Discovering Ranking Scores in the Web of Data

Cyril De Runz, Arnaud Giacometti, Béatrice Markhoff, Arnaud Soulet

Université de Tours, LIFAT, Blois

université de TOURS

LIFAT EA 63 00

What are the most important painters?

theartwolf

Rank	Painter
1	**Pablo Picasso**
2	Giotto Di Bondone
3	Leonardo da Vinci
4	Paul Cézanne
5	Rembrandt
6	Diego Velasquez
7	Wassily Kandinsky
8	Claude Monet
9	Caravaggio
10	Joseph M. W. Turner

Ranker

Rank	Painter
1	Leonardo da Vinci
2	Vincent van Gogh
3	Michelangelo
4	Rembrandt
5	**Pablo Picasso**
6	Claude Monet
7	Caravaggio
8	Johannes Vermeer
9	Raphael
10	Salvador Dalí

ART CYCLOPEDIA

Rank	Painter
1	**Pablo Picasso**
2	Vincent van Gogh
3	Leonardo da Vinci
4	Claude Monet
5	Salvador Dalí
6	Henri Matisse
7	Rembrandt
8	Andy Warhol
9	Georgia O'Keeffe
10	Michelangelo

Many answers…

What are the most important painters?

theartwolf

Rank	Painter
1	**Pablo Picasso**
2	Giotto Di Bondone
3	Leonardo da Vinci
4	Paul Cézanne
5	Rembrandt
6	Diego Velasquez
7	Wassily Kandinsky
8	Claude Monet
9	Caravaggio
10	Joseph M. W. Turner

Ranker

Rank	Painter
1	Leonardo da Vinci
2	Vincent van Gogh
3	Michelangelo
4	Rembrandt
5	**Pablo Picasso**
6	Claude Monet
7	Caravaggio
8	Johannes Vermeer
9	Raphael
10	Salvador Dalí

ART CYCLOPEDIA

Rank	Painter
1	**Pablo Picasso**
2	Vincent van Gogh
3	Leonardo da Vinci
4	Claude Monet
5	Salvador Dalí
6	Henri Matisse
7	Rembrandt
8	Andy Warhol
9	Georgia O'Keeffe
10	Michelangelo

Problem: Is it possible to discover ranking scores for painters?

What are the most important painters?

theartwolf

Rank	Painter
1	**Pablo Picasso**
2	Giotto Di Bondone
3	Leonardo da Vinci
4	Paul Cézanne
5	Rembrandt
6	Diego Velasquez
7	Wassily Kandinsky
8	Claude Monet
9	Caravaggio
10	Joseph M. W. Turner

Ranker

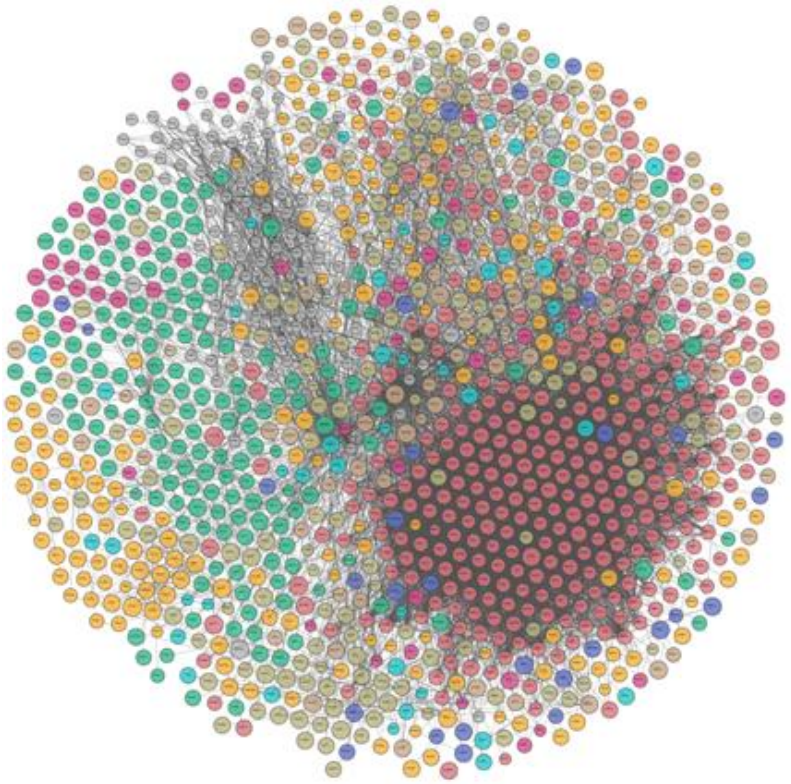
Rank	Painter
1	Leonardo da Vinci
2	Vincent van Gogh
3	Michelangelo
4	Rembrandt
5	**Pablo Picasso**
6	Claude Monet
7	Caravaggio
8	Johannes Vermeer
9	Raphael
10	Salvador Dalí

ART CYCLOPEDIA

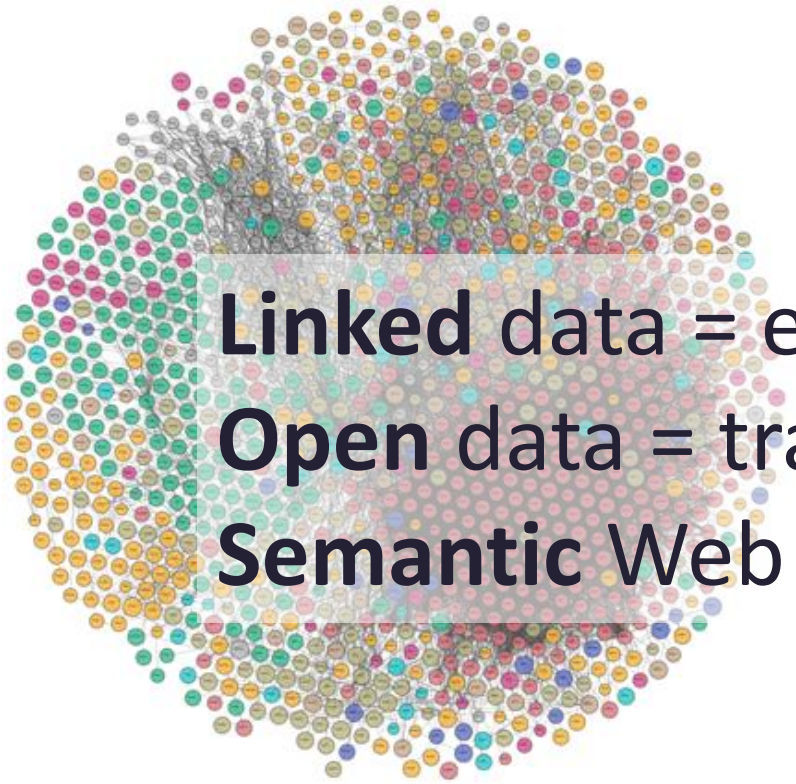
Rank	Painter
1	**Pablo Picasso**
2	Vincent van Gogh
3	Leonardo da Vinci
4	Claude Monet
5	Salvador Dalí
6	Henri Matisse
7	Rembrandt
8	Andy Warhol
9	Georgia O'Keeffe
10	Michelangelo

Problem: Is it possible to discover ranking scores for ~~painters~~ everything?

Web of data = digital twin of the world

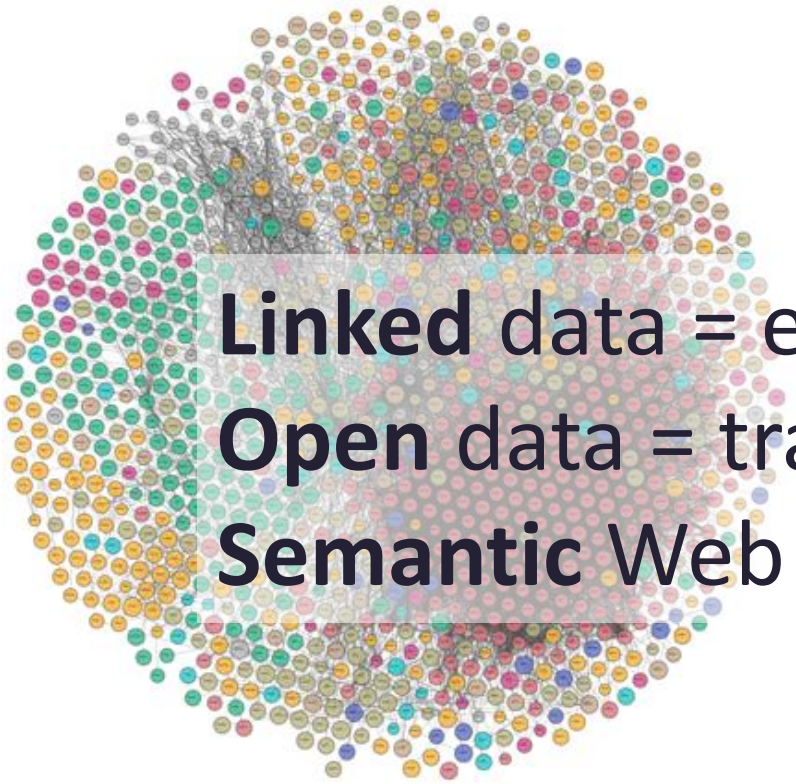


Web of data = digital twin of the world



Linked data = easy to process
Open data = transparency and reproducibility
Semantic Web = reasonnable

Web of data = digital twin of the world



Linked data = easy to process
Open data = transparency and reproducibility
Semantic Web = reasonnable

Challenge: How to deal with the heterogeneity of the Web of Data?

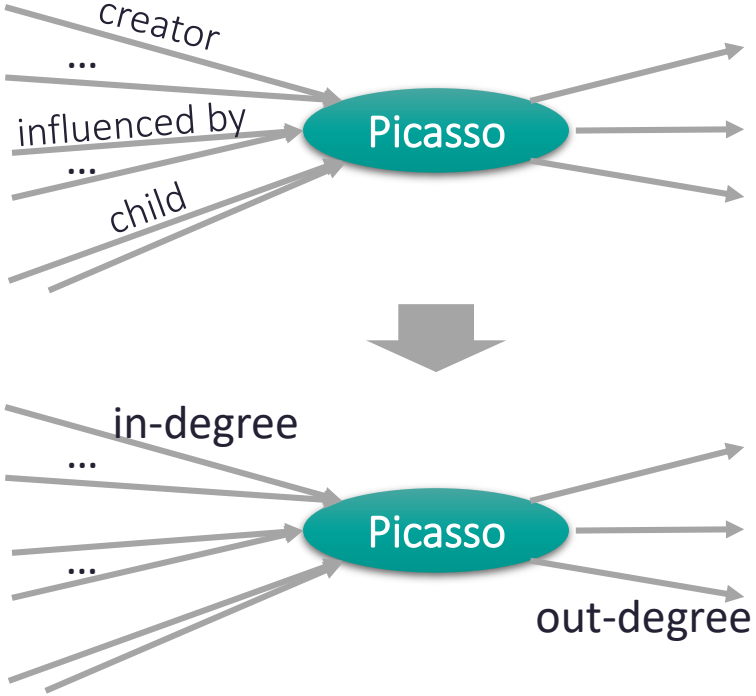
Entity Ranking in literature

- **Webometrics** *[Thelwall et al, 05]*

= infometrics applied to the Web

- **Entity ranking for searching** *[Hogan et al, 11;Delbru et al, 12]*

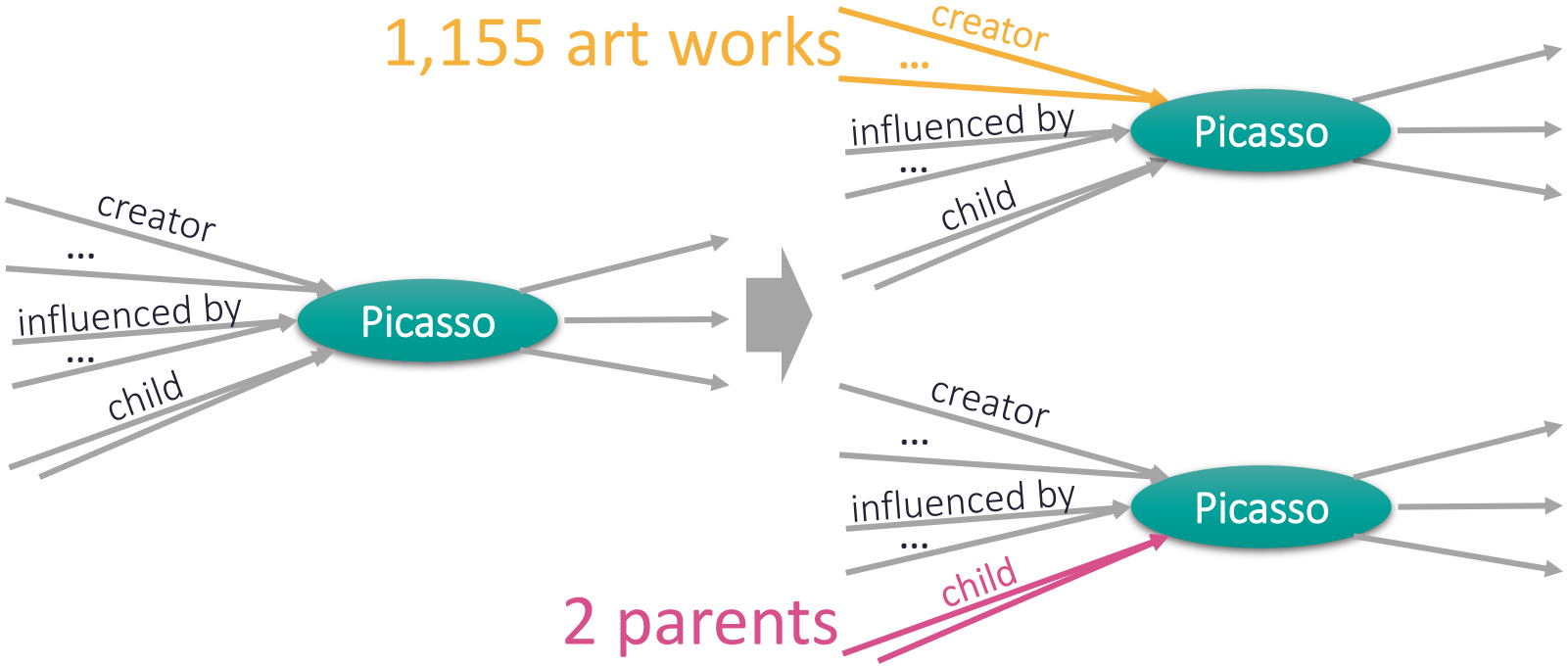
= information retrieval applied to the Web of Data



Meaningless ranking scores in the real-world + Representation dependent

Semantic ranking score

= ranking score based on counting the number of links for a property



The number of paintings is relevant score (`creator`), but not the number of parents (`child`)

What is a relevant counting property for ranking?

Ranking score = Power Law ?

An index to quantify an individual's scientific research output

J. E. Hirsch*

Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319

Communicated by Manuel Cardona, Max Planck Institute for Solid State Research, Stuttgart, Germany, September August 15, 2005)

I propose the index *h*, defined as the number of papers with citation number ≥*h*, as a useful index to characterize the scientific output of a researcher.

citations | impact | unbiased

For the few scientists who earn a Nobel prize, the impact and relevance of their research is unquestionable. Among the rest of us, how does one quantify the cumulative impact and relevance of an individual's scientific research output? In a world of limited resources, such quantification (even if potentially distasteful) is often needed for evaluation and comparison purposes (e.g., for university faculty recruitment and advancement, award of grants, etc.).

The publication record of an individual and the citation record clearly are data that contain useful information. That information includes the number (N_p) of papers published over n years, the number of citations (N_c) for each paper (j), the journals where the papers were published, their impact parameter, etc. This large amount of information will be evaluated with different criteria by different people. Here, I would like to propose a single number, the "h index," as a particularly simple and useful way to characterize the scientific output of a researcher.

A scientist has index h if h of his or her N_p papers have at least h citations each and the other ($N_p - h$) papers have ≤h citations each.

The research reported here concentrated on physicists; however, I suggest that the h index should be useful for other scientific disciplines as well. (At the end of the paper I discuss some observations for the h index in biological sciences.) The highest h among physicists appears to be E. Witten's h, which is 110. That is, Witten has written 110 papers with at least 110 citations each. That gives a lower bound on the total number of citations to Witten's papers at $h^2 = 12{,}100$. Of course, the total number of citations ($N_{c,tot}$) will usually be much larger than h^2, because h^2 both underestimates the total number of citations of the h most-cited papers and ignores the papers with <h citations. The relation between $N_{c,tot}$ and h will depend on the detailed form of the particular distribution (1), and it is useful to define the proportionality constant a as

$$N_{c,tot} = ah^2. \quad [1]$$

I find empirically that a ranges between 3 and 5.

Other prominent physicists with high hs are A. J. Heeger (h = 107), M. L. Cohen (h = 94), A. C. Gossard (h = 94), P. W. Anderson (h = 91), S. Weinberg (h = 88), M. E. Fisher (h = 88), M. Cardona (h = 86), P. G. deGennes (h = 79), J. N. Bahcall (h = 77), Z. Fisk (h = 75), D. J. Scalapino (h = 75), G. Parisi (h = 73), S. G. Louie (h = 70), R. Jackiw (h = 69), F. Wilczek (h = 68), C. Vafa (h = 66), M. B. Maple (h = 66), D. J. Gross (h = 66), M. S. Dresselhaus (h = 62), and S. W. Hawking (h = 62). I argue that h is preferable to other single-number criteria commonly used to evaluate scientific output of a researcher, as follows:

(*i*) Total number of papers, ductivity. Disadvantage impact of papers.

(*ii*) Total number of citations, total impact. Disadvantage by a small number of sentative of the individual many others on those Eq. 1 will imply a very disadvantage is that N_c review articles versus

(*iii*) Citations per paper (*i* allows comparison of tage: hard to find, rew high productivity.

(*iv*) Number of "significant papers with ≥*y* citations eliminates the disadvantage an idea of broad and arbitrary and will rank and *y* needs to be adjusted

(*v*) Number of citations to example, q = 5). Advantages of the ch a single number, make compare. Also, q is arbitrary disfavor individuals.

Instead, the proposed h index individual's work, avoids all listed above, usually can be by "times cited" in the The (http://isiknowledge.com), total number of citations (

Thus, I argue that two in rable in terms of their overall number of papers or their different. Conversely, comparing scientific age) with a similar citation count and very different h is likely to be the more a For a given individual, approximately linearly with assume that the researcher each published paper earns subsequent year. The total is then

$$N_{c,tot} = \sum$$

*E-mail: jhirsch@ucsd.edu.

†Of course, the database used must be by the individual's publications.

© 2005 by The National Academy of S

www.pnas.org/cgi/doi/10.1073/pnas.0507655102

PNAS | November 15,

Forbes

Filter list by: Reset Filters Oldest Youngest Women Country/Territory

Rank	Name	Net Worth
1	JEFF BEZOS	$113 B
2	BILL GATES	$98 B
3	BERNARD ARNAULT & FAMILY	$76 B
4	WARREN BUFFETT	$67.5 B
5	LARRY ELLISON	$59 B
6	AMANCIO ORTEGA	$55.1 B
7	MARK ZUCKERBERG	$54.7 B
8	JIM WALTON	$54.6 B
9	ALICE WALTON	$54.4 B

Power Law = probability distribution proportional to $x^{-\alpha}$

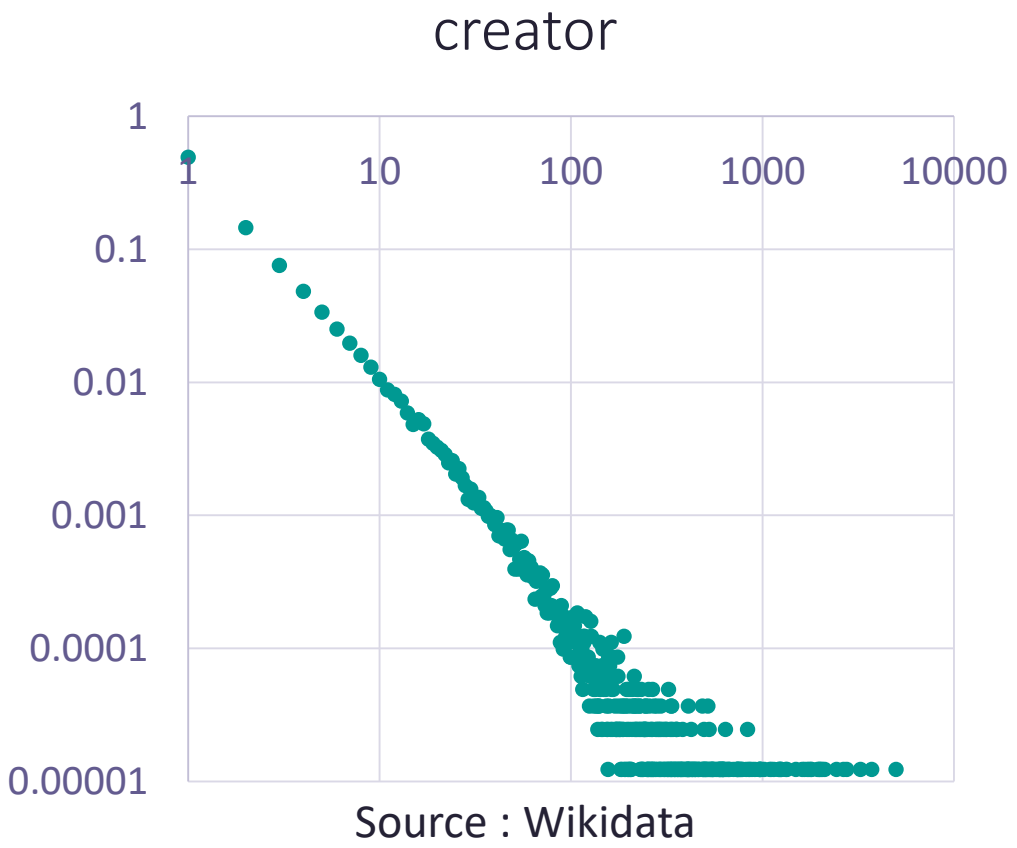
- **Infometry** *[Garfield, 1972; Egghe, 2005]*

➔ Lotka's law / preferential attachment model *[Barabasi et Albert, 1999]*

- **Econometry** *[Yule, 1924]*

➔ Yule-Simon distribution

Ranking score = Power Law ?



Source : Wikidata

Power Law = probability distribution proportional to $x^{-\alpha}$

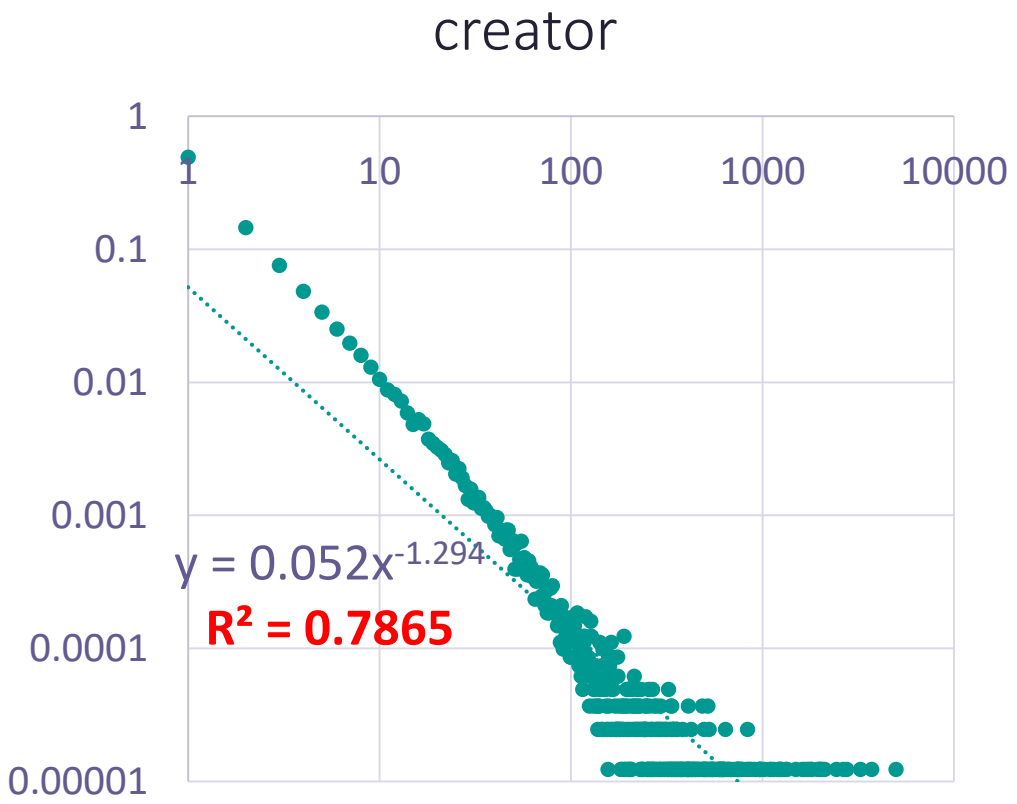
- **Infometry** *[Garfield, 1972; Egghe, 2005]*

➔ Lotka's law / preferential attachment model *[Barabasi et Albert, 1999]*

- **Econometry** *[Yule, 1924]*

➔ Yule-Simon distribution

Ranking score = Power Law ?



Power Law = probability distribution proportional to $x^{-\alpha}$

- **Infometry** *[Garfield, 1972; Egghe, 2005]*

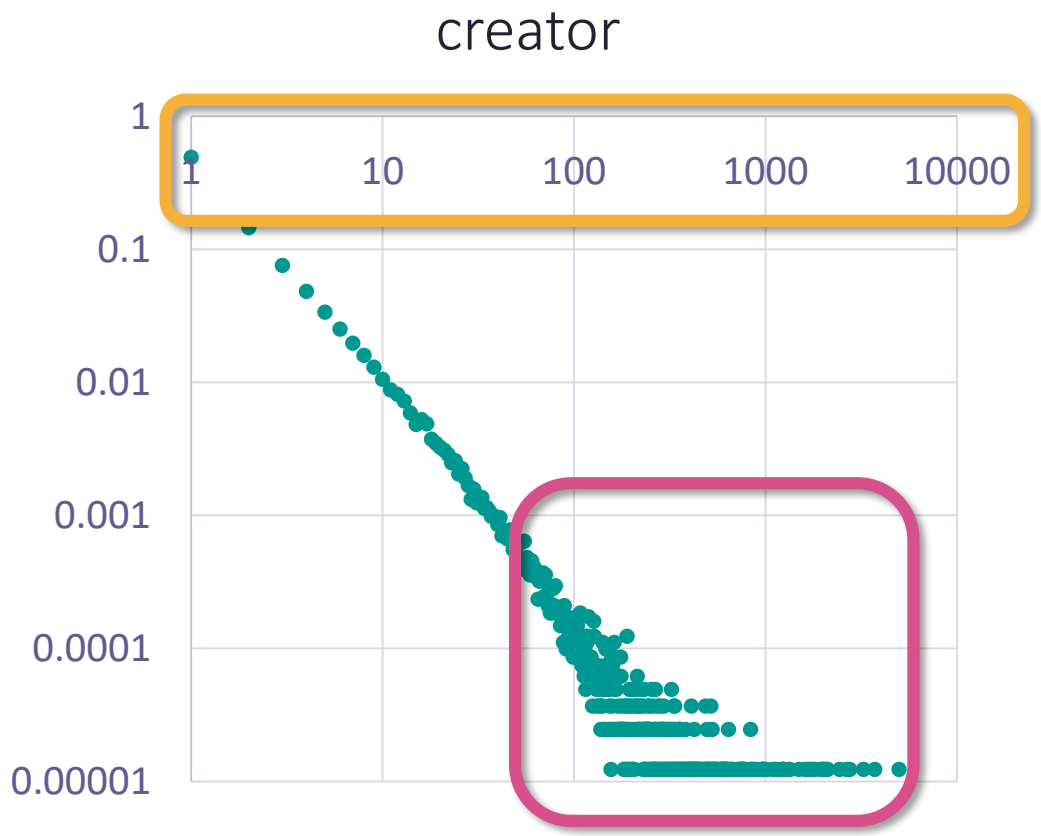
➔ Lotka's law / preferential attachment model *[Barabasi et Albert, 1999]*

- **Econometry** *[Yule, 1924]*

➔ Yule-Simon distribution

Many relevant counting properties do not conform to a power law

Ranking score = inequality



- **C1: Range**

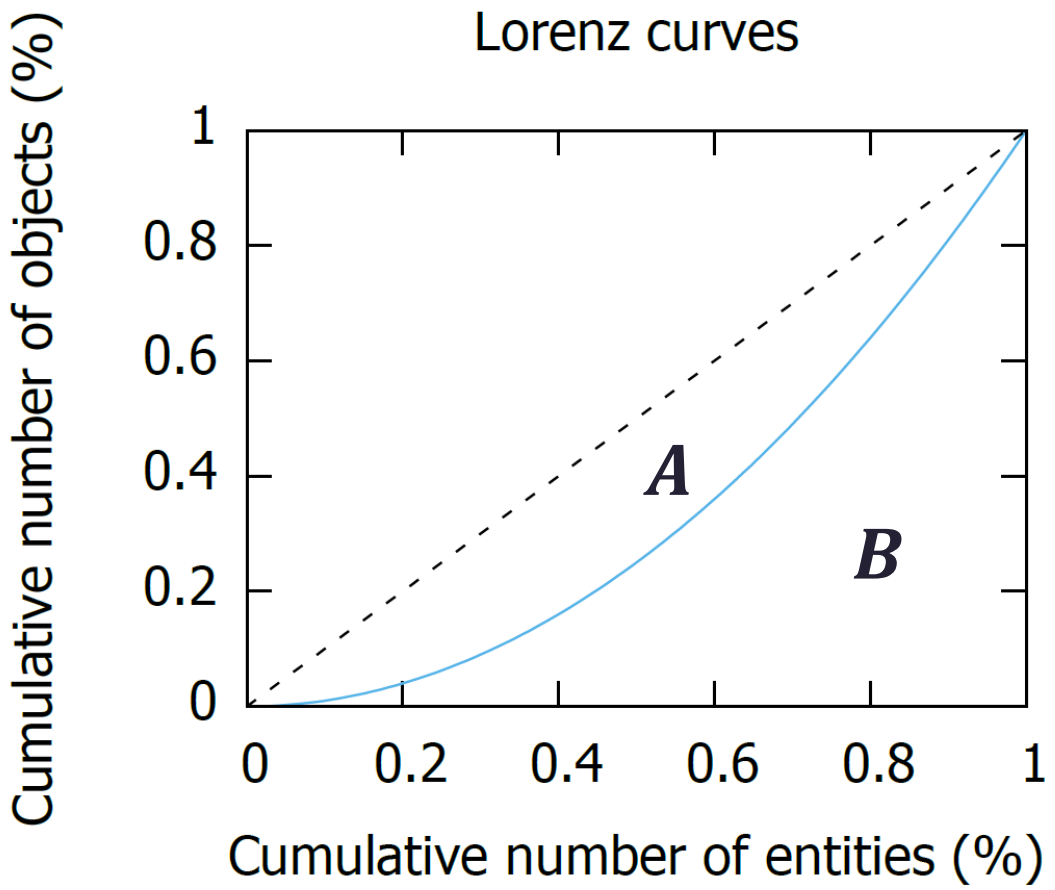
The gap between the lowest and the highest counts is large.

- **C2: Concentration**

Inequality is higher than a uniform distribution.

A relevant counting property for ranking leads to inequality!

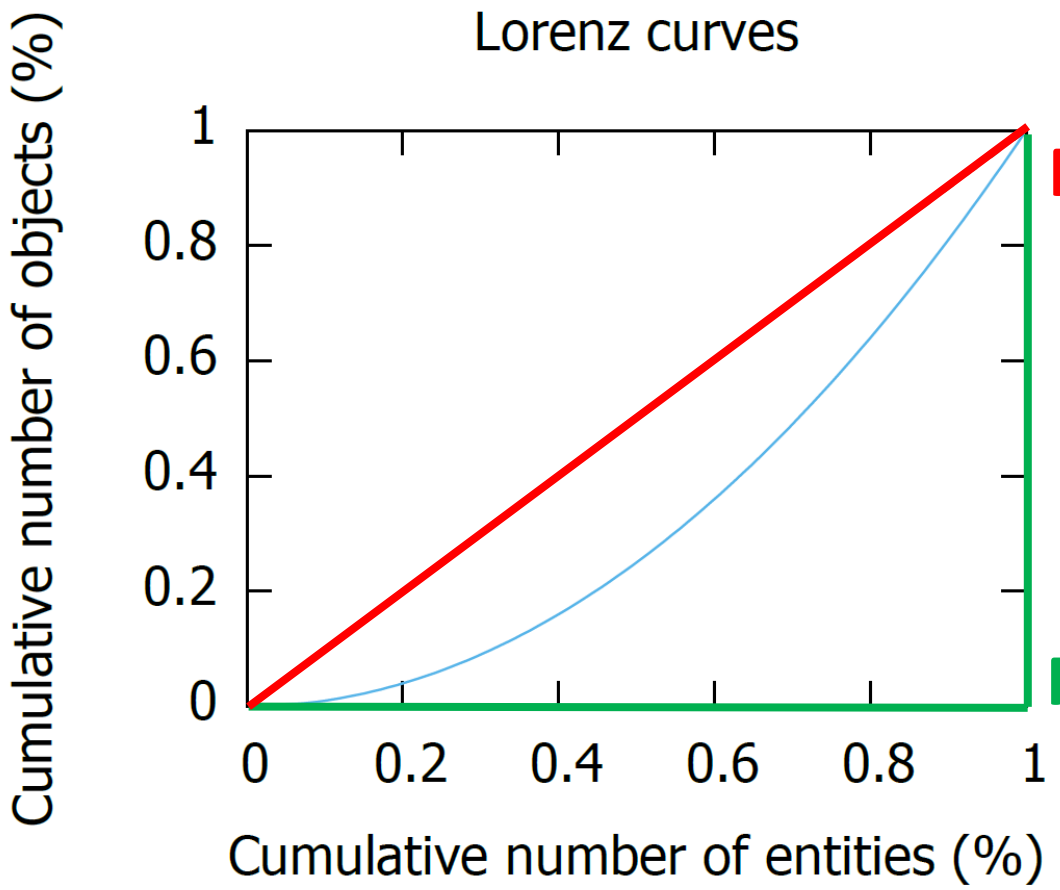
Ranking score = inequality



$$G = \frac{A}{A + B}$$

We use the Gini coefficient as inequality measure.

Ranking score = inequality

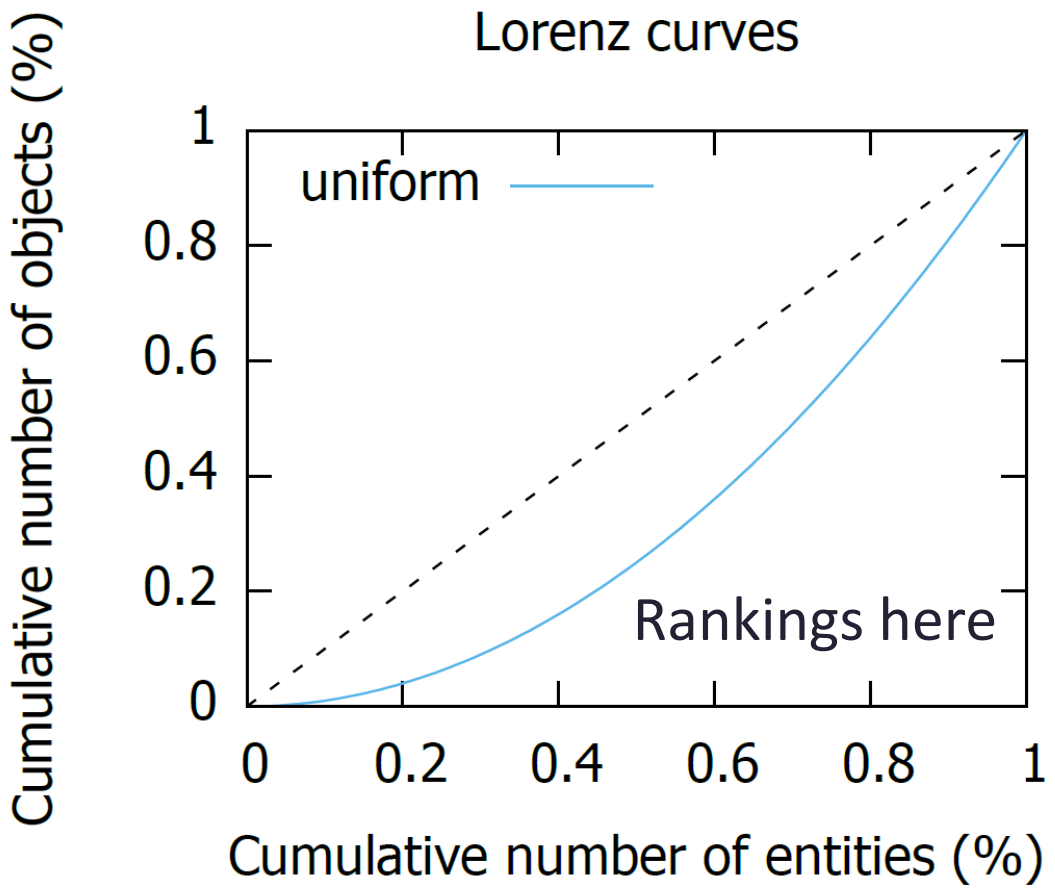


Perfect equality

Perfect inequality

We use the Gini coefficient as inequality measure.

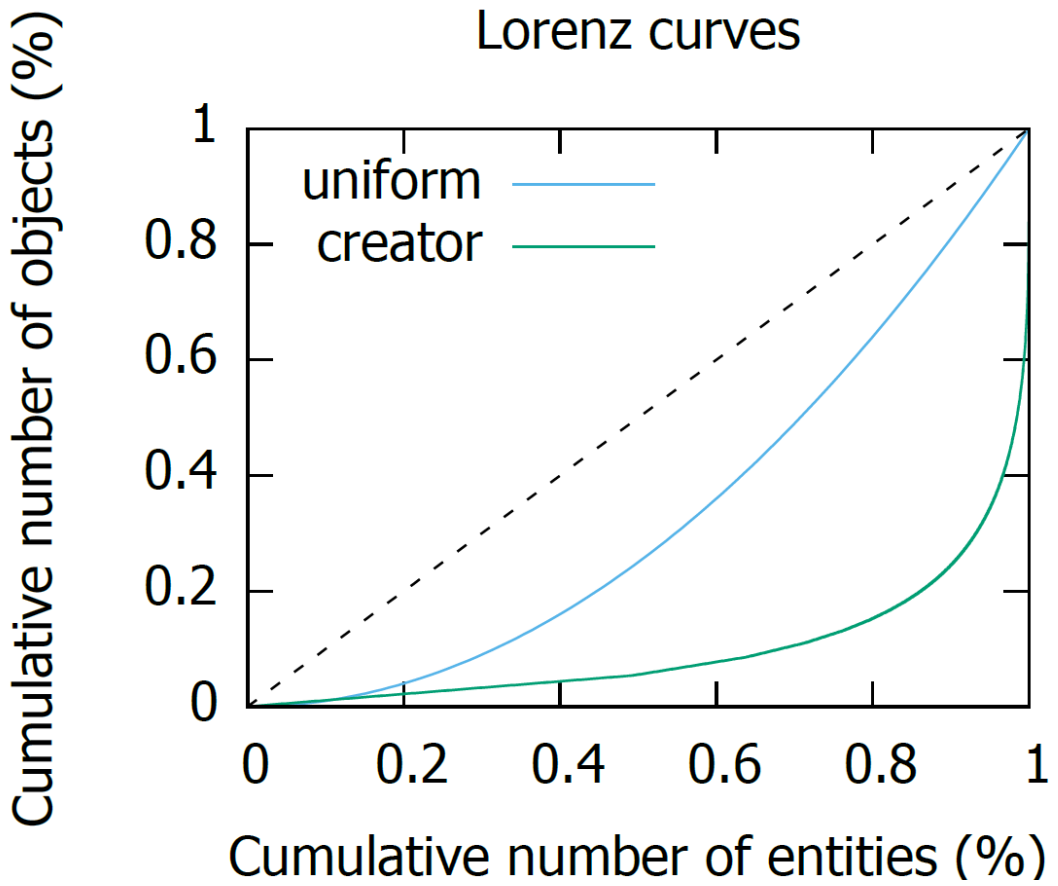
Ranking score = inequality



$$G = \frac{1}{3}$$

Criteria C1 Range + C2 Concentration = Gini coefficient ≥ 1/3

Ranking score = inequality



$$G = \frac{1}{3}$$

$$\geq$$

$$G = 0,805$$

Ranking score = inequality

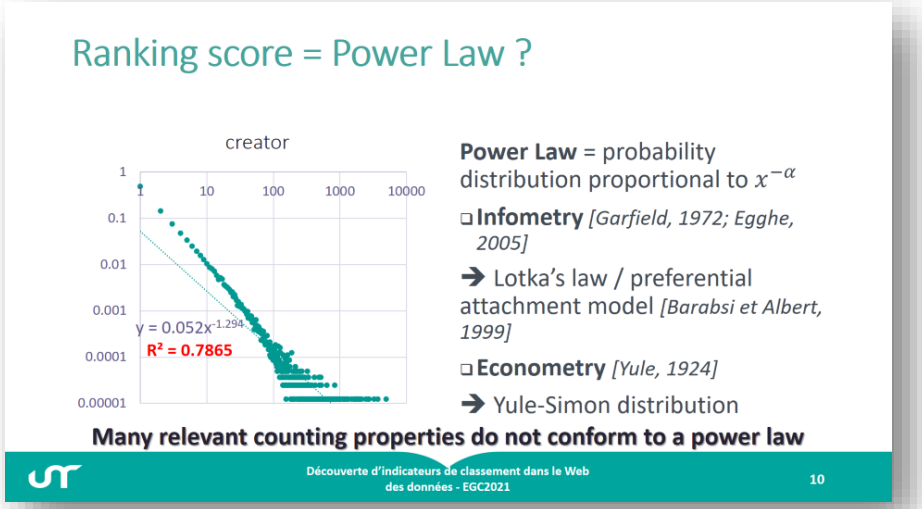
Ranking of the most important painters by the number of paintings:

Rang	Peintre	Nbr. peintures	Rang	Peintre	Nbr. peintures
1	Kalervo Palsa	1 940	11	Mary Vaux Walcott	787
2	Edvard Munch	1 789	12	David Teniers the Younger	718
3	Anthony van Dyck	1 215	13	Pablo Picasso	705
4	John Everett	1 021	14	Rembrandt	690
5	Peter Paul Rubens	1 001	15	Paul Cézanne	656
6	George Catlin	968	16	Camille Pissarro	644
7	Claude Monet	940	17	Panos Terlemezian	631
8	Pierre-Auguste Renoir	907	18	Eliseu Visconti	590
9	Vincent van Gogh	890	19	Edwin Austin Abbey	585
10	Jean-Baptiste-Camille Corot	879	20	Emanuel Fohn	549

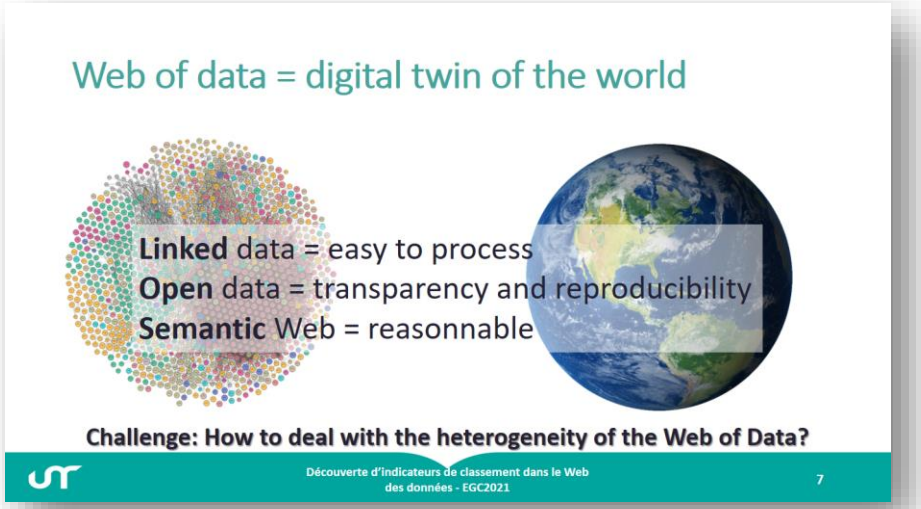
Use case on Wikidata

- **Knowledge Base: Wikidata (June 2020)**
- **659 inverse properties (having URI as range)**

1 Gini coefficient vs PL compliance?

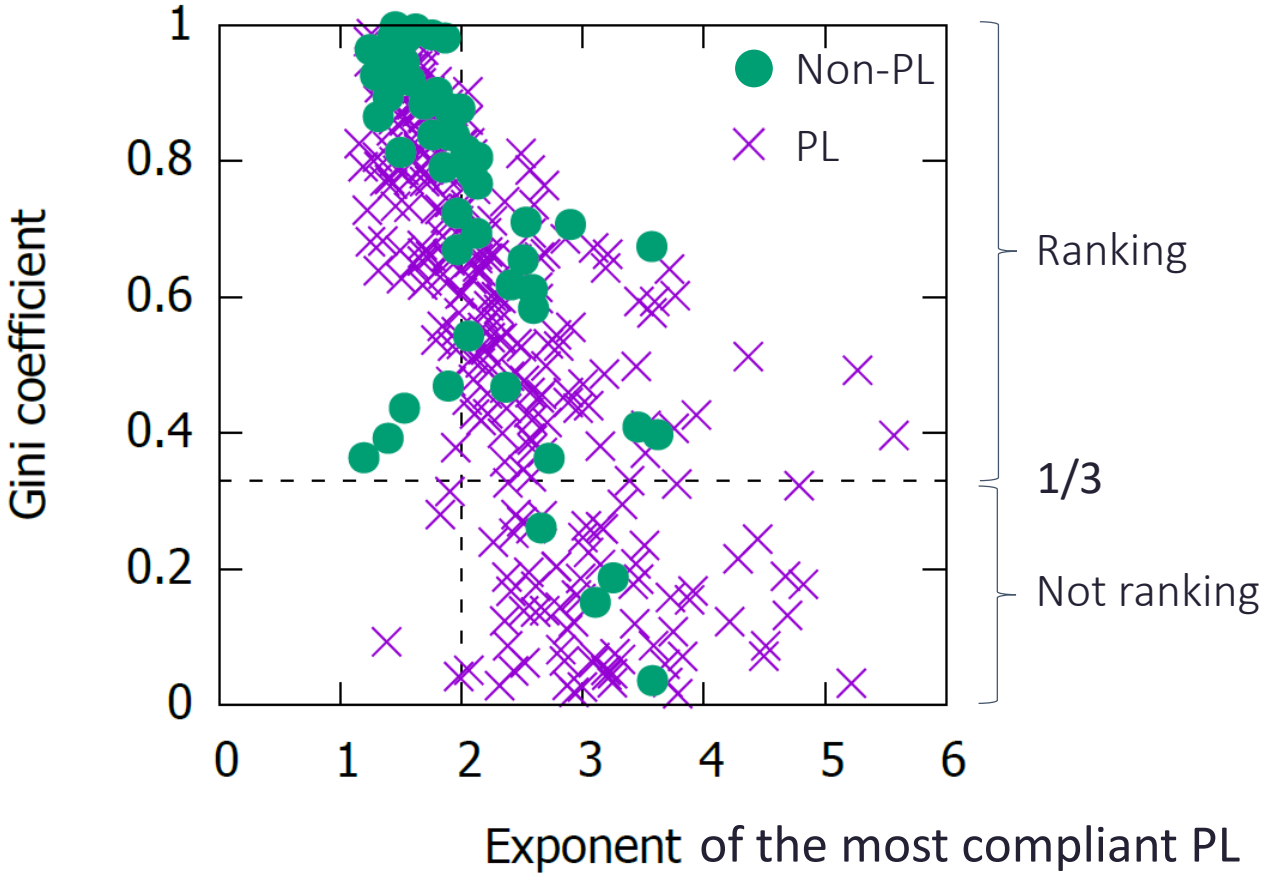


2 Relevant rankings for everything?



1

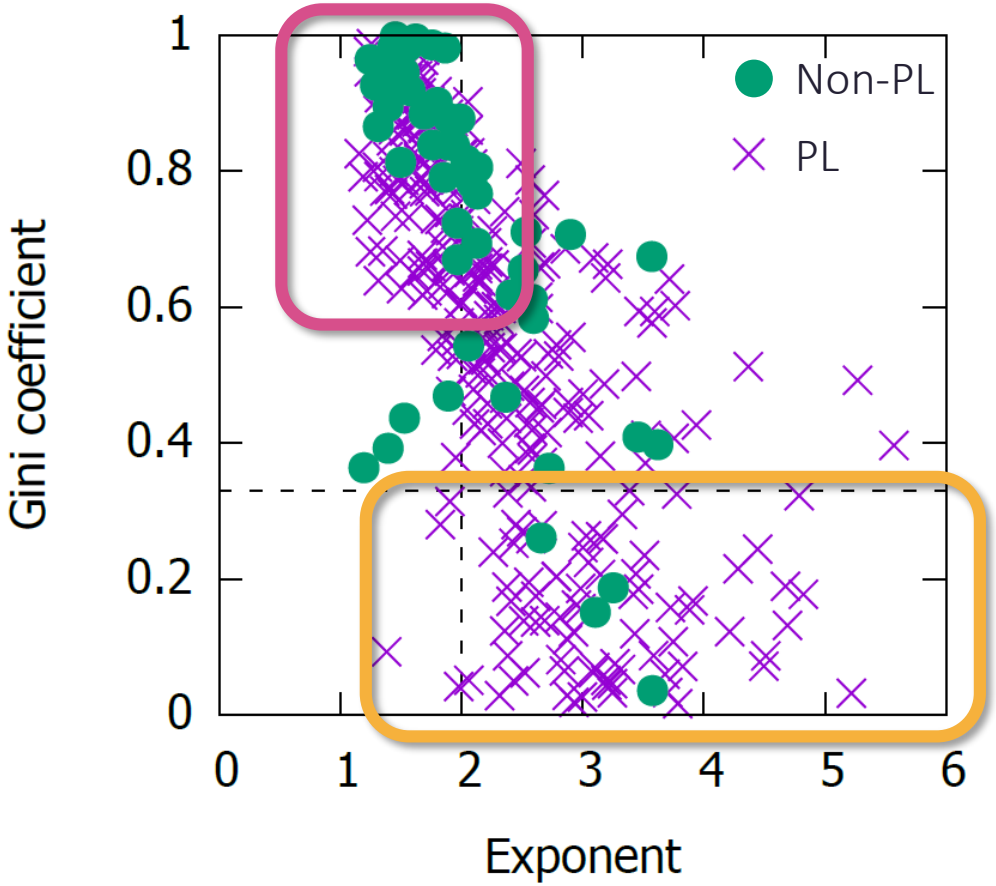
Gini coefficient vs Power Law compliance



1

Gini coefficient vs Power Law compliance

Many relevant properties are not PL!



Many irrelevant properties are PL!

1

Top 20 properties for ranking

Propriété p	Gini G_{p-1}	#entités	Propriété p	Gini G_{p-1}	#entités
*lang. of work or name (P407)	0.997	1172	*occupation (P106)	0.982	12 147
instance (P31)	0.995	65 777	country of citizenship (P27)	0.981	3 093
*writing system (P282)	0.994	447	*religion (P140)	0.981	1 245
epoch (P6259)	0.994	168	country of origin (P495)	0.981	1 252
sex or gender (P21)	0.987	113	*described by source (P1343)	0.981	28 828
color (P462)	0.986	242	*instrument (P1303)	0.98	818
*manner of death (P1196)	0.985	309	*collection (P195)	0.98	7 457
*material used painting (P186)	0.985	4 163	*subject has role (P2868)	0.979	878
lang. spoken or written (P1412)	0.984	1 374	honorific prefix (P511)	0.978	196
country (P17)	0.983	2 553	native language (P103)	0.976	784

Non-PL

PL

PL

*Not compliant with a PL

Power law inconsistent: native language, but not language of work

The best ranking by class (P31)

Classe	Propriété p pour classer	Gini G_{p-1}	#classés	#instances
Human (Q5)	creator (P170)	0.805	90	718 384
Taxon (Q16521)	parent taxon (P171)	0.781	91	88 720
position (Q4164871)	position held (P39)	0.877	88	81 046
Scientific journal (Q5633421)	published in larger work (P1433)	0.913	99	30 146
Communes of France (Q484170)	destination point (P1444)	0.531	17	29 629
Surname (Q101352)	second family name (P1950)	0.657	69	29 175
Business (Q4830453)	manufacturer (P176)	0.811	57	24 398
Band (rock and pop) (Q215380)	performer (P175)	0.672	14	22 774
River (Q4022)	mouth of the watercourse (P403)	0.614	52	21 642
scholarly article (Q13442814)	described by source (P1343)	0.981	12	19 256
association football club (Q476028)	participating team (P1923)	0.787	68	17 253
Gene (Q7187)	regulates (molecular biology) (P128)	0.63	57	14 458
Road (Q34442)	connects with (P2789)	0.487	36	13 784
Immortalised cell line (Q21014462)	parent cell line (P3432)	0.86	60	13 013
Organization (Q43229)	member (P463)	0.838	10	10 977
Award (Q618779)	award received human (P166)	0.896	17	9 534
Season (sports) (Q27020041)	participant (P1344)	0.787	13	8 005
language (Q34770)	language of work or name (P407)	0.997	88	7 788
Comune (Q747074)	ancestral home (P66)	0.419	11	7 731
University (Q3918)	affiliation (P1416)	0.827	59	7 678

Many different fields: biology, sport, geography, music,…

2

The best ranking by class (P31)

Classe	Propriété p pour classer	Gini G_{p-1}	#classés	#instances
Human (Q5)	creator (P170)	0.805	90	718 384
Taxon (Q16521)	parent taxon (P171)	0.781	91	88 720
position (Q4164871)	position held (P39)	0.877	88	81 046
Scientific journal (Q5633421)	published in larger work (P1433)	0.913	99	30 146
Communes of France (Q484170)	destination point (P1444)	0.531	17	29 629
Surname (Q101352)	second family name (P1950)	0.657	69	29 175
Business (Q4830453)	manufacturer (P176)	0.811	57	24 398
Band (rock and pop) (Q215380)	performer (P175)	0.672	14	22 774
River (Q4022)	mouth of the watercourse (P403)	0.614	52	21 642
scholarly article (Q13442814)	described by source (P1343)	0.981	12	19 256
association football club (Q476028)	participating team (P1923)	0.787	68	17 253
Gene (Q7187)	regulates (molecular biology) (P128)	0.63	57	14 458
Road (Q34442)	connects with (P2789)	0.487	36	13 784
Immortalised cell line (Q21014462)	parent cell line (P3432)	0.86	60	13 013
Organization (Q43229)	member (P463)	0.838	10	10 977
Award (Q618779)	award received human (P166)	0.896	17	9 534
Season (sports) (Q27020041)	participant (P1344)	0.787	13	8 005
language (Q34770)	language of work or name (P407)	0.997	88	7 788
Comune (Q747074)	ancestral home (P66)	0.419	11	7 731
University (Q3918)	affiliation (P1416)	0.827	59	7 678

Existing measures in scientometrics

2

The best ranking by occupation (P106)

Profession	Propriété p pour classer	Gini G_{p-1}	#classés	#instances
Actor (Q33999)	cast member (P161)	0.674	50	79 386
Writer (Q36180)	screenwriter (P58)	0.577	31	55 002
Painter (Q1028181)	creator (P170)	0.805	75	45 771
Film director (Q2526255)	director (P57)	0.643	61	40 620
film actor (Q10800557)	cast member (P161)	0.674	64	38 444
Screenwriter (Q28389)	director (P57)	0.643	58	36 938
Singer (Q177220)	performer (P175)	0.672	43	29 127
television actor (Q10798782)	cast member (P161)	0.674	40	25 231
Composer (Q36834)	composer (P86)	0.736	72	24 346
university teacher (Q1622272)	author (P50)	0.859	22	24 076
Journalist (Q1930187)	presenter (P371)	0.457	21	22 884
Film producer (Q3282637)	director (P57)	0.643	50	18 325
Catholic priest (Q250867)	consecrator (P1598)	0.528	48	17 619
sport cyclist (Q2309784)	classification of race participants (P2321)	0.694	72	17 540
Architect (Q42973)	architect (P84)	0.544	42	17 268
Poet (Q49757)	author (P50)	0.859	19	16 417
stage actor (Q2259451)	cast member (P161)	0.674	33	15 919
Musician (Q639669)	composer (P86)	0.736	26	15 130
Historian (Q201788)	editor (P98)	0.378	15	10 814
sculptor (Q1281618)	creator (P86)	0.804	21	9 784

Very coherent rankings!

Conclusion

- **Ranking score = inequality**
  - A transdisciplinary model based on 2 criteria (range+concentration)
  - Use case on Wikidata: 386 properties for ranking >2 M entities (≈everything)
- **Future work:** Combining ranking for advanced scores like h-index

Rg	Peintre	h
1	Vincent van Gogh	17
2	Leonardo da Vinci	16
3	Johannes Vermeer	16
4	Raphael	15
5	Pieter Bruegel the Elder	15
6	Caravaggio	15
7	Rembrandt	12
8	Titian	12
9	Édouard Manet	12
10	Diego Velázquez	11

Rg	Peintre	h
11	Hieronymus Bosch	11
12	Jan van Eyck	11
13	Pierre-Auguste Renoir	11
14	J.-A.-D. Ingres	11
15	Gustave Courbet	10
16	Joseph Wright of Derby	10
17	El Greco	10
18	Sandro Botticelli	10
19	Francisco Goya	10
20	Salvador Dalí	10

Rg	Peintre	h
21	Jacques-Louis David	10
22	Paul Gauguin	10
23	Caspar David Friedrich	9
24	Rogier van der Weyden	9
25	Pablo Picasso	9
26	Claude Monet	9
27	W.-A. Bouguereau	9
28	Albrecht Dürer	9
29	Gustav Klimt	8
30	Jan Matejko	8

- **We need more efforts for exploiting the Web of Data as strategic data for transdisciplinary method for real-world analytics!**

EGC 2022
BLOIS
24-28 janvier

egc2022.univ-tours.fr

Sihem Amer-Yahia

Présidente du comité de programme

Dates de soumission

Résumé : 8 octobre

Article : 15 octobre

https://twitter.com/egc2022