Et si on comprenait la structure de graphes de connaissances comme Wikidata ?

Hassan Abdallah, Béatrice Markhoff, and Arnaud Soulet

Université de Tours, LIFAT, Blois

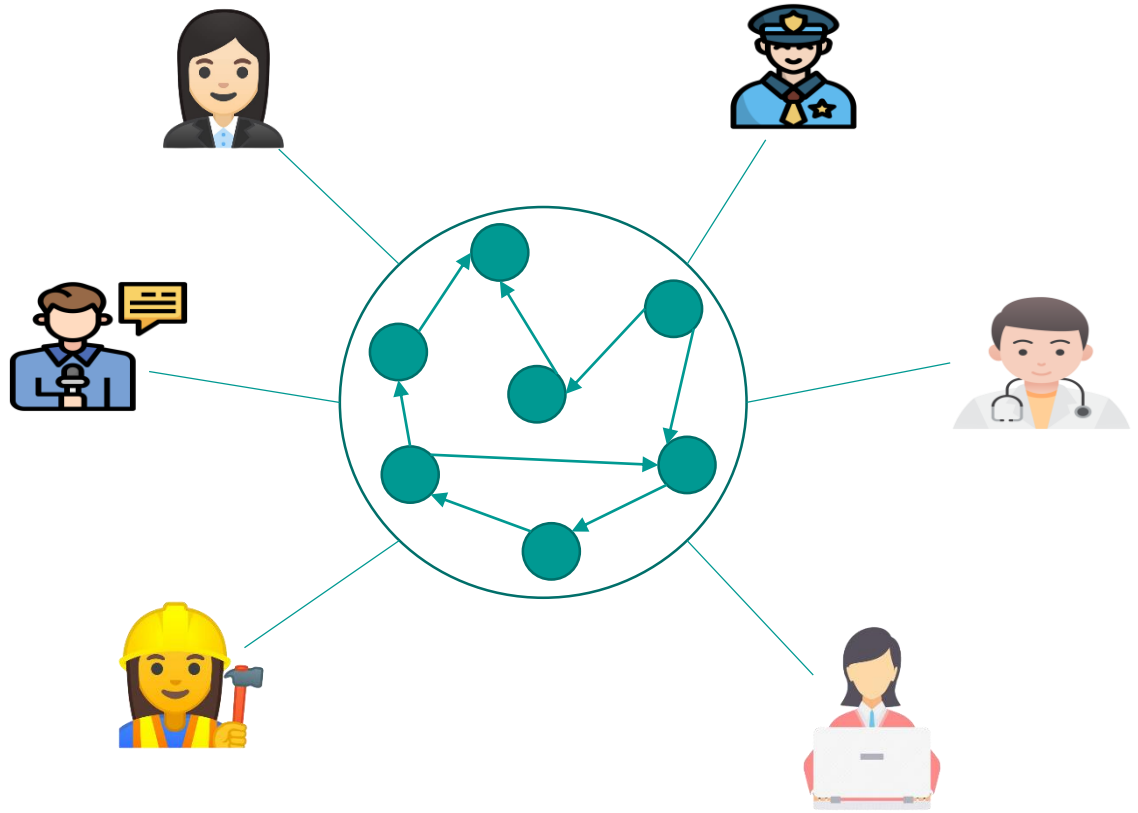
IC 2023 / 4-July-2023

Content

1. **Positioning idea**
2. **Benefits**
3. **Challenges**

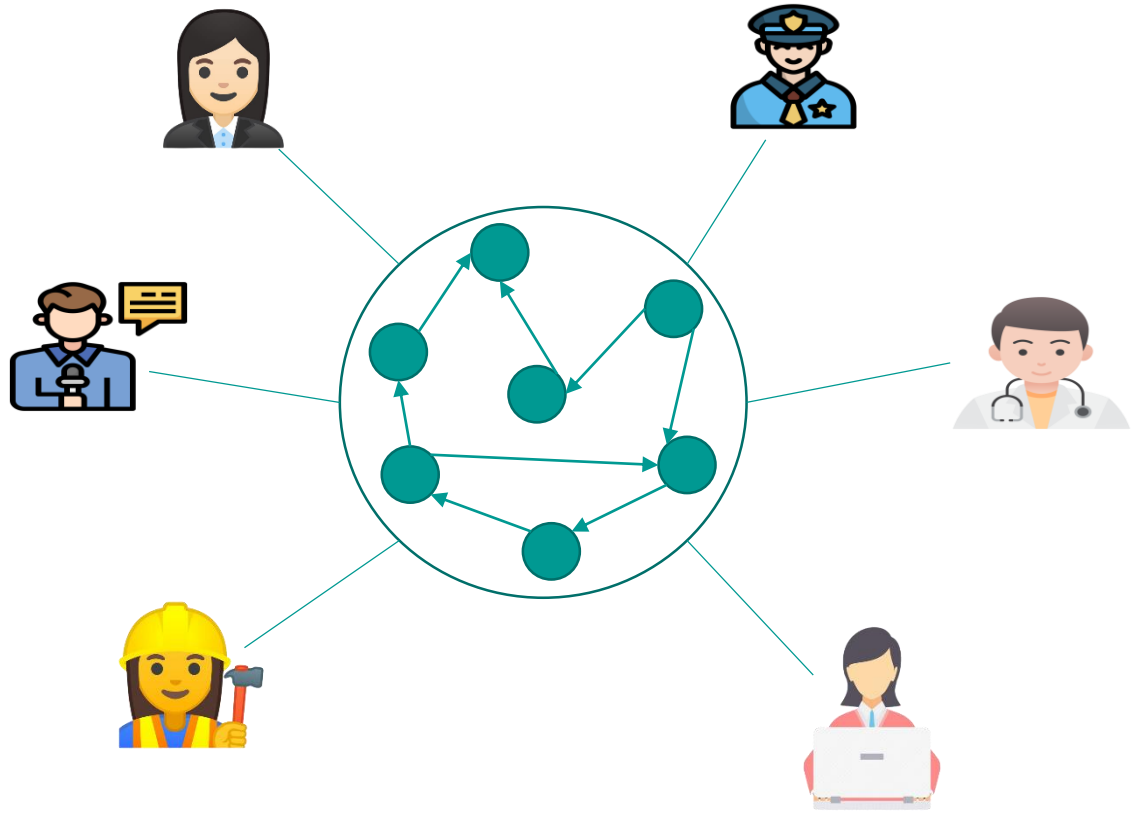
Crowd-sourced Knowledge graphs

- Created through the collective efforts of a diverse group of individuals
- The graph that emerges is a multidisciplinary knowledge graph
- Examples of crowdsourced KGs: **DBpedia**, **YAGO4**, **Wikidata**



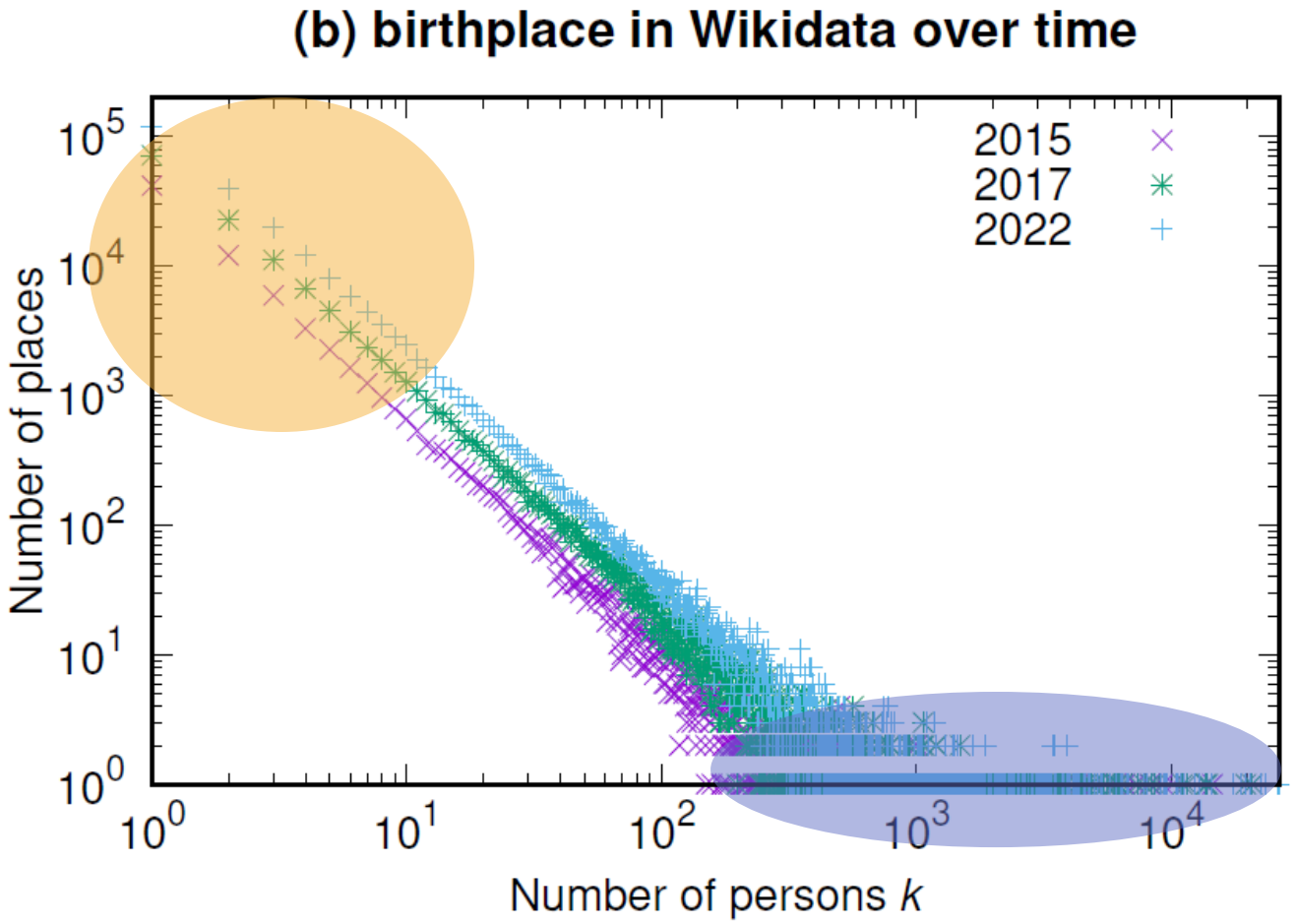
Crowd-sourced Knowledge graphs

- Created through the collective efforts of a diverse group of individuals
- The graph that emerges is a multidisciplinary knowledge graph
- Examples of crowdsourced KGs: **DBpedia**, **YAGO4**, **Wikidata**

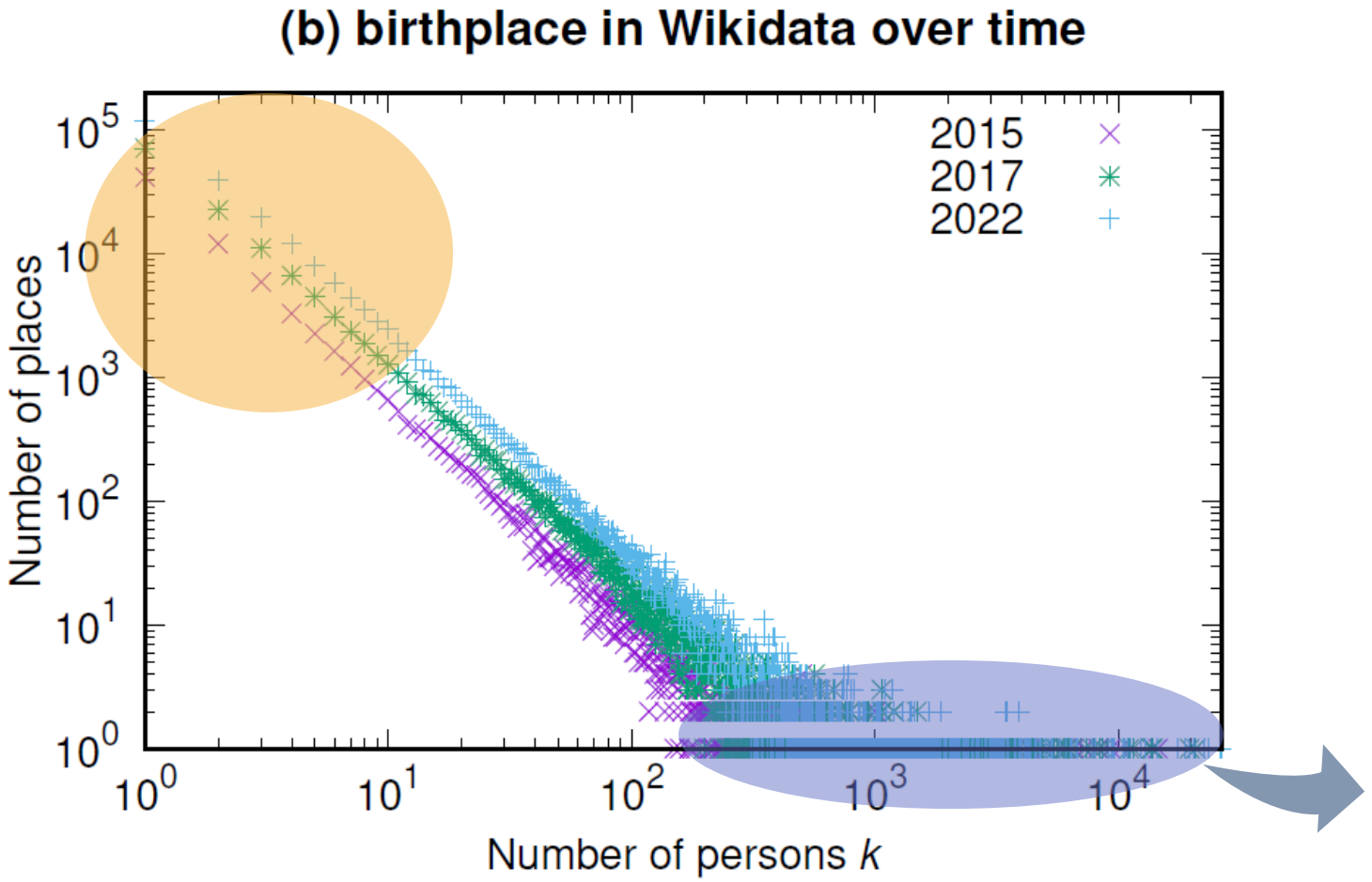


Does a structure emerge from crowdsourced knowledge graphs?

Example of the emergence of a regular structure



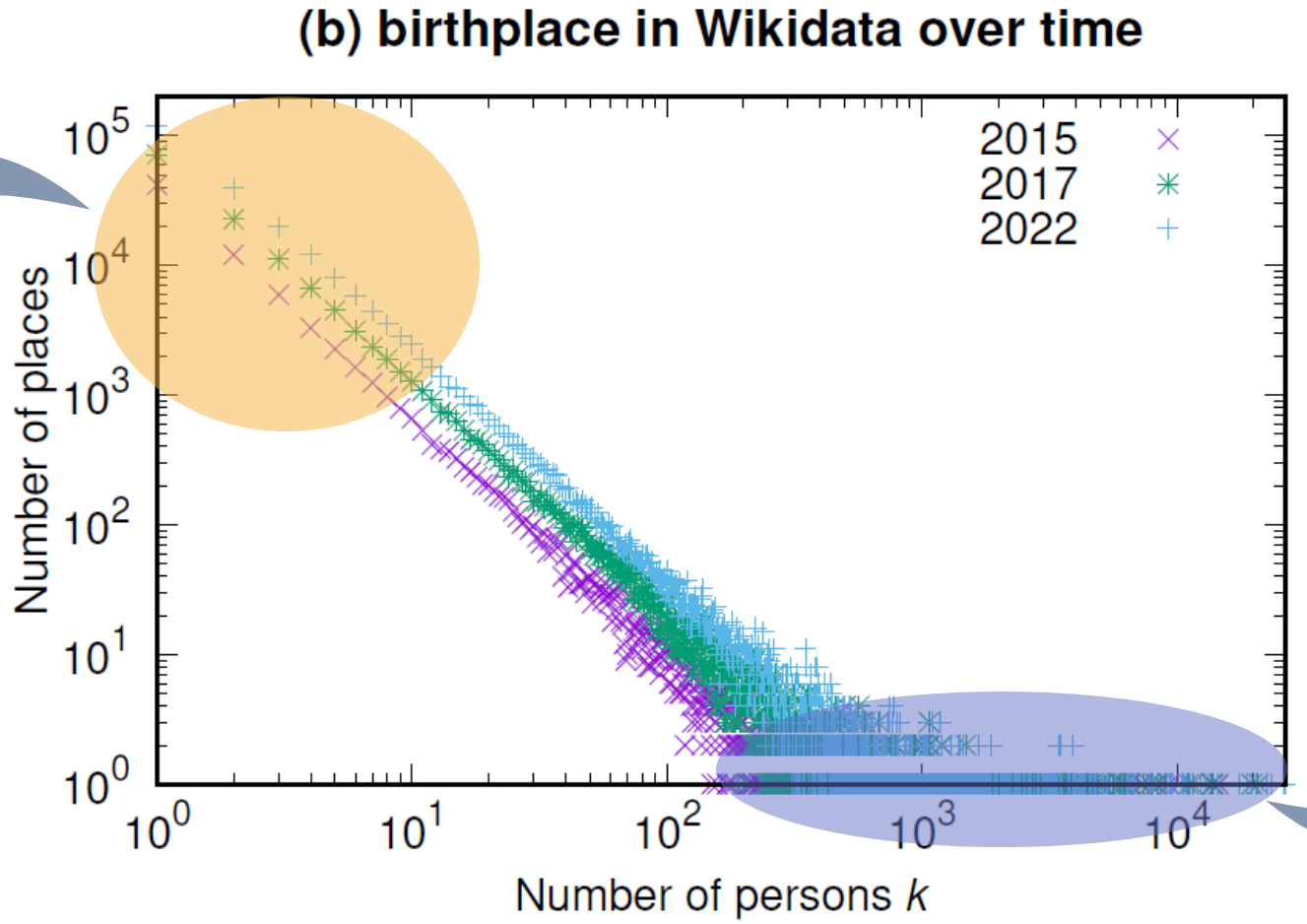
Example of the emergence of a regular structure



Paris, London, New-York…

Example of the emergence of a regular structure

Parthenay,
Lucens, Saida,…



Paris, London,
New-York…

Positioning

Knowledge representation

Modeling of concepts

- Description logics and ontologies
- Repercussions on the form of the facts

Modeling of fact organization

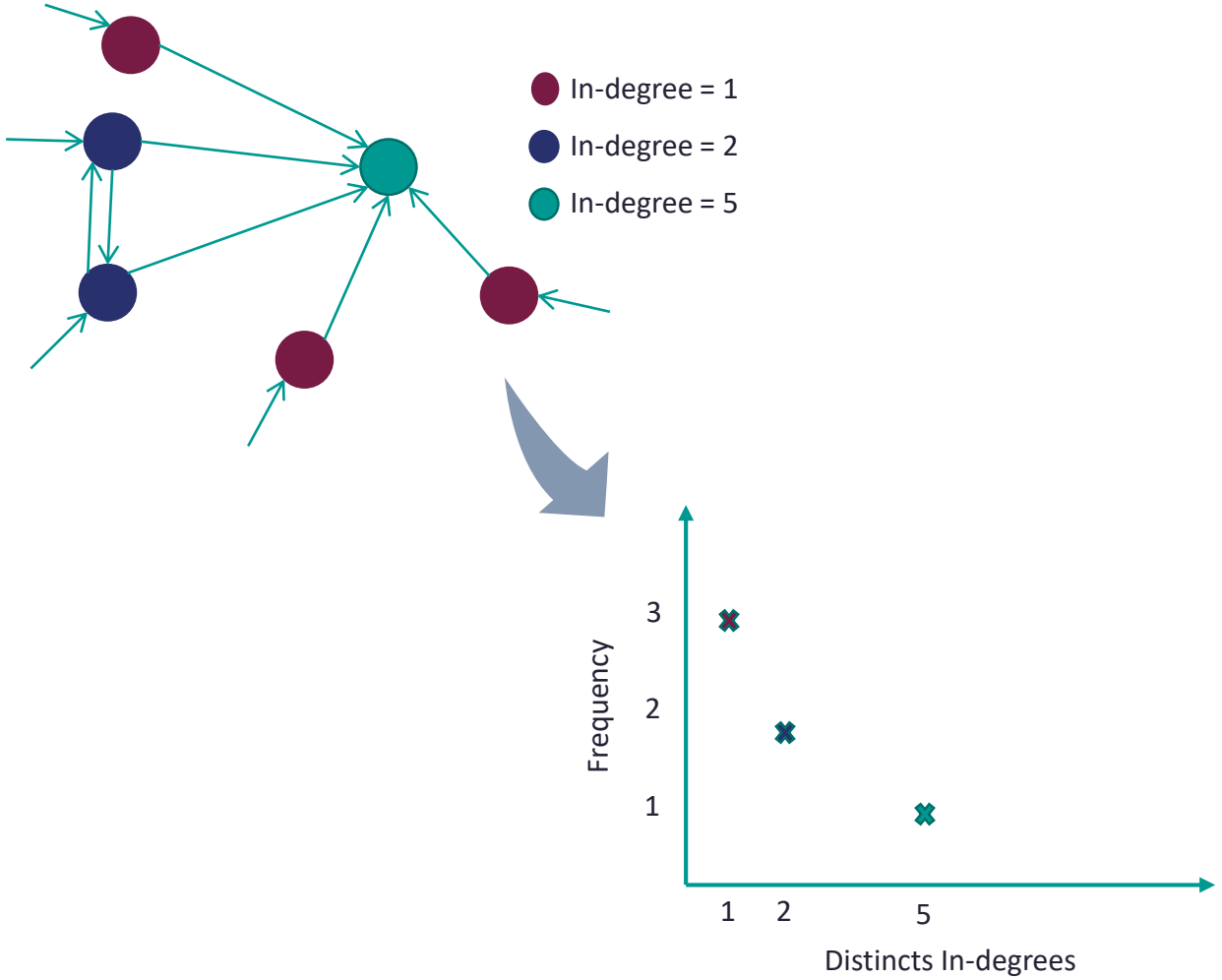
- Description of the structuring of entities and relationships
- Study of the way in which facts are organized

Developing models that generate synthetic data that closely resembles real graphs

Degree distribution

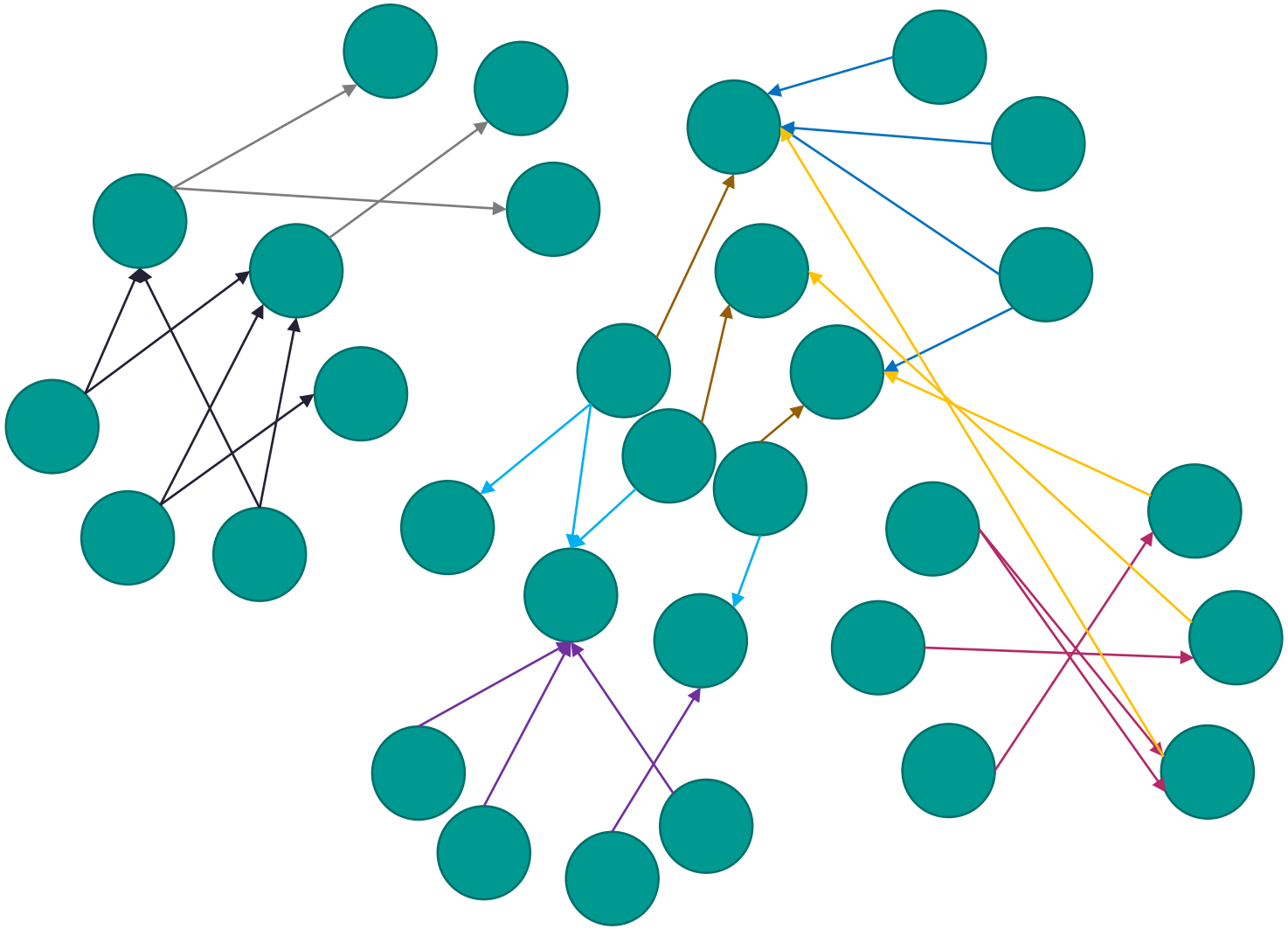
❑ Why we use it?

- Indicating the distribution of facts within the graph
- Highlighting entities that concentrate the majority of knowledge



Knowledge Graphs

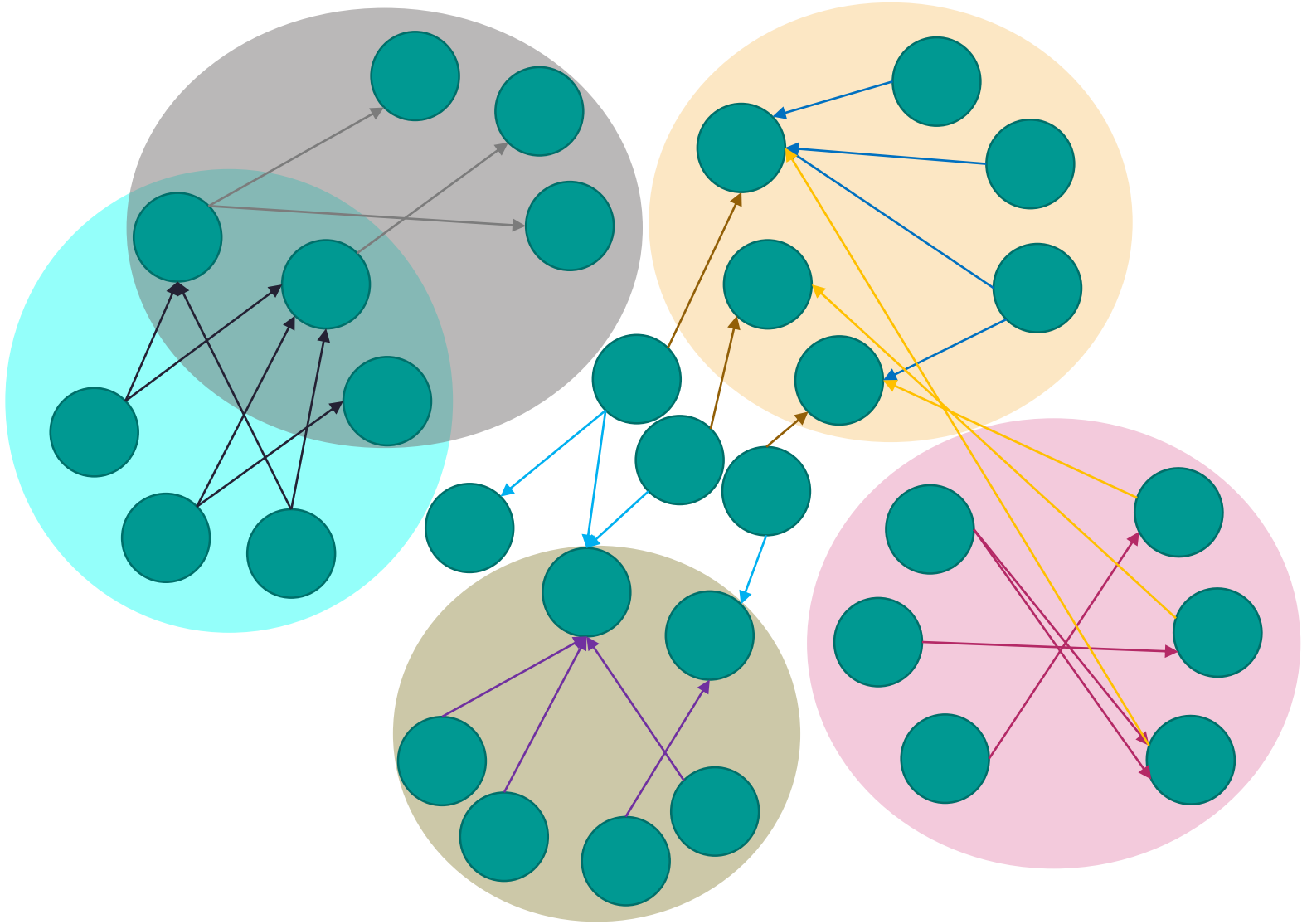
- Ingredients of knowledge graphs :
  - Entities
  - Relationships



Knowledge Graphs

- Ingredients of knowledge graphs :
  - Entities
  - Relationships

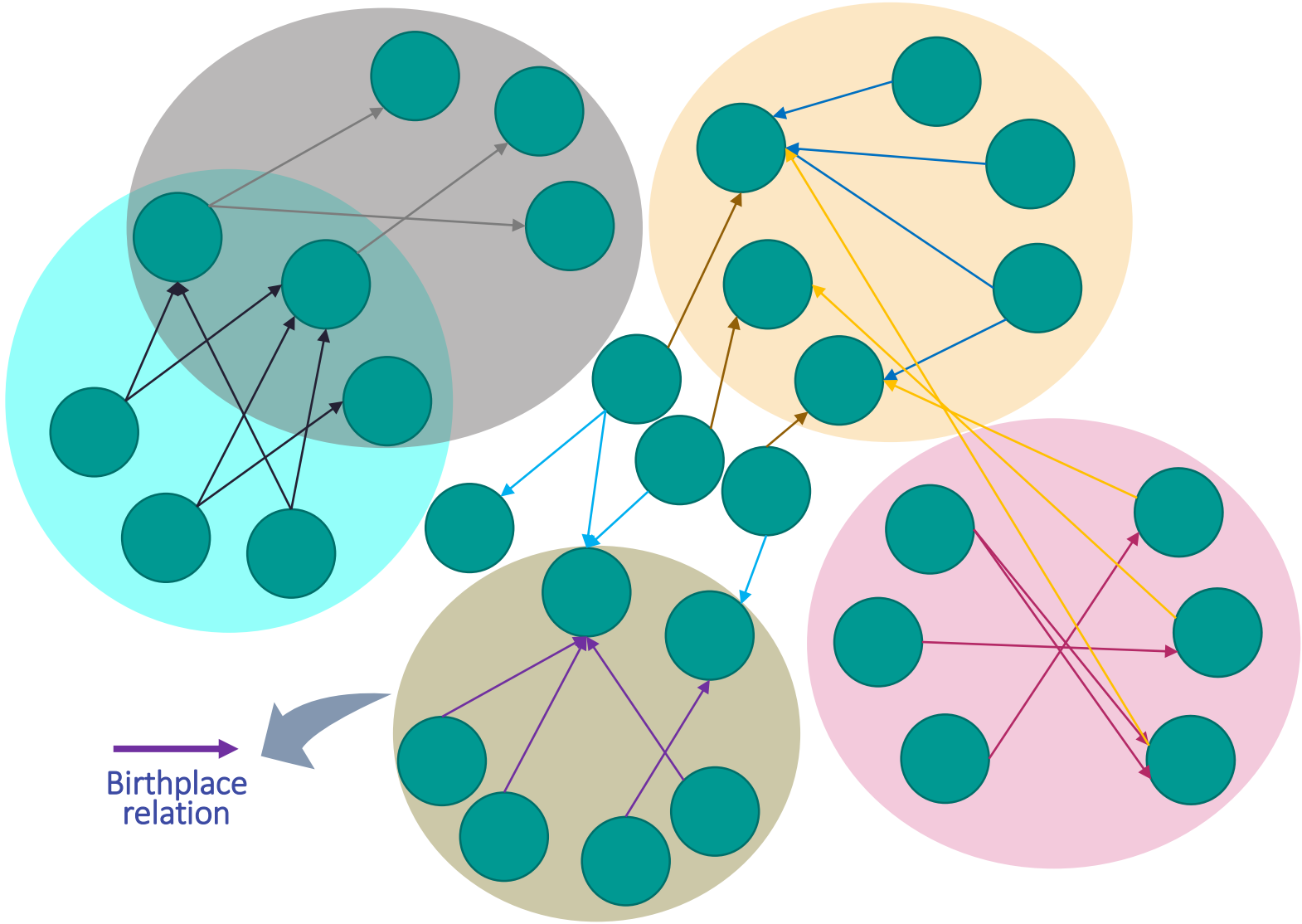
- We can see KGs as a set of bipartite graphs, where each one contains:
  - Only one type of relationship
  - Two sets of entities: an objects set, and a subjects set



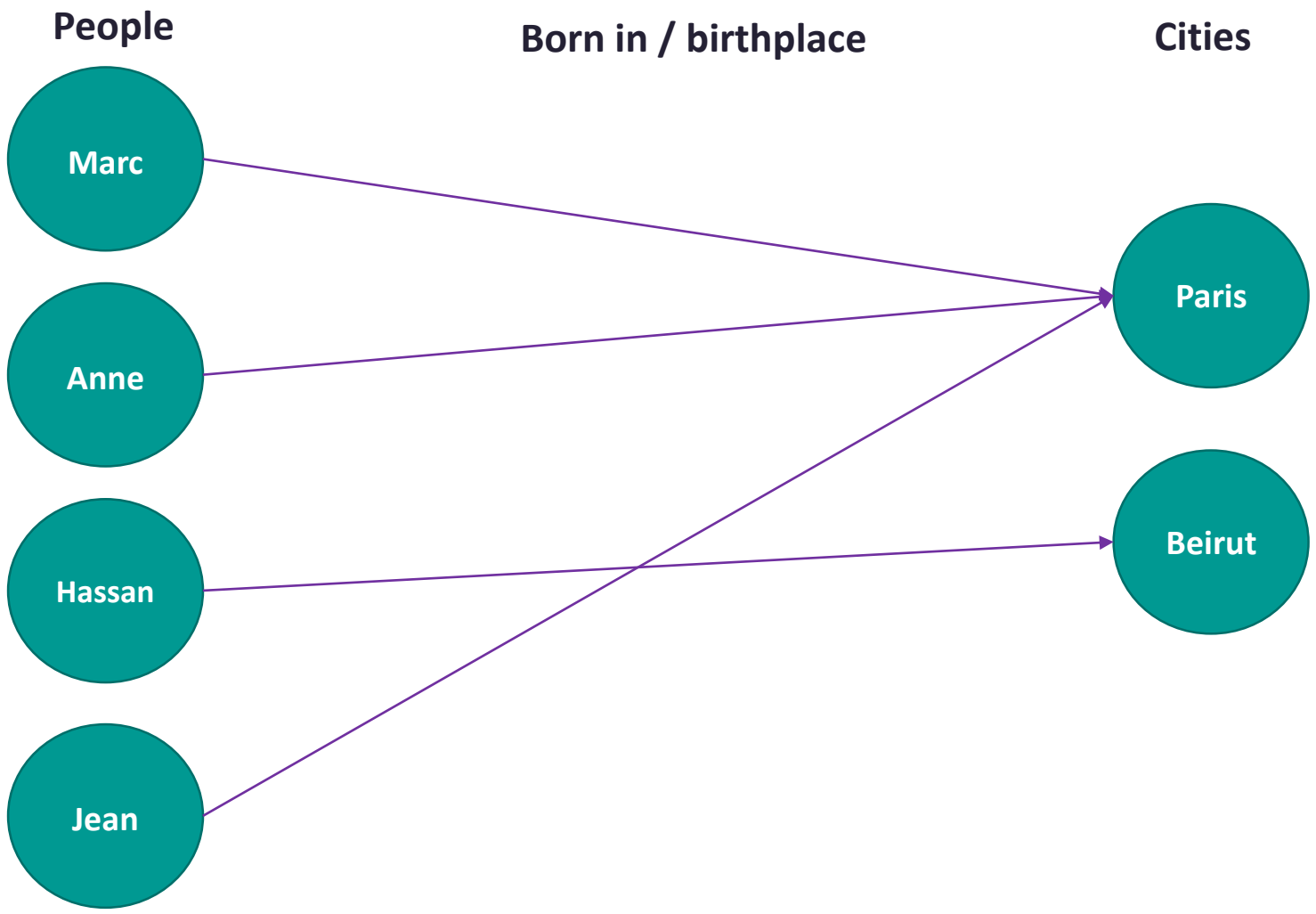
Knowledge Graphs

- Ingredients of knowledge graphs :
  - Entities
  - Relationships

- We can see KGs as a set of bipartite graphs, where each one contains:
  - Only one type of relationship
  - Two sets of entities: an objects set, and a subjects set

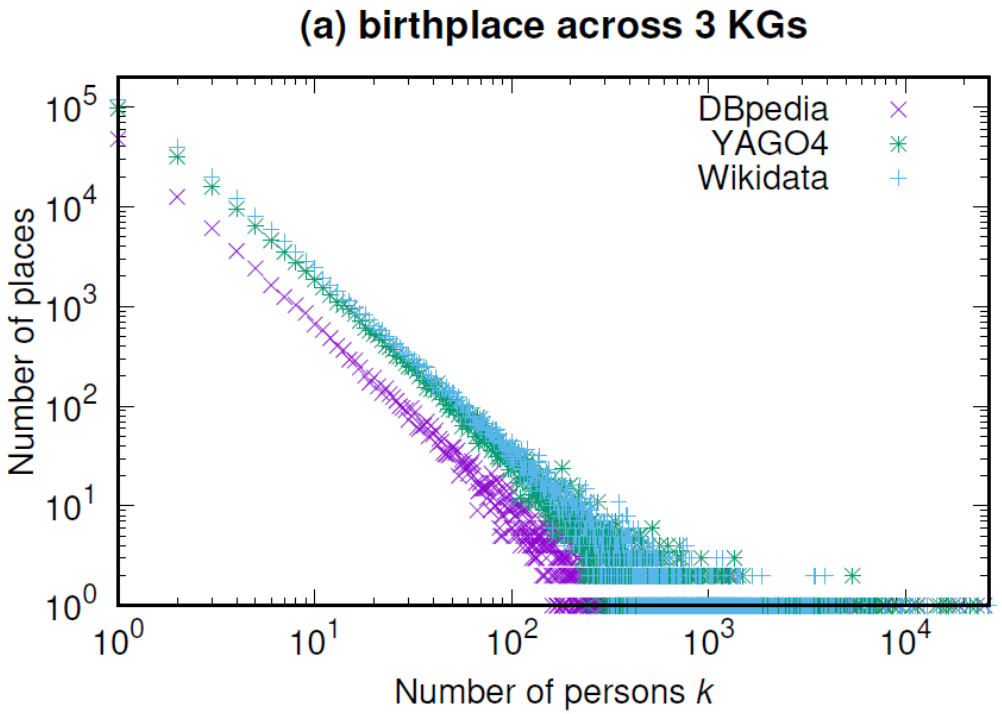


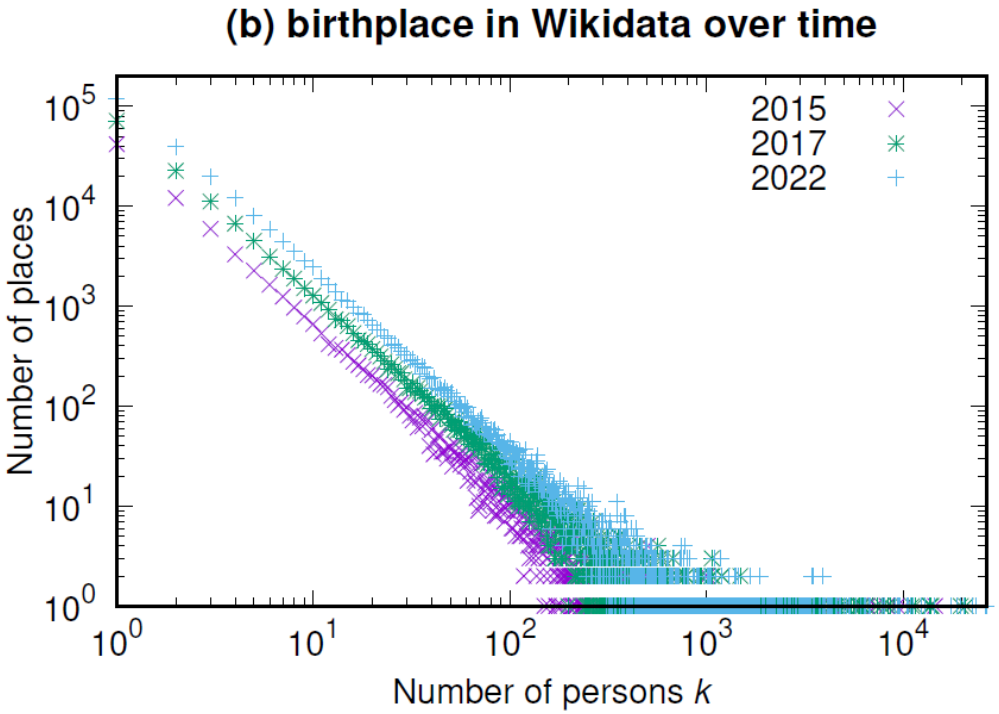
Example: Birthplace relation Graph



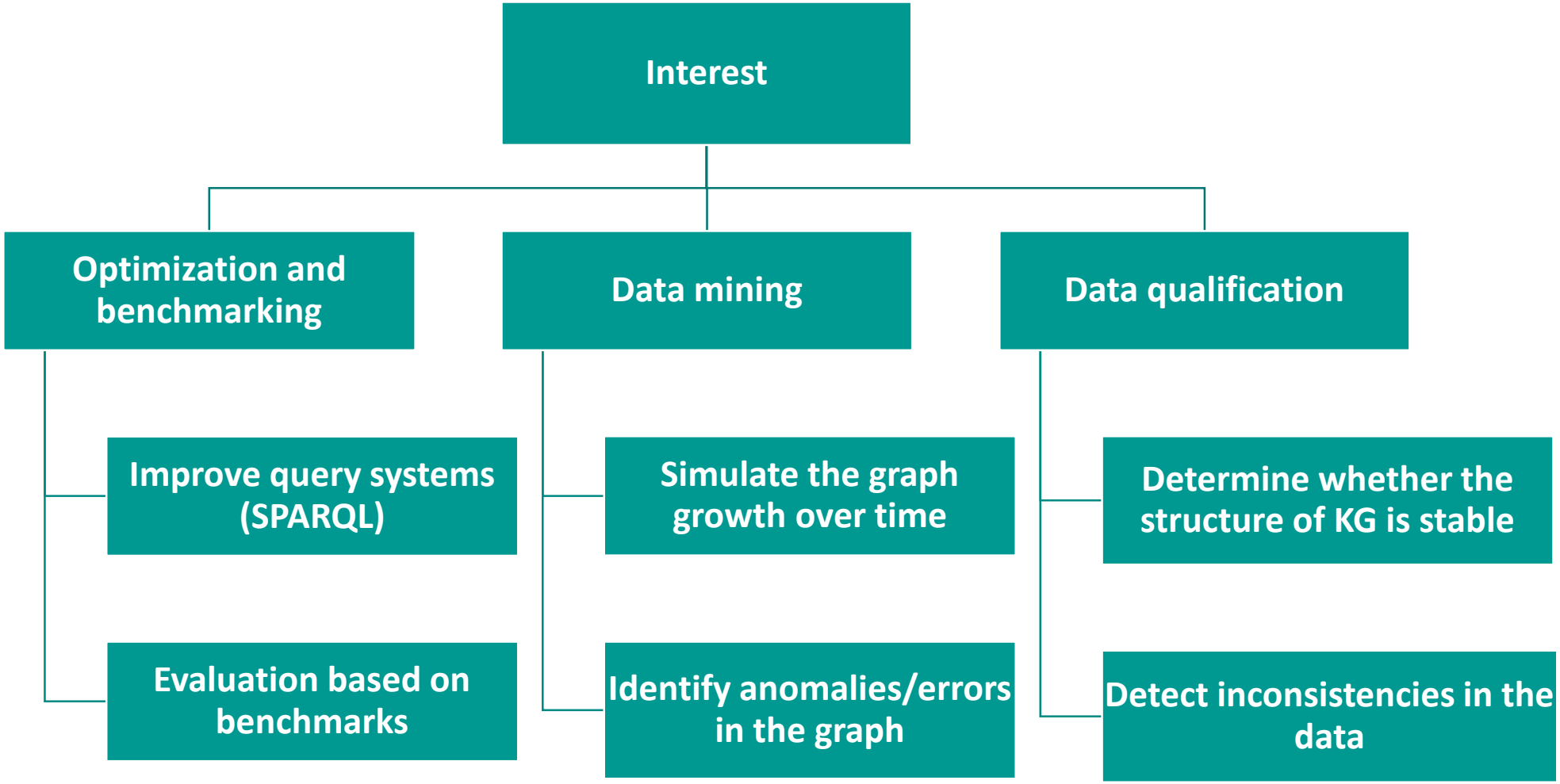
Birthplace relation distributions

The degree distribution follows a power law with almost a similar skewness for data gathered from three different data sources and over time.

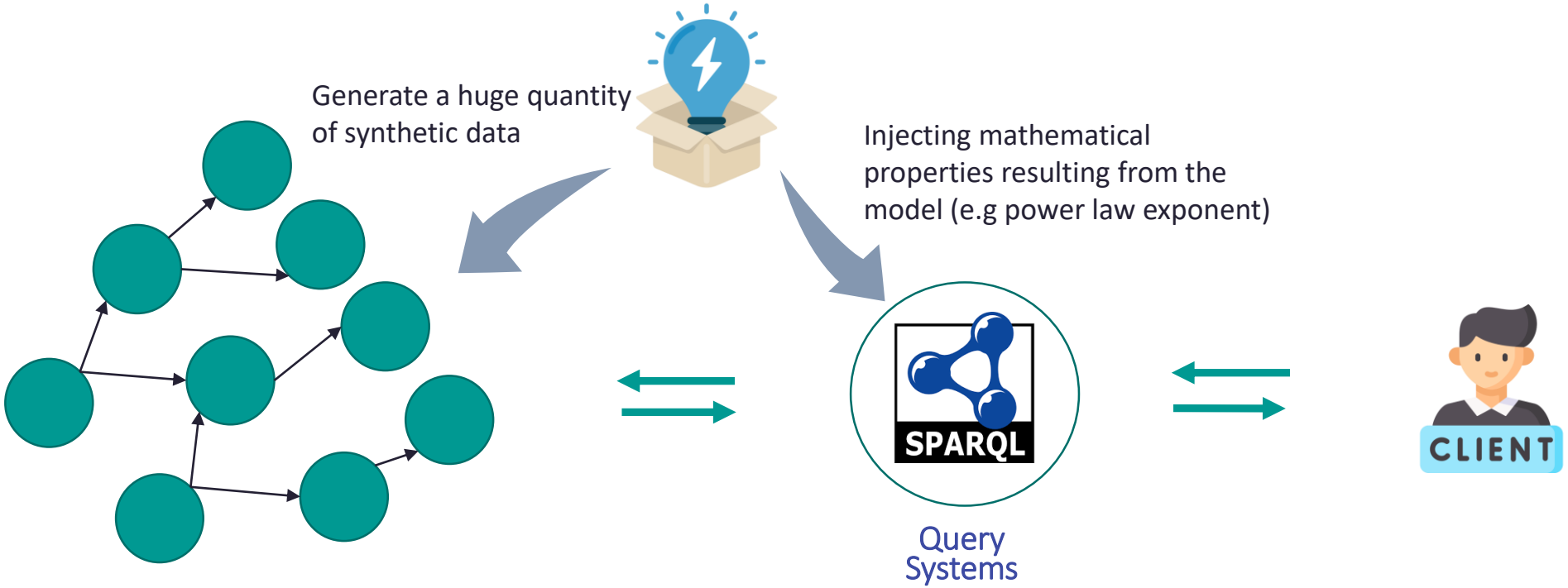




Why we build generative models?



Optimization and benchmarking

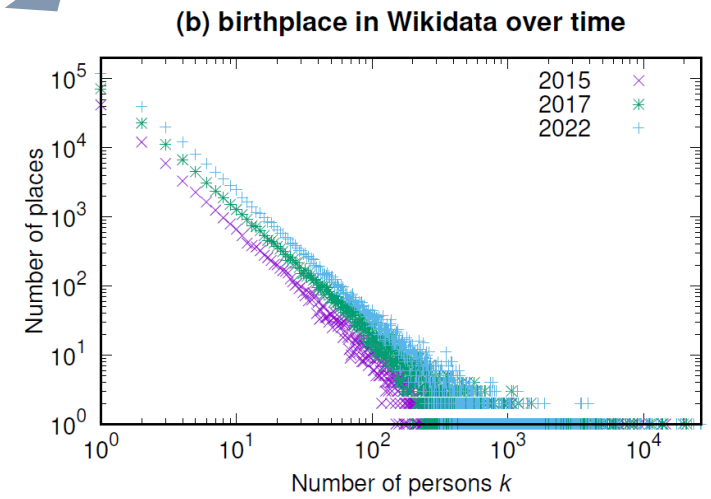


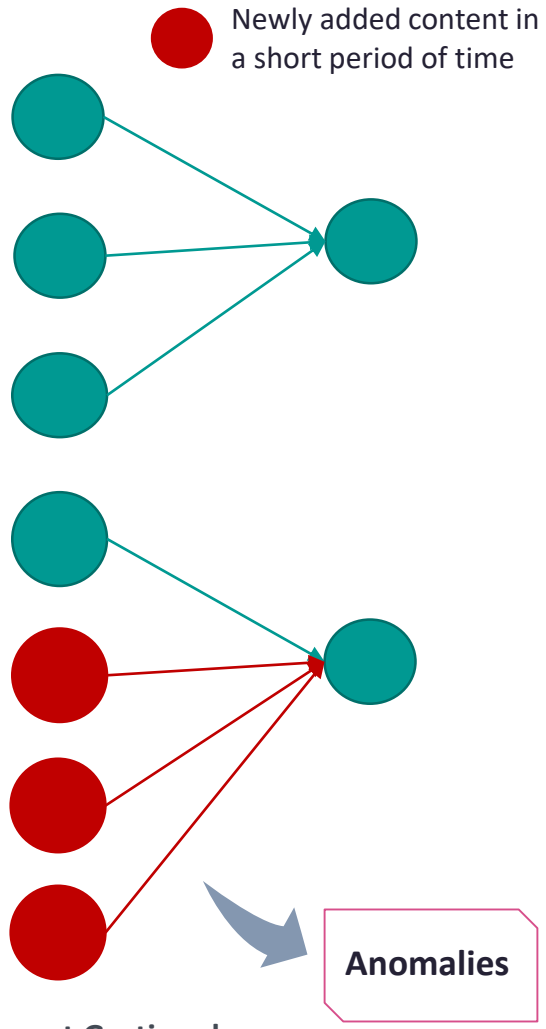
Theoharis, Y., Georgakopoulos, G., & Christophides, V. (2012). PoweRGen: A power-law based generator of RDFS schemas. Information Systems, 37(4), 306–319.

Data mining

Creator relationship		
Rank	Painter	#Paintings
1	**Nandalal Bose**	**3511**
2	Kalervo Palsa	1940
3	Edvard Munch	1783
4	Peter Paul Rubens	1692
5	Anthony van Dyck	1557
6	Jef Diederen	1208
7	Claude Monet	1125
8	Pierre-Auguste Renoir	1117
9	John Everett	1062
10	Salvador Dalí	1057

- Finding appropriate ranking indicators [4]: e.g **Creator** is an appropriate relation to classify painters.
- Identify anomalies in the graph.
- Simulate graph growth overtime

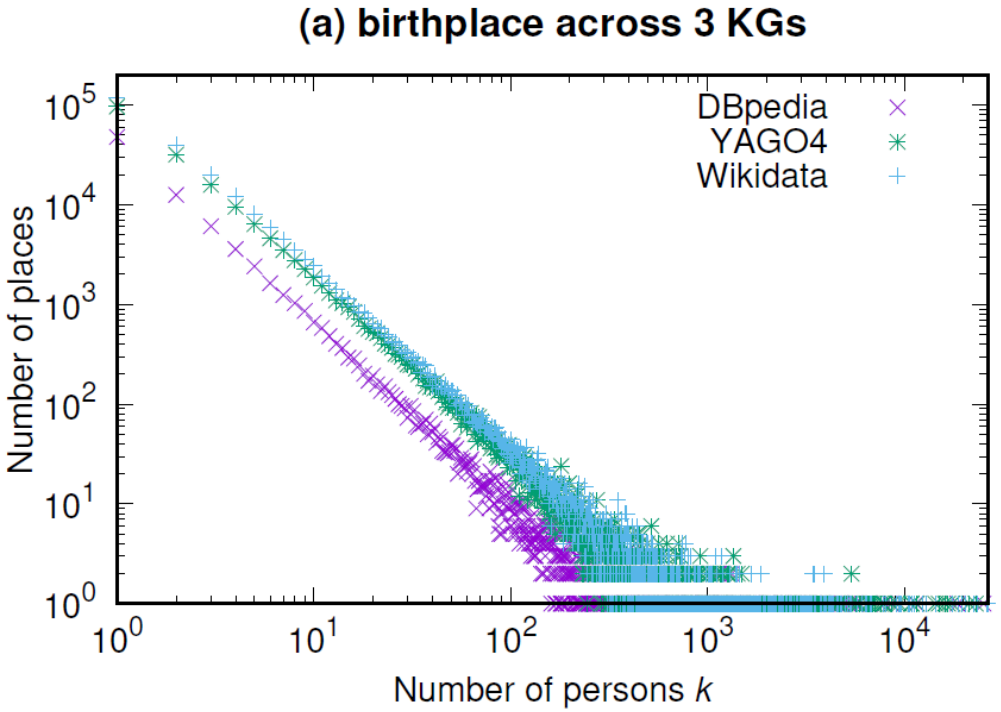




De Runz, C., Giacometti, A., Markhoff, B. and Soulet, A., 2021. Découverte d'indicateurs de classement dans le Web des données. Extraction et Gestion des Connaissances: Actes EGC'2021

Data qualification

- Determine whether the structure is stable or not
- Knowledge completeness
- Inconsistencies detection in the data and improving its quality



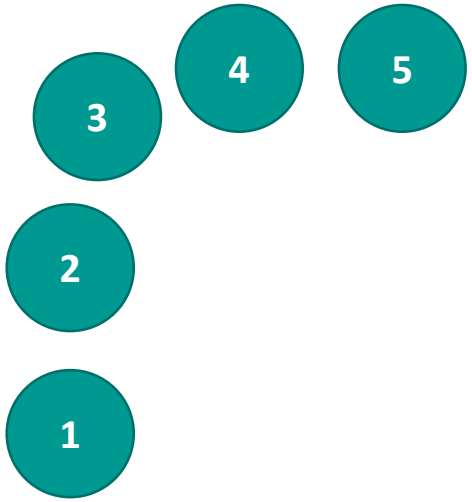
Suhas Shrinivasan and Simon Razniewski. How stable is knowledge base knowledge ? arXiv preprint arXiv :2211.00989, 2022.

Challenges

How can network science be used to model such different semantic relationships?

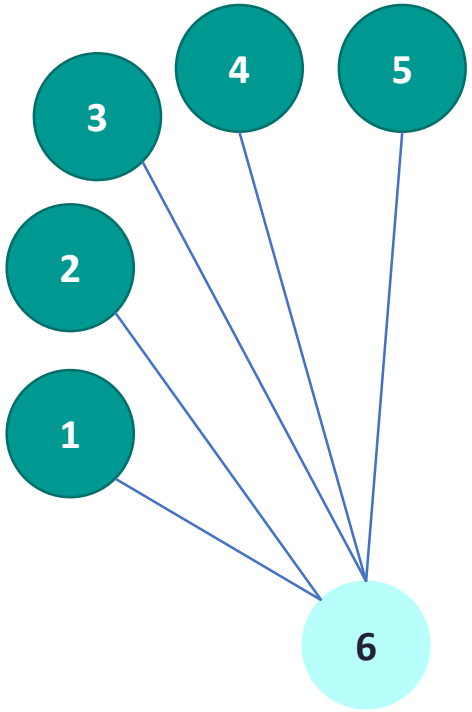
Data veracity	Data volume	Data variety
Incompleteness and errors in knowledge	High data volume like Wikidata	KGs contain multidisciplinary knowledge, thus we need an transdisciplinary model
Dynamic nature of the world represented	Experimental, algorithmic and memory management challenges	The diversity of entities and relationships even in the same discipline

Barabási–Albert model



Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512

Barabási–Albert model



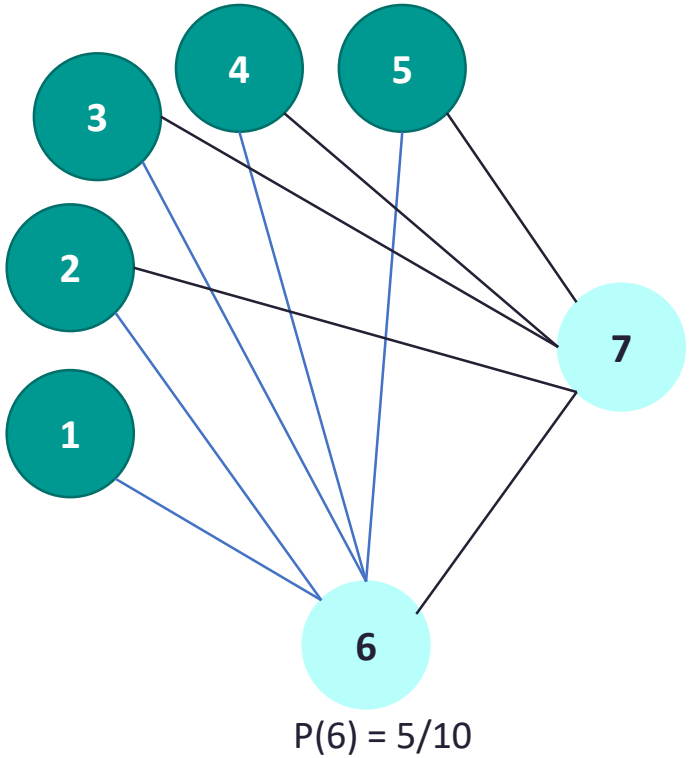
P(6) = 5/10

Continuous growth

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512

Barabási–Albert model

P(1) = P(2) = P(3) = P(4) = P(5) = 1/10



P(6) = 5/10

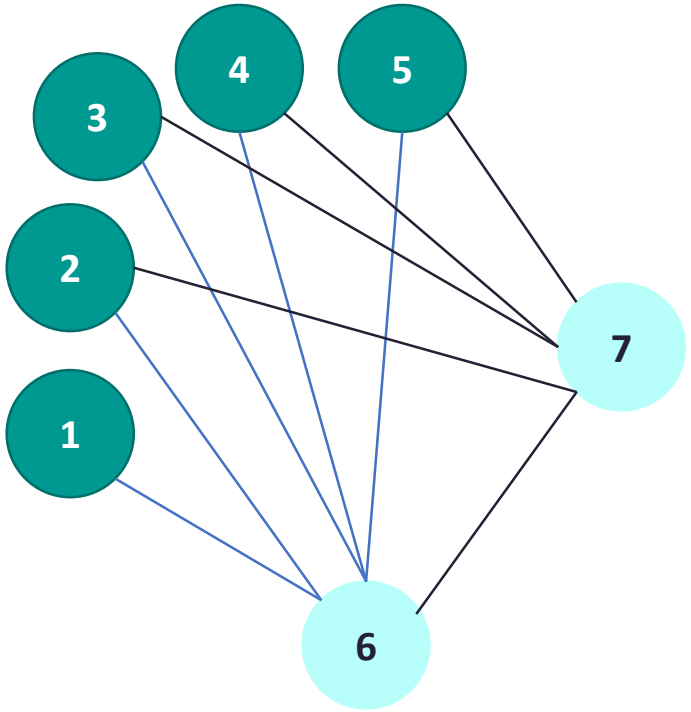
Continuous growth

Preferential attachment

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512

Barabási–Albert model

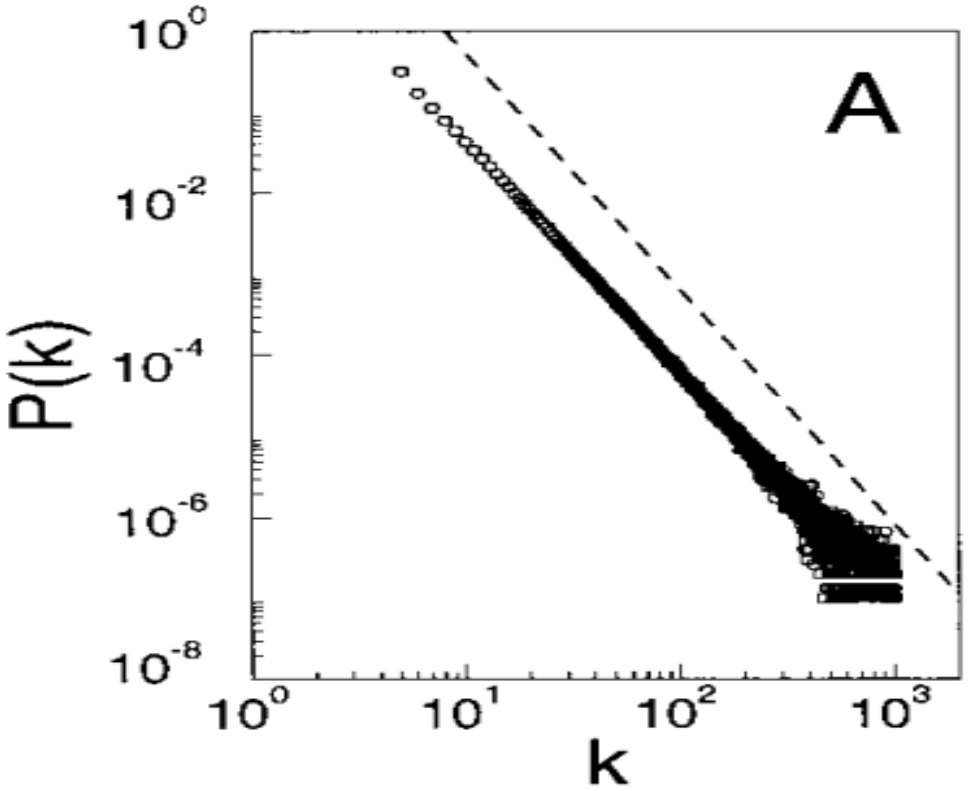
P(1) = P(2) = P(3) = P(4) = P(5) = 1/10



P(6) = 5/10

Continuous growth

Preferential attachment

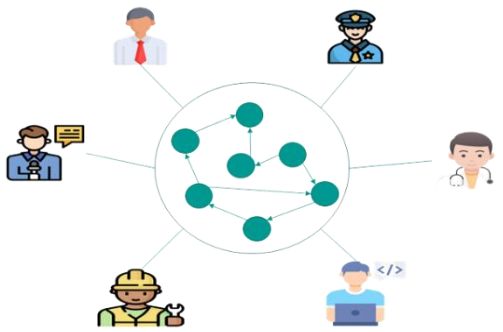


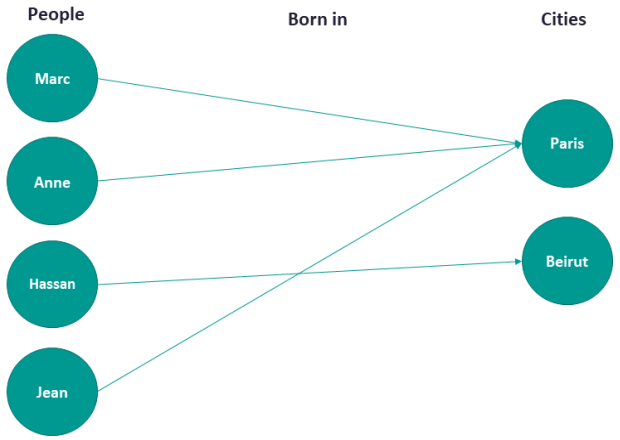
Connectivity distribution at t=150,000 (○) and t=200,000 (□). using m_0=5 m=5. The law exponent=2.9.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512

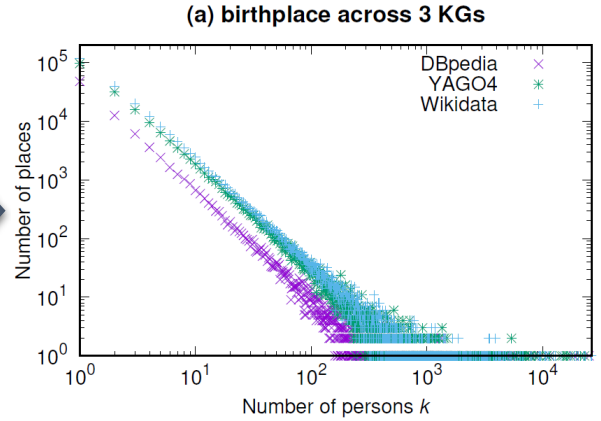
Conclusion

Does a structure emerge from a crowdsourced knowledge graph?





How to model the structure of a knowledge graph?



Benefits

- Optimization and benchmarking
- Data mining
- Data qualification

Challenges

- Data Veracity: Incompleteness, dynamic changes
- Data volume: huge KGs
- Data Variety: multidisciplinary knowledge

References

- [1] Ibrahim Abdelaziz, Razen Harbi, Zuhair Khayyat, and Panos Kalnis. A survey and experimental comparison of distributed sparql engines for very large rdf data. VLDB, 10(13) :2049–2060, 2017.
- [2] Waqas Ali, Muhammad Saleem, Bin Yao, Aidan Hogan, and Axel-Cyrille Ngonga Ngomo. A survey of rdf stores & sparql engines for querying knowledge graphs. The VLDB Journal, pages 1–26, 2022.
- [3] Theoharis, Y., Georgakopoulos, G., & Christophides, V. (2012). PoweRGen: A power-law based generator of RDFS schemas. Information Systems, 37(4), 306–319.
- [4] De Runz, C., Giacometti, A., Markhoff, B. and Soulet, A., 2021. Découverte d'indicateurs de classement dans le Web des données. Extraction et Gestion des Connaissances: Actes EGC'2021.
- [5] Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512.
- [6] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In WWW, pages 211–220, 2007.
- [7] Suhas Shrinivasan and Simon Razniewski. How stable is knowledge base knowledge ? arXiv preprint arXiv :2211.00989, 2022.

THANK YOU