Hunting Breakthroughs in Science

The case of *Quantum Computing in OpenAlex*

Thomas Maillart, Thibaut Chataing & David Dosu

CyberAlp Retreat – 19-20 June 2025 – Davos, Switzerland

OpenAlex

In a nutshell :

- OpenAlex is a free, open index of scholarly papers and related entities.
- Covers authors, institutions, journals, concepts, and more.
- Offers structured, regularly updated data via API
  - ~50 000 new entries added daily

Key facts :

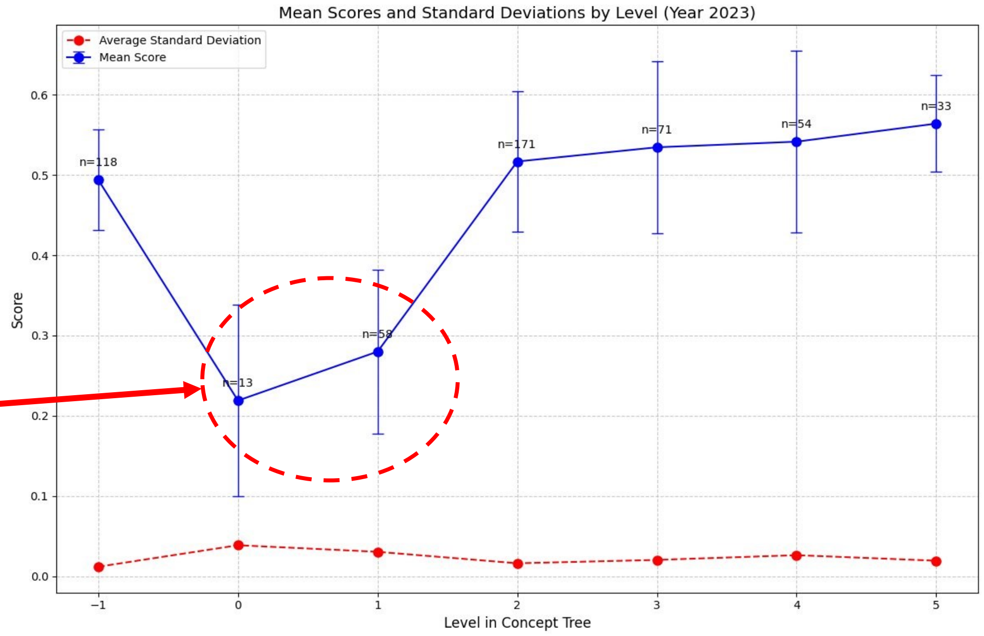
- **~ 240 million works (papers, books, datasets)**
- **~ 213 million author profiles**
- **~ 109 000 institutions**
- **~ 65 000 concepts:** structured as a knowledge network

Importance and limitations of concepts

- Linked to **Wikidata items** for semantic interoperability and external referencing
- Each paper is tagged with multiple concepts, each given a **relevance score (0–1)**
- Scores reflect how central a concept is to a work, enabling precise filtering and analysis
- Helps map the **thematic importance** of topics across millions of papers

⚠ Note : Concepts are marked as deprecated for future development

- Hard to maintain and update
- Tagging can be imprecise or too broad
- Structure doesn't align well with modern ontologies
- OpenAlex is shifting toward more flexible, user-driven classification



Finding a workable definition of *"breakthrough"*

🧠 Qualitative / Formal Criteria

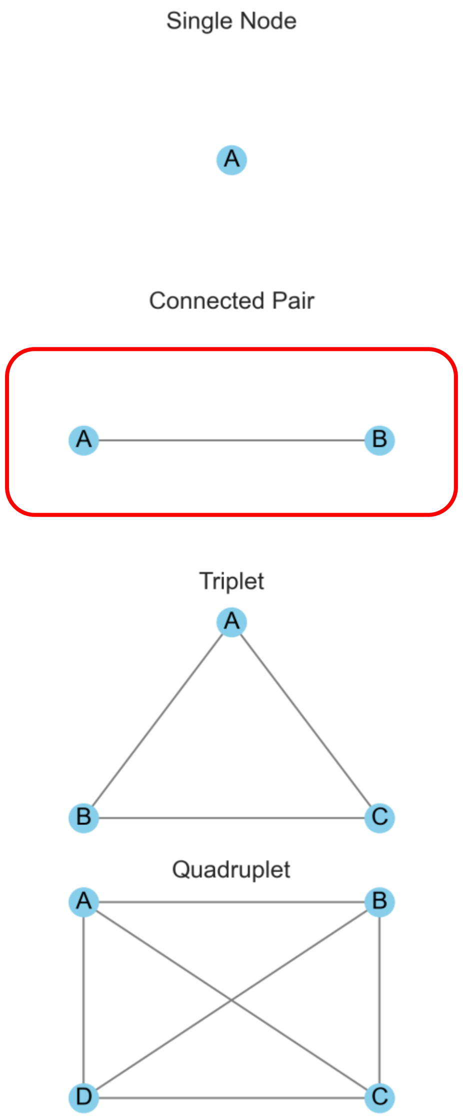
- Introduction of a **new concept** in the semantic network ?
- Novel **combination** of existing concepts (pair, triplet, etc.) ?

=> All may work as they signal **shifts in knowledge structure**

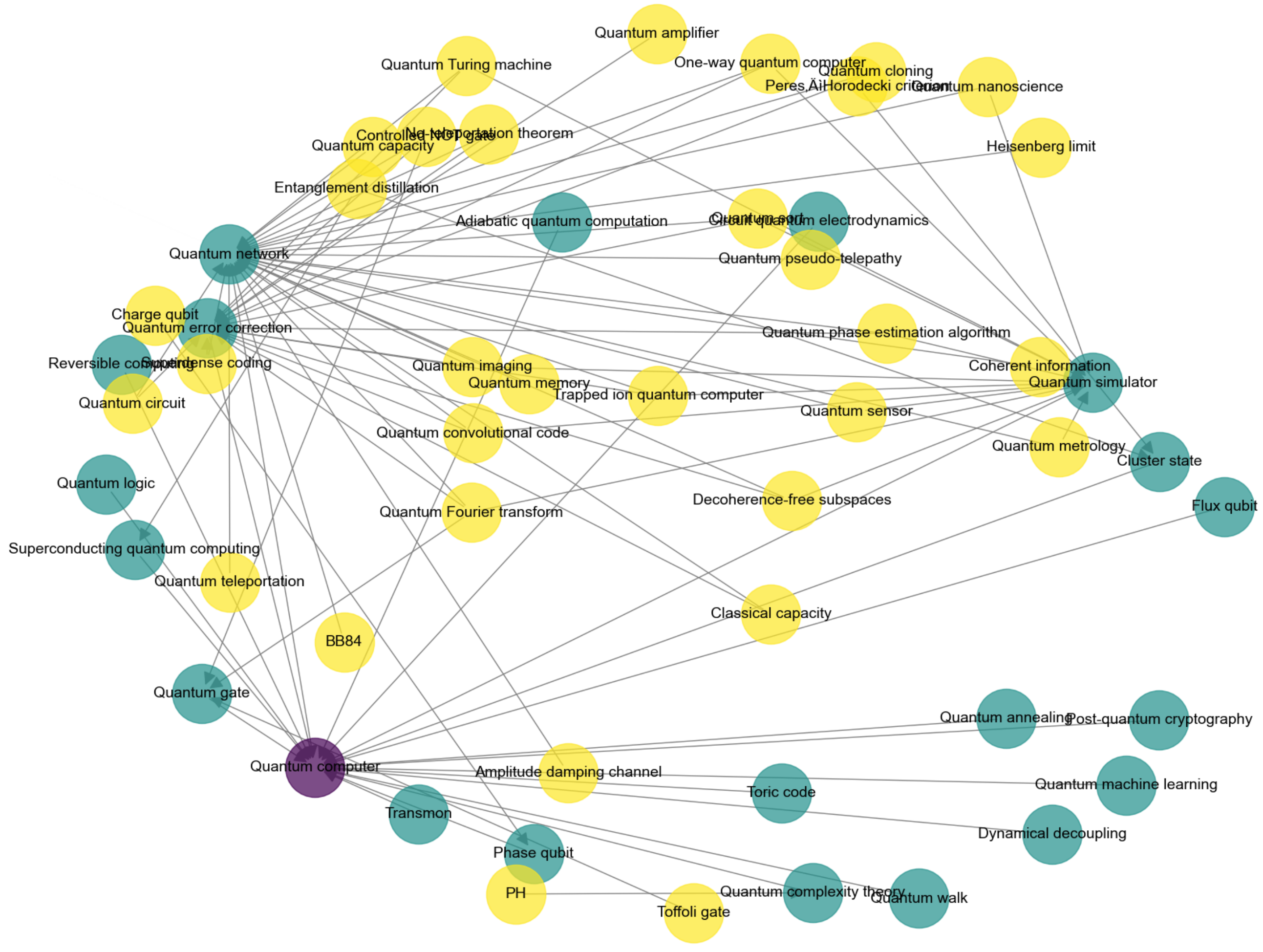
📊 Quantitative Criteria

- **Persistence**: An entity that endures and evolves over time
- **Influence**: Sparks or shapes future breakthroughs

Today, we focus on predicting connected pairs



Case study : Quantum computer (level 3)



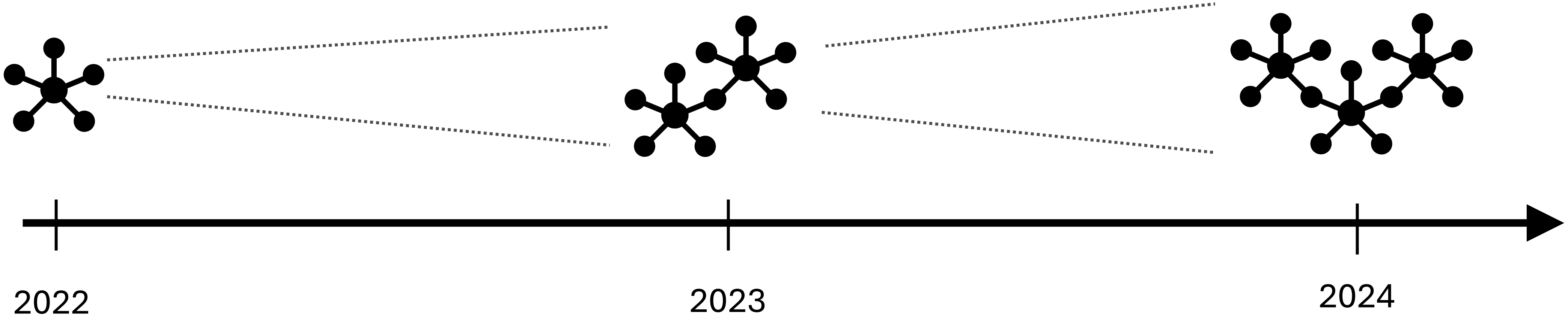
Semantic network of concepts (quantum computer subtree)

- Level 3 : quantum computer
- Level 4
- Level 5

~85 000 papers in "quantum computer" over 33 years (1990 – 2023)

Modeling breakthroughs as pairs of concepts over time

- Concepts → **Nodes**
- Number of papers with 2 concepts (per year, min. 4 papers) → **Weighted edges**
- 1 snapshot per year (1990 – 2023) → **Time**



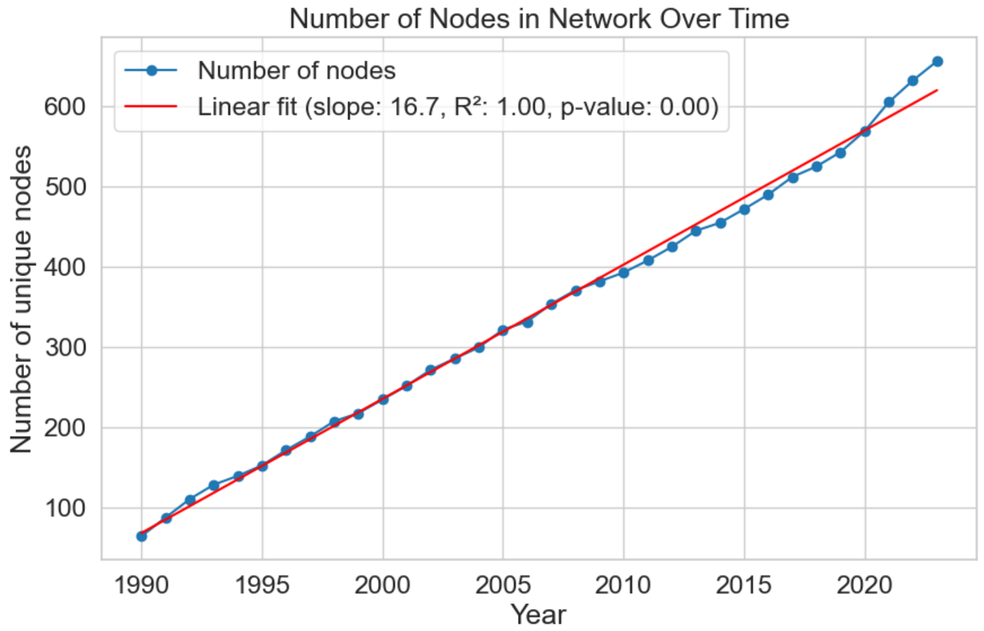
Quantum computer network evolution

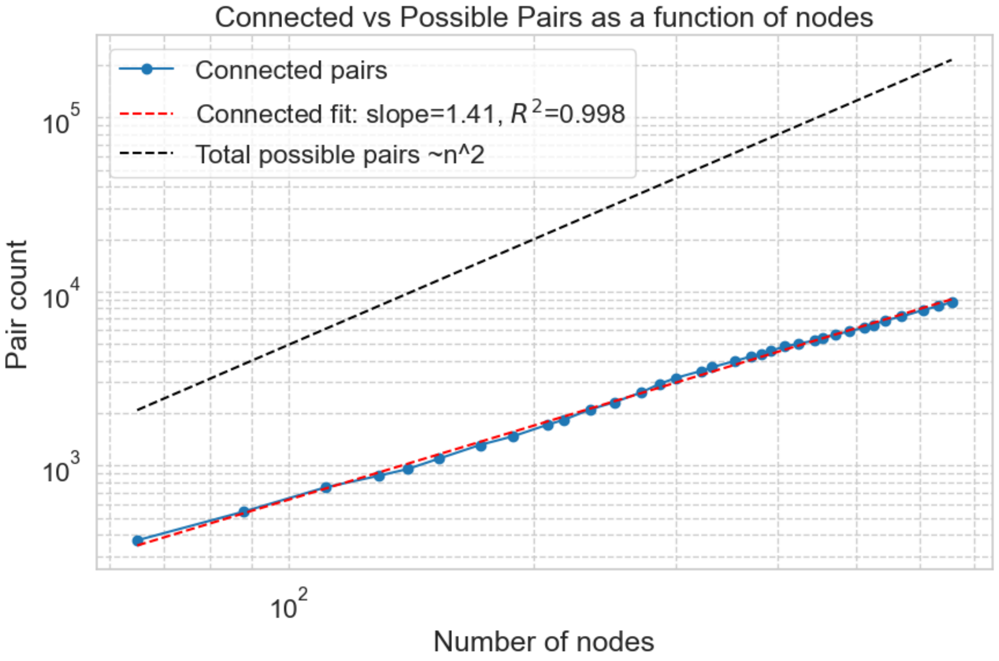
- Nodes have increased linearly
  → 16 nodes/year

- Edges have increased in scaling ***~n^α***,
  with ***1 < α = 1.41 < 2***
  →as the field develops the number of possible pairs largely overshoots the number of pairs actually created.

→ How do we find which “breakthrough” pair will form ?

RESEARCH OBJECTIVE : From the graph at time ***t*** predict links for the graph at ***t+1, t+2, t+ 3,…***





Predict breakthroughs as pairs of concepts

- Objectives :
  - Predict links between two concepts (i.e. edges)
  - Identify key network metrics, reflecting network structures associated with breakthrough

- Strategy :

  Target a wide set of network metrics, including weighted edges, to enrich the understanding of network structures as features to predict the appearance of concept pairs.

Diverse metrics

Graph metrics

- Extract high-level network structures
- Reveal patterns that often persist or evolve in predictable ways.

	Node metrics				**Link metrics**	
Category	Bridging & Flow Centrality	Neighborhood Structure	Structural Importance	Topological Proximity	Connectivity Heuristics	Neighborhood Overlap
Count	3	6	10	4	2	6
Example	Betweenness centrality	Clustering coefficient	Page rank	Closeness centrality	Preferential attachment	Jaccard coefficient

Link prediction : strategy

1. **Compute metrics to qualify the graph**
   - 23 Node metrics (e.g., degree)
   - 8 Link metrics (e.g., Jaccard index)

2. **Feature engineering for node metrics**
   - Absolute difference of node characteristics (between 2 nodes)
   - Hadamar product*:
     - multiplicative interactions between features
     - => reveal nonlinear relationships which cannot be not captured by node features

3. **Drop highly correlated Features**
   - => 59 features

*Hadamar product is the element-wise multiplication of two matrices

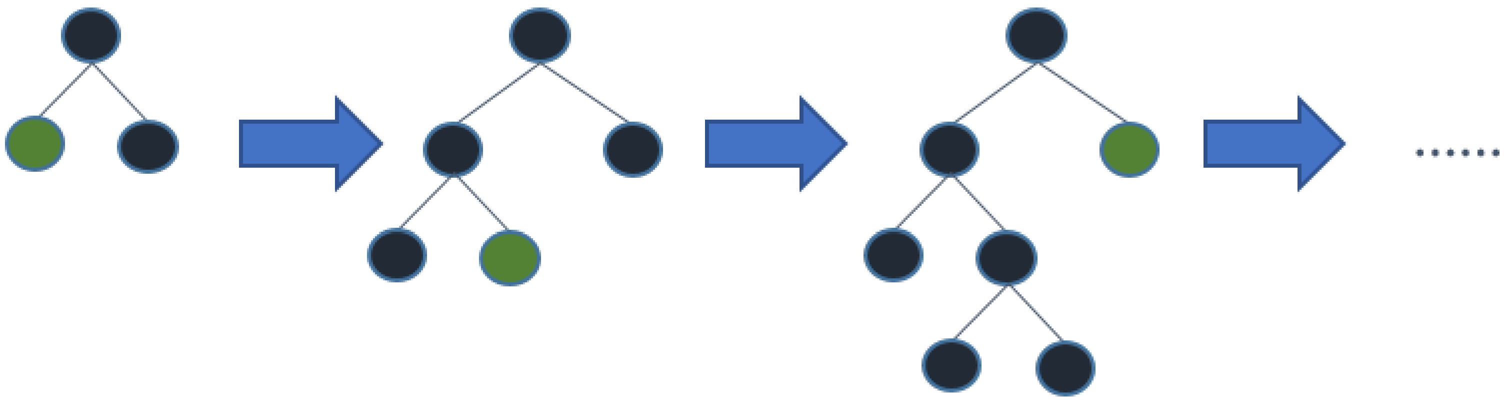
Link prediction : binary classification

- Model : **LGBMClassifier** with dart optimization

  gradient boosting machine to build an ensemble of decision trees

- Fine tuning with **Optuna**

  automatic hyperparameter optimization framework that efficiently searches for the best model parameters using a define-by-run approach



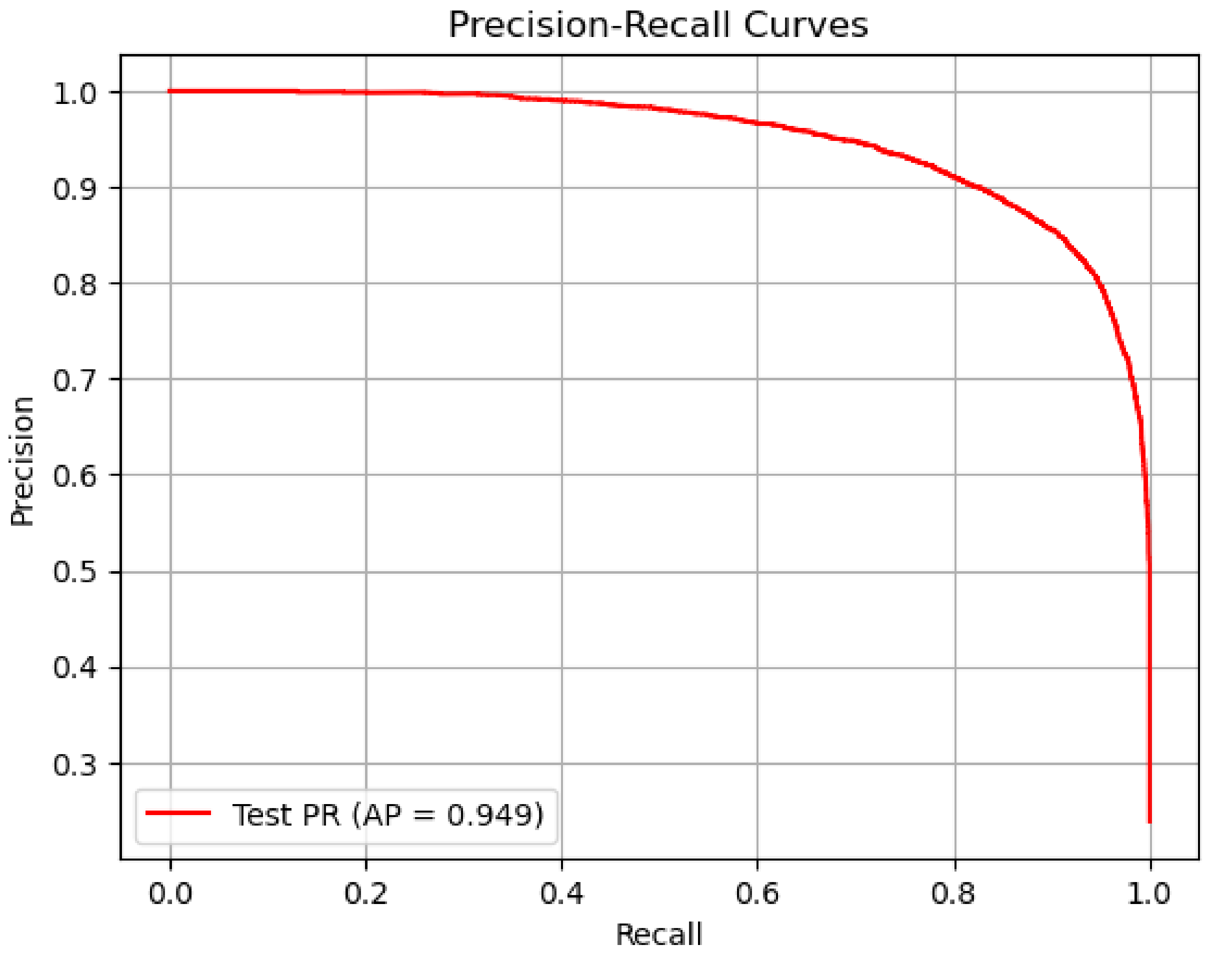
Leaf-wise tree growth

Results I : link prediction t+1

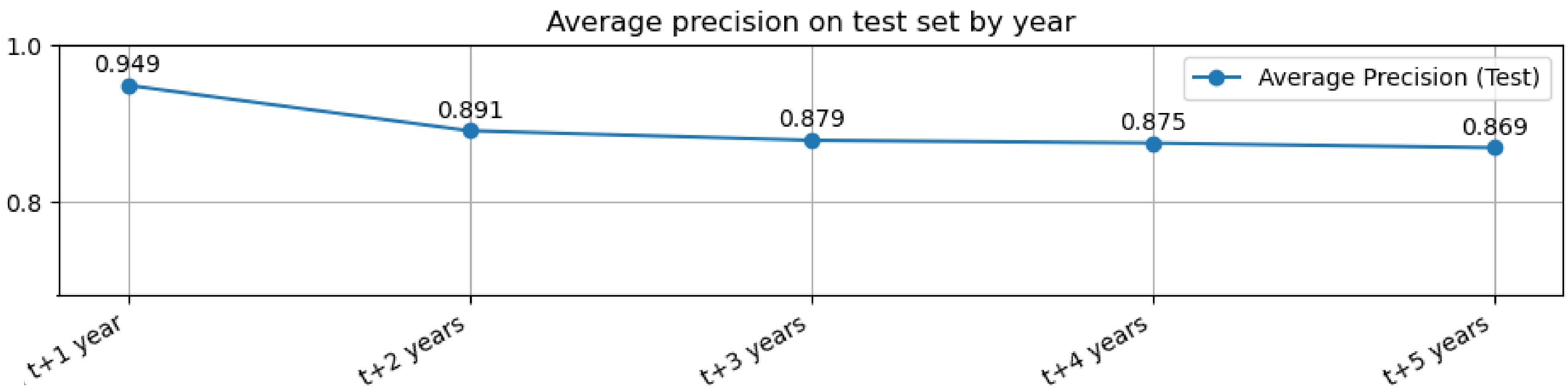
Metrics	Scores
AUC-ROC	0.983
Average Precision	**0.949**
Precision	0.830
Recall	0.926
F1 Score	0.875

✅ **High predictive performance** with AP of 0.9849 and F1 score of 0.875 confirms the model's ability to accurately forecast future concept links.

🔍 **Strong recall (0.926)** ensures most true links are captured, while solid precision (0.830) limits false positives.



Results I : link prediction up to t +5



→ Powerful prediction even over 5 years.

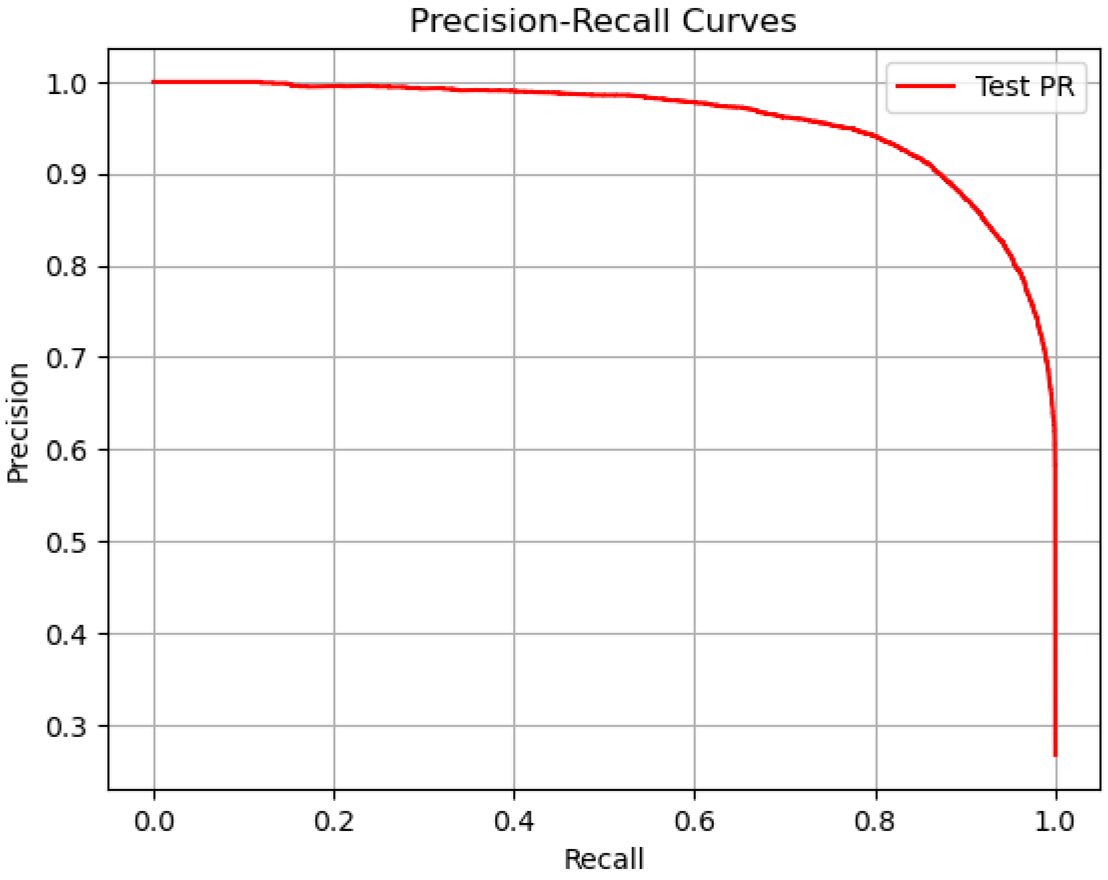
→ Suggests that network structures tend to be relatively deterministic

→ What about are the performance on predicting the appearance/disappearance of link ?

Results I : Focus on state switching link

Over the testing set :

Metrics	Scores
AUC-ROC	0.905
Average Precision	**0.816**
Precision	0.923
Recall	0.836
F1 Score	0.878



✅ **The model finds well when a link appears or disappears**

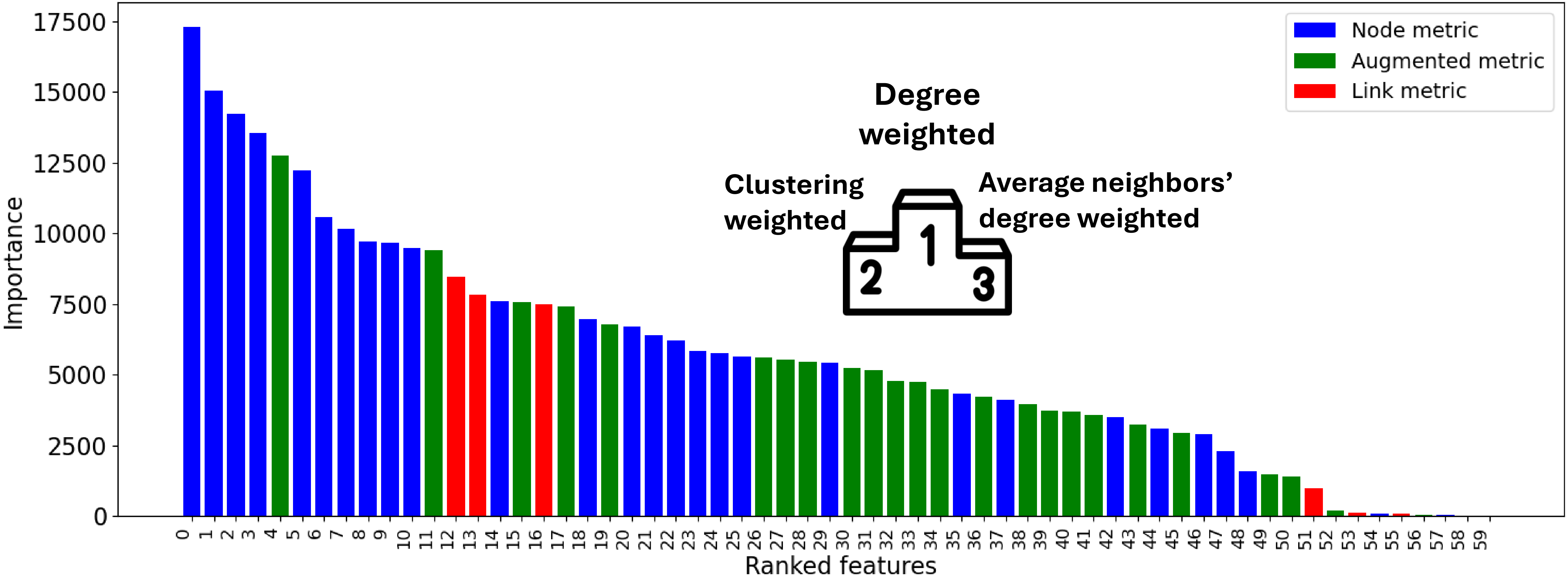
Results I : discussion

- ➢ **Predicting new** or **persistent links** is **easier** and **more accurate**, **showing** that **stable structural patterns** in the graph.

- ➢ **Predicting state switches** (appearance/disappearance) is **more complex**, as it involves subtle or transient dynamics in concept relationships.

- ➢ Overall, the graph's structure provides **stronger signals for continued or emerging links**, while **disappearing links are less structurally predictable**.

Results II : most predictive features



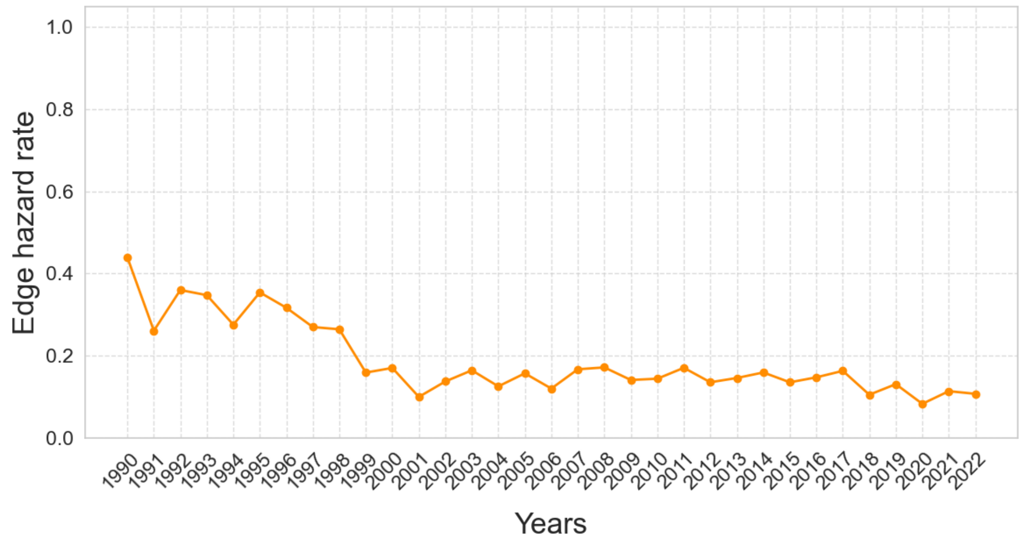
Results II : discussion

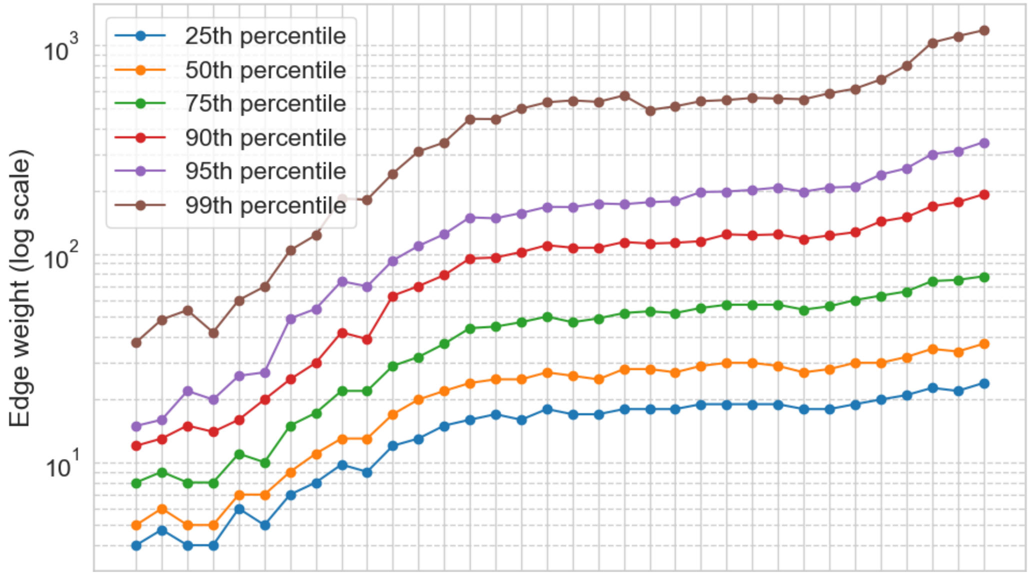
- Top 3 metrics provide complementary insights:
  - **Degree and neighbor-based metrics** signal centrality and influence,
  - **Clustering metrics** highlight local cohesiveness and research stability,
  - **Average neighbors' degree** uncovers hidden thematic structures.

=> These results suggest a strong emphasis on **local cohesion** (clustering), **tie strength**, and **higher-order connectivity** as key indicators that a node is embedded in an **active and tightly-knit subnetwork**, which increases its likelihood of forming new links.

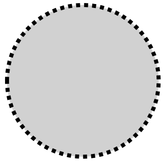
Ongoing work & next steps

- link/node (dis)appearance (hazard rate)
- pair weights
- triplets/quadruplets
- upstream/downstream citations
- waiting times between publications
- refined understanding of network structures associated with breakthroughs
- breakthroughs across fields

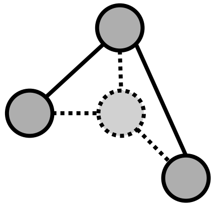




Next step predicting node (dis)appearance : A structural perturbation analysis



New nodes (concepts) cannot be directly predicted, as they represent **novel entries** in the system with no historical trace



Structural disturbances in the graph can act as **early signals** of **emerging** or **disappearing** concepts.

➢ **Hypothesis**: Rather than forecasting the node itself, we model the **structural imbalance it provokes**, enabling indirect detection of conceptual innovation or decline.

Hunting Breakthroughs in Science

The case of *Quantum Computing in OpenAlex*

Thomas Maillart, Thibaut Chataing & David Dosu

CyberAlp Retreat – 19-20 June 2025 – Davos, Switzerland