	1996–1998	90	—	Foundational experiments on manure management and simple agricultural digesters.
tr44	absorbed	convergence	2006–2008	88	c1g3 (2009)	Regional bioenergy potential mapping and spatial logistics planning for energy crops.
tr45	absorbed	convergence	2018–2018	87	c1g4 (2019)	Early research on antibiotic inhibition in manure and microbial stability markers.
tr46	absorbed	convergence	2009–2009	86	c1g8 (2010)	Foundational biogas engine modification, performance, and dual-fuel mode characteristics.
tr47	absorbed	divergence	2004–2004	82	tr7 (2005)	Early metagenomic analysis of microbial consortia and methanogenic pathways.
tr48	absorbed	convergence	2010–2010	82	c1g1 (2011)	Socio-economic mapping of distributed energy potential in developing nations.
tr49	extinct	dormancy	2022–2022	81	—	Advanced research on hydrothermal carbonization of biomass and digestate.
tr50	extinct	dormancy	2020–2020	76	—	Optimization of mixing quality and reactor scale-up strategies using CFD.
tr51	absorbed	divergence	2003–2005	69	tr23 (2006)	Early core studies on biochemical methane potential (BMP) standardization.
tr52	absorbed	divergence	2004–2005	67	c1g8 (2006)	Early focus on cogeneration, waste heat utilization, and electricity generation.
tr53	absorbed	divergence	2009–2009	65	tr33 (2010)	Technical limits and performance monitoring for high-rate anaerobic filters.
tr54	absorbed	divergence	2009–2009	65	tr23 (2010)	Foundational research on sewage sludge management and process failure dynamics.

ID	Type	State	Years	Size	Into	Description
tr55	absorbed	divergence	2006–2007	59	tr17 (2008)	Experiments on syngas production, catalytic reforming, and hydrogen-enriched fuels.
tr56	extinct	dormancy	2015–2016	59	—	Early research on antibiotic resistance genes and microbial community succession.
tr57	extinct	dormancy	2015–2015	57	—	Optimization of micro-gas turbine performance and emission characteristics.
tr58	absorbed	divergence	2001–2002	56	tr31 (2003)	Spatial succession of bacteria in mesophilic reactors and process indicator identification.
tr59	extinct	dormancy	2023–2023	49	—	Evaluation of greenhouse gas mitigation and energy recovery from urban landfills.
tr60	absorbed	convergence	2019–2019	45	tr26 (2020)	Brazilian bioenergy estimates and distributed generation models for the Bioethanol industry.
tr61	extinct	dormancy	2023–2023	45	—	Analysis of the circular bioeconomy for industrial and agricultural residues.
tr62	absorbed	divergence	1995–2001	43	tr58 (2002)	Foundational studies on the microbiology of thermophilic anaerobic systems.
tr63	absorbed	divergence	2002–2002	43	tr29 (2003)	Early research on industrial cogeneration systems and high-efficiency heat recovery.
tr64	absorbed	divergence	2008–2008	41	c1g4 (2009)	Early focus on siloxane removal and trace compound filtration in biogas.
tr65	absorbed	convergence	2013–2014	40	tr5 (2015)	Specialized research path for hazardous waste stabilization and leachate treatment.
tr66	extinct	dormancy	2013–2013	38	—	Early path investigating zero-valent iron for enhancing anaerobic methane production.
tr67	extinct	dormancy	2017–2017	38	—	Optimization of compression ratios and ignition timing for hydrogen-enriched biogas engines.
tr68	extinct	dormancy	1996–1996	37	—	Foundational experiments on cattle manure and local waste starters for digestion.
tr69	absorbed	divergence	2001–2001	36	tr17 (2002)	Early technical feasibility of integrated wastewater-to-energy systems in South Asia.
tr70	absorbed	divergence	2002–2002	36	tr37 (2003)	Early path for grassland conservation, mowing management, and solid fuel production.
tr71	absorbed	divergence	2002–2003	35	tr27 (2004)	Foundational research on technology dissemination and rural energy adoption factors.
tr72	absorbed	divergence	2004–2004	35	tr17 (2005)	Early core studies on catalytic steam reforming and syngas for fuel cells.
tr73	extinct	dormancy	2020–2020	35	—	Evaluation of landfill gas potential and environmental risk management in Asia.
tr74	absorbed	divergence	1999–1999	33	tr62 (2000)	Early path for identifying acetate-oxidizing bacteria and their syntrophic interactions.
tr75	absorbed	divergence	2005–2005	32	tr25 (2006)	Early experiments on phase-separated reactors and substrate characterization.
tr76	extinct	dormancy	2015–2015	32	—	Integrated assessment of microgrid frequency stability and renewable energy integration.

ID	Type	State	Years	Size	Into	Description
tr77	extinct	dormancy	1989–1989	31	—	Foundational research on household digester temperature control and rural management.
tr78	extinct	dormancy	1996–1996	30	—	Early core studies on the nutrient value and application of biogas digestate.
tr79	absorbed	convergence	2006–2006	30	tr36 (2007)	Early path for syngas conversion, reforming, and integration with high-efficiency burners.

Appendix E. Lineage structure of the nineteen communities

Figures E.12 and E.13 extend the view of Figure 9 to every community: the central spine against its tributaries as size-by-year curves.

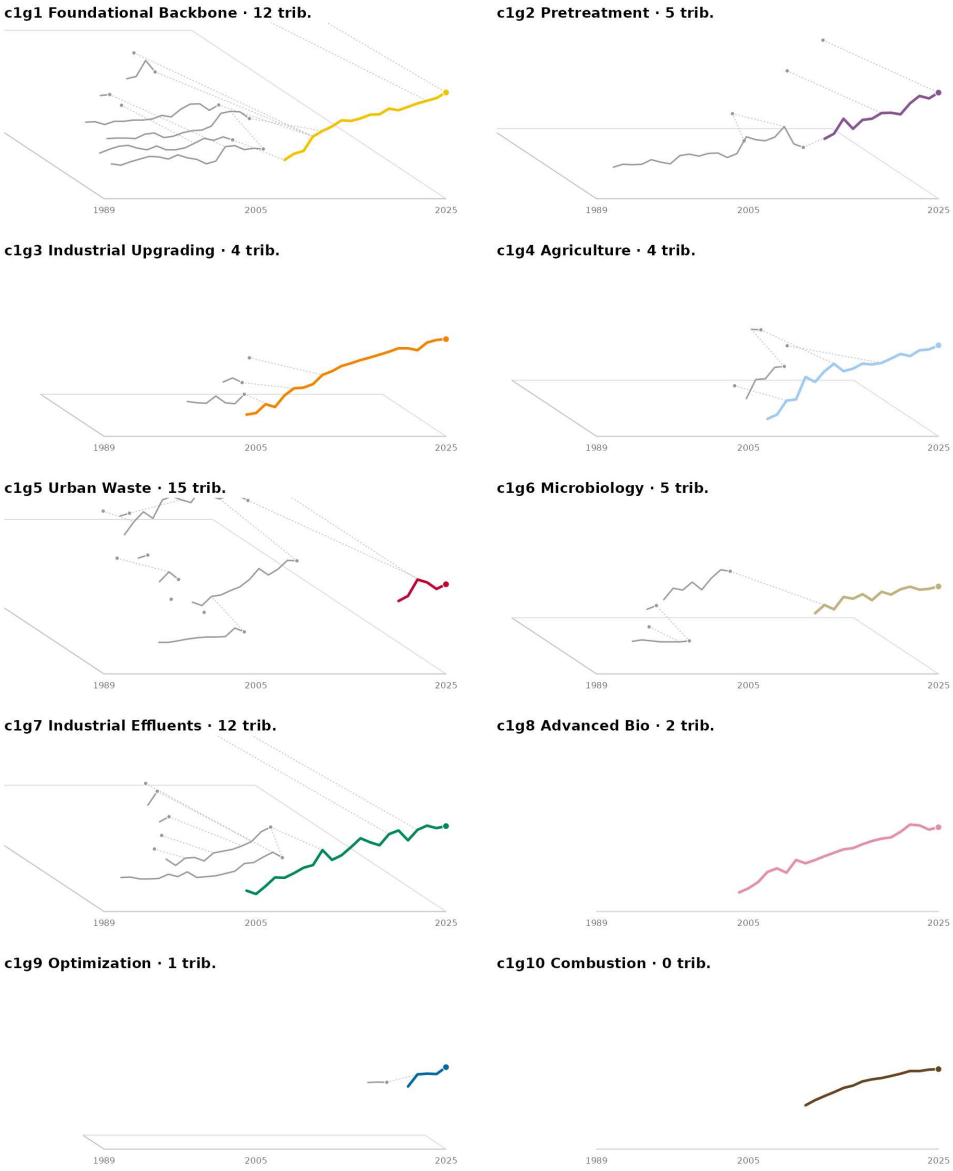


Figure E.12: Lineage growth curves for communities c1g1–c1g10, in the view of Figure 9: the central spine (panel colour) and its feeders (grey) as competing curves on receding lanes, cumulative size by year on the same logarithmic scale and a common 1989–2025 time axis; dotted connectors mark handoffs, dots mark chain ends. As in Figure 9, only feeders delivering more than 1.5% of their cumulative size to the final community are drawn, while each panel's title carries the full tributary count of the braid absorption tree.

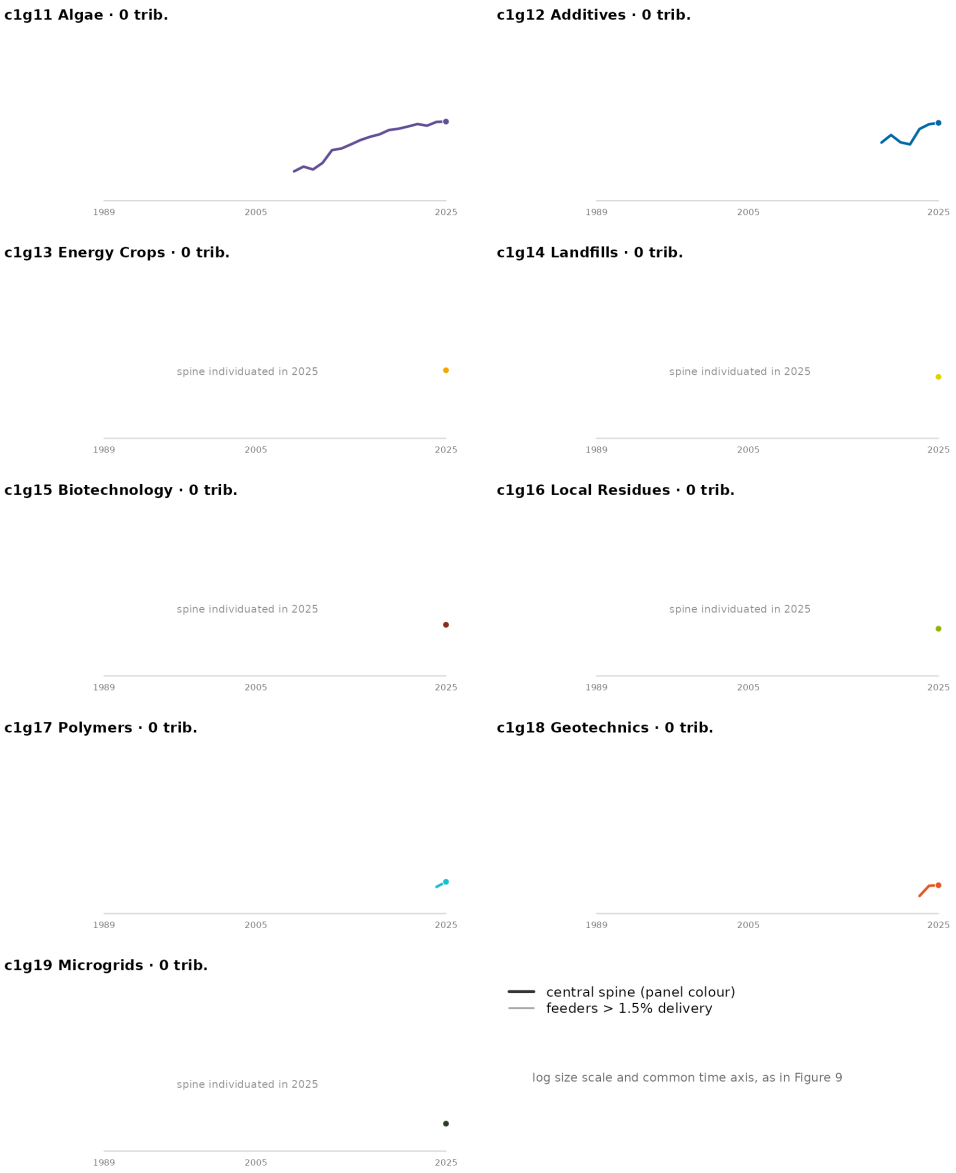


Figure E.13: Lineage growth curves for communities c1g11–c1g19, continuing Figure E.12. The five spines individuated in 2025 appear as single points. Among these smallest final communities, the zero tributary count reflects the admission thresholds rather than isolated origins: their earlier incarnations sit below the $n_{\min}$ node floor, so the DAG admits no roots for a chain to absorb, and the material that built them arrives from lineages absorbed elsewhere (Section 5.6).

Appendix F. Emergence ownership

Table F.9 credits the authors who own the living fronts. The highest total emergence ownership accrues to M. J. Taherzadeh, Budiyono, K. Karimi, I. Angelidaki, and J. D. Murphy. The volume-free normalized score reorders the top of the list only slightly (Karimi, Budiyono, Taherzadeh), while showing that Angelidaki's ownership rides on the largest document count rather than on per-document intensity.

Table F.9: Emergence ownership over the living fronts: the ten authors with the highest total emergence score. The total sums each author's contribution to the emergence of living central trajectories; the normalized score divides by the square root of the author's document count, removing volume (Garner et al 2017).

Author	Total score	Documents	Normalized
M. J. Taherzadeh	247.7	136	21.2
Budiyono	223.9	105	21.8
K. Karimi	211.9	92	22.1
I. Angelidaki	160.3	313	9.1
J. D. Murphy	152.5	125	13.6
V. Dubrovskis	146.7	117	13.6
H. Oechsner	146.7	142	12.3
J. Dach	145.7	80	16.3
H. B. Møller	144.0	116	13.4
X. Li	140.3	94	14.5

Appendix G. Hub documents and their reach across lineages

The Z_iP_i plane of Step 5 (Section 4) places the 36,716 documents of the final communities that have received at least one citation. Both coordinates are computed on citations received (Guimera and Amaral 2005): the within-module degree Z_i standardizes the citations a document receives from its own community, and the participation coefficient P_i measures how evenly its citations spread across communities. A document is a hub when $Z_i \geq 2.5$, and its role is provincial (R5, $P_i \leq 0.30$), connector (R6, $0.30 < P_i \leq 0.75$), or kinless (R7, $P_i > 0.75$). Of the 36,716 documents, 853 (2.3%) are hubs: 553 connectors, 253 provincial hubs, and 47 kinless hubs. Seven in ten hubs (600 of 853) have $P_i > 0.30$, so the papers that anchor a community are mostly recognized by other communities as well. The full list, with the citations each hub receives from every final group and every trajectory, is given in Online Resource 1.

The plane reads citations by final group, which is a static view. The decomposition adds a temporal one: a hub influences a trajectory when documents of that trajectory cite it, each citing document being assigned to the trajectory of the first cluster–year node it joins. On this reading, the median hub is cited from 11 of the 98 trajectories (interquartile range 7–16), and 101 hubs are cited from 20 or more. Reach follows the hub role. The median provincial hub is cited from 4 final groups and 6 trajectories, the median connector from 9 and 13, and the median kinless hub from 12 and 20. These counts include the hub's own group and trajectory. No hub is cited from all nineteen groups; 303 are cited from at least ten.

Hub roles also differ across trajectory states. Here the documents of a trajectory are all those that pass through its nodes. Among the nineteen extinct lineages, the median share of documents that are hubs is 1.1%, against 2.2% for the mature centrals, 3.6% for converging tributaries, and 4.7% for diverging ones; hubs with $P_i > 0.30$ (R6 and R7) carry most of the difference (1.0% against 3.2% between extinct and diverging lineages). Part of this gradient is age and size. Diverging tributaries are the oldest lineages (median start 2004), extinct ones the youngest and smallest (median start 2015, 57 documents), and a document needs time to accumulate the citations that make it a hub. With start year and size held constant, the differences among the three fates of the 79 tributaries keep their sign but are not significant at conventional levels. The hub side shows the same timing. Of the 101 hubs cited from twenty or more trajectories, 57 (56%) entered the network through a tributary that later diverged, against 15% of the remaining hubs, and three in four of them were published by 2010. The field's widely shared base is largely its early literature, and that literature entered through lineages that were later absorbed and whose cohorts dispersed across several communities. These are descriptive medians over 19 to 30 trajectories per state, not tests of a hypothesis.

Four hubs illustrate the range.

- Weiland (2010), a review of biogas production, is the widest-reaching hub in the field: a kinless hub ($P_i = 0.88$) in c1g6 with 1,404 citations from the final communities, cited from eighteen final groups and 41 trajectories, among them tributaries that converged, diverged, and went extinct. It works as a common ancestor of the field rather than of one community.

- Angelidaki et al (2009), the protocol for biomethane-potential batch

assays, has the strongest home anchoring of the four ($Z_i = 22.5$, 422 of its 783 citations from c1g2) and still reaches fifteen groups and 31 trajectories. A measurement standard written in one community and adopted across many is the document-level trace of selection settling on a common practice.

- Hansen et al (1998), on ammonia inhibition in swine-manure digestion, is the oldest of the four. It is a connector hub of c1g6 cited from eleven groups and 30 trajectories, and it carries the inhibition problem from microbiology into pretreatment (121 citations from c1g2) and urban waste (32 from c1g5).

- Chen et al (2010), on household biogas in rural China, is a provincial hub ($P_i = 0.25$): 161 of its 211 citations come from the c1g1 spine itself. It is cited from 15 trajectories, but the other fourteen contribute 50 citations between them. It retains knowledge inside one lineage rather than brokering it across lineages.

Appendix H. A trajectory in depth: biomethane upgrading (c1g3)

The key-route main path is an optional extension module of Step 10 (Section 4); this appendix shows what it adds when one community is read end to end. The community is c1g3, biogas upgrading to biomethane (4,438 documents, spine 2004–2025, recent growth +25%, a net knowledge importer), and it matters beyond the method: once biogas is stripped of CO_2 and contaminants it is a drop-in renewable substitute for fossil natural gas, so this trajectory is the trajectory of a candidate energy-transition pathway.

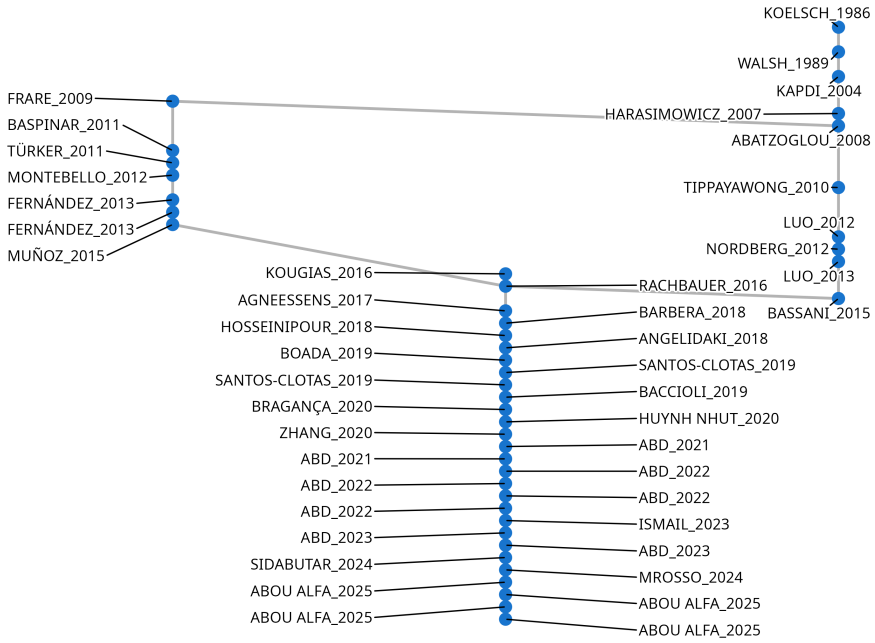


Figure H.14: Within-group key-route main path of c1g3 (biomethane upgrading). Nodes are documents on the highest-traffic citation route (Liu and Lu 2012); labels give the first author and year. The route runs from cogeneration (1986) and biogas cleaning (2004–2008), forks into a desulfurization line (left) and a biological CO_2-methanation line (right) that rejoin in 2016, and ends in biomethane upgrading by pressure-swing adsorption on low-cost sorbents (2021–2025).

The key route (Figure H.14) runs through four phases. It opens with cogeneration from raw biogas (1986–1989) and a cleaning phase that treats biogas as a gas to be stripped of contaminants by scrubbing, membranes, and purification (2004–2008). In 2009 the route forks into two parallel lines. One continues the subtractive programme as desulfurization, increasingly by biological means in anoxic biotrickling filters (2009–2015). The other turns the surplus CO_2 into a feedstock: hydrogen is added so that hydrogenotrophic methanogens convert CO_2 to methane, first inside the digester and then also in separate reactors, in a line of work from I. Angelidaki's group, whose papers on the route are first-authored by Luo, Bassani, and Kougias (2012–2017). The two lines rejoin in 2016. The last phase (2018–2025) turns to grid-

and transport-grade biomethane: trace contaminants such as siloxanes and volatile organic compounds are removed in biotrickling filters (2019–2020), and CO_2 is separated by pressure-swing adsorption (2021–2025), increasingly on low-cost sorbents made from waste such as spent coffee, eggshell, and biochar, with one study integrating upgrading with steam-methane reforming for hydrogen. Over four decades the route treats the CO_2 in biogas first as an impurity, then for a time as a feedstock, and finally as a stream to be separated at lower cost.

Read as a trajectory, c1g3 has strong directionality within a stabilizing paradigm: over the spine's life (2004–2025) its citation cycle time falls from nine to six years by 2011 and stays near six thereafter, knowledge renewing faster as the front industrializes, and its keyword entropy falls steadily from 0.87 to 0.66, the thematic consolidation of a maturing community. With self-sufficiency 0.92 and only four small, early tributaries (the last merging in 2012), c1g3 grew its own line rather than absorbing others: a Dosi trajectory made observable, a coherent, self-propelled line of technical advance that carries biogas from a raw fuel toward a grid- and transport-grade renewable gas.

Appendix I. Reproducibility

Data. OpenAlex (Priem et al 2022), query `title_and_abstract.search = ‘‘( biogas )’’`, retrieved via the `openalexR` interface in June 2026; 77,347 documents published through 2025. The citation network is built on the deduplicated articles and reviews (63,155 documents).

Software. R (R Core Team 2025), `birddog` 2.0.0 (Souza 2026), with the analysis orchestrated as a `targets` pipeline.

Parameters. Louvain community detection with seed 888 and a 30-document community floor, applied with identical settings to the final network and to every cumulative year; temporal DAG with a 10-document node floor ($n_{\min} = 10$), Jaccard floor $\tau_J = 0.05$, two successors kept per node, and document-overlap similarity; growth window 2000–2025. The detector was selected by content coherence (Section 5.9) with a minimum margin of 0.02; braid was retained.

Function map. Table I.10 maps the pipeline steps of Figure 1 to the `birddog` functions that implement them.

Table I.10: Pipeline steps and the `birddog` functions that implement them.

Step	Function(s)
4. Collect data	`read_openalex()` (OpenAlex API), `read_wos()` (exported files)
5. Citation network and groups	network: `sniff_network()`; giant component: `sniff_components()`; groups: `sniff_groups()`; attributes: `sniff_groups_attributes()`; hubs: `sniff_groups_hubs()`; influence: `sniff_groups_influence()`
6. Content (NLP)	`sniff_groups_terms()` (`udpipe` phrases, group-level tf–idf)
7. Subtopics	`sniff_groups_stm_prepare()`, `sniff_groups_stm_run()` (not used in this application)
8. Trajectories	reverse cumulative clustering: `sniff_groups_cumulative()`; lineage: `sniff_groups_lineage()`; DAG: `sniff_trajectory_dag()`; detectors: `sniff_trajectory_braid()`, `sniff_trajectory_channel()`; confluence: `sniff_trajectory_confluence()`; readers: `sniff_trajectory_destination()`, `sniff_trajectory_formation()`, `sniff_trajectory_self_sufficiency()`; states: `sniff_trajectory_dynamics()`; validation: `sniff_trajectory_coherence()`, `sniff_trajectory_comparison()`
9. Indexes	`sniff_cct()`, `sniff_entropy()`; per trajectory: `sniff_trajectory_cct()`, `sniff_trajectory_entropy()`; ownership: `sniff_trajectory_emergence_owners()`; influential documents: `sniff_trajectory_hubs()`
10. Main paths	`sniff_key_route()`, the within-group key-route main path (SPC) (Liu and Lu 2012; Batagelj 2003)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.xlsx