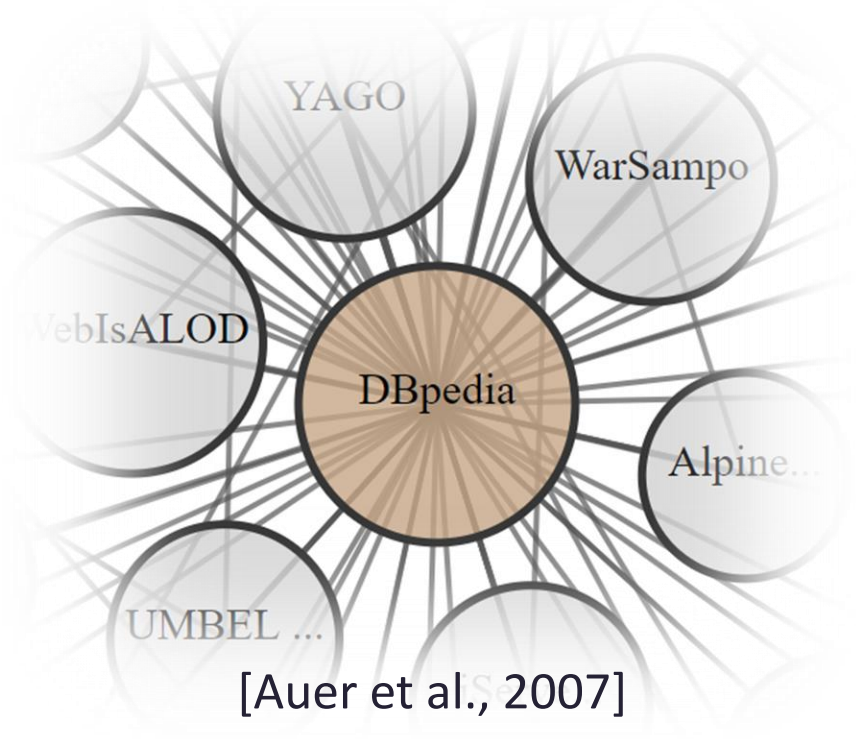
Representativeness of Knowledge Bases with the Generalized Benford's Law

Arnaud Soulet, Arnaud Giacometti, Béatrice Markhoff and Fabian M. Suchanek

University of Tours

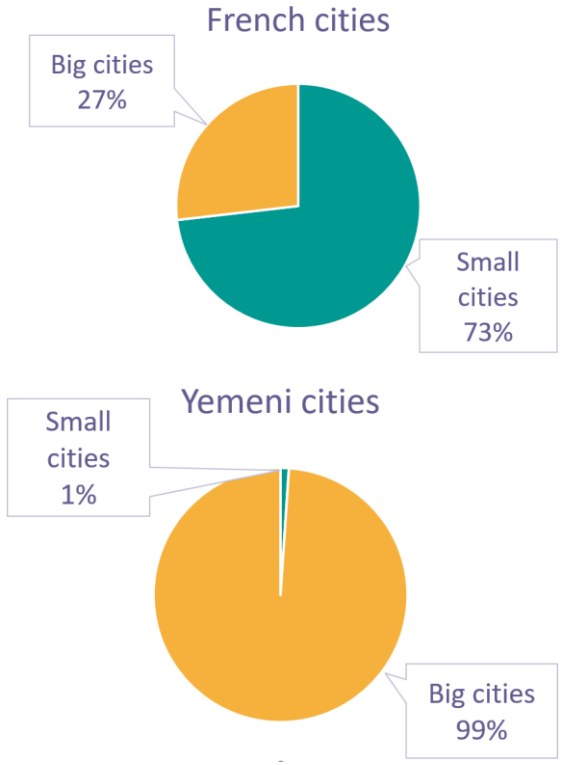
Telecom ParisTech

Reliability of queries on Knowledge Bases



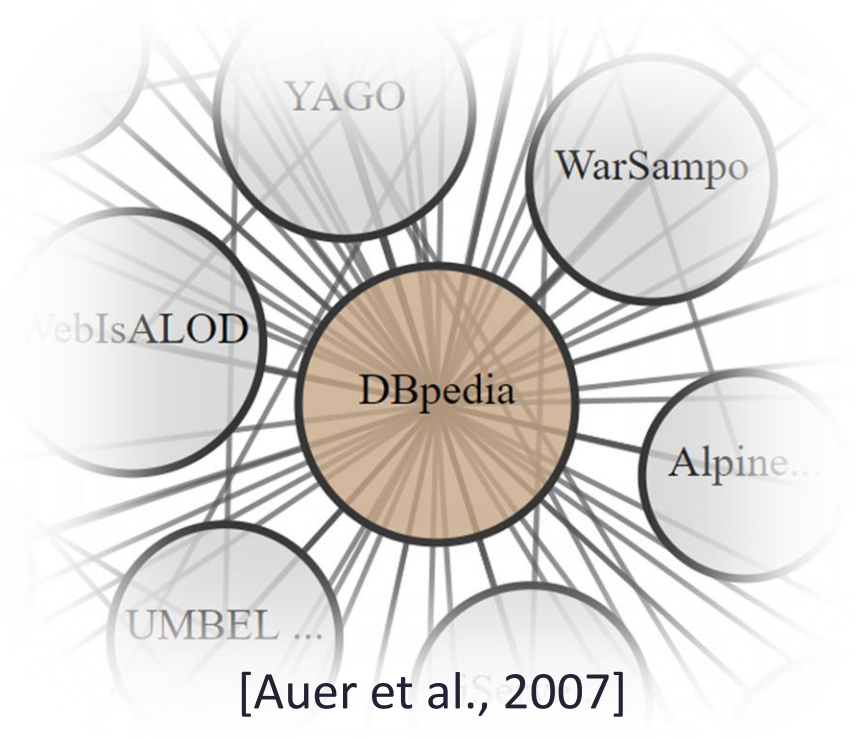
[Auer et al., 2007]

statistical query



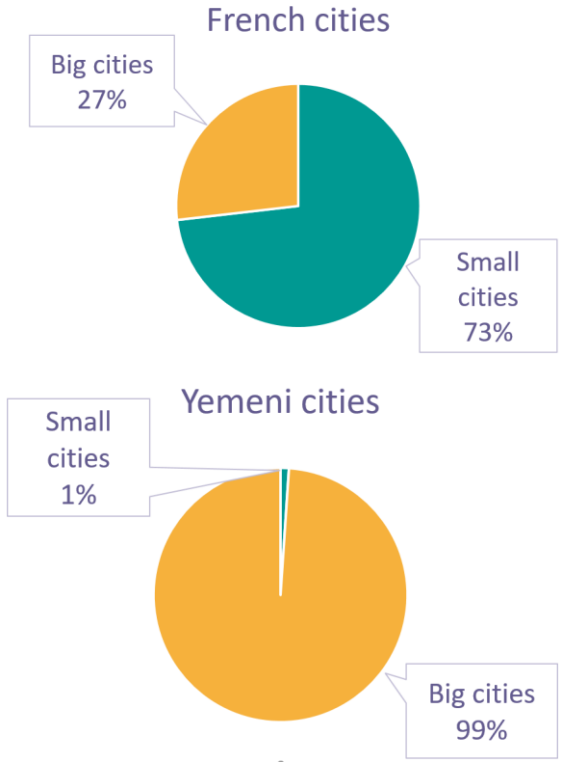
How many cities are small (<1k inhabitants) in France/Yemen?

Reliability of queries on Knowledge Bases



[Auer et al., 2007]

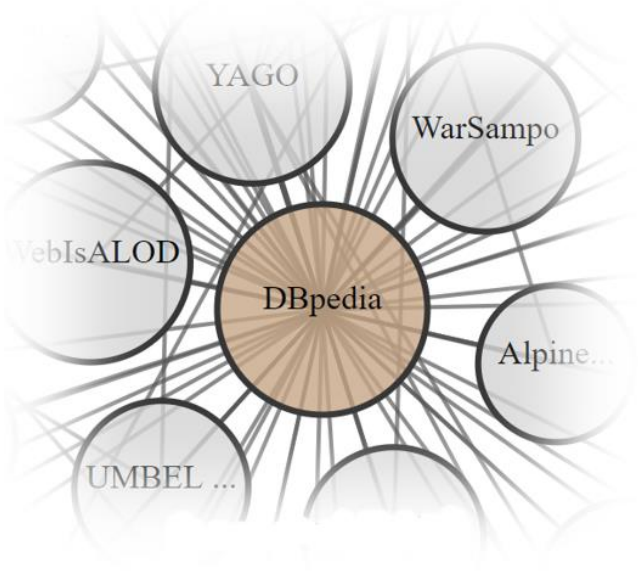
statistical query



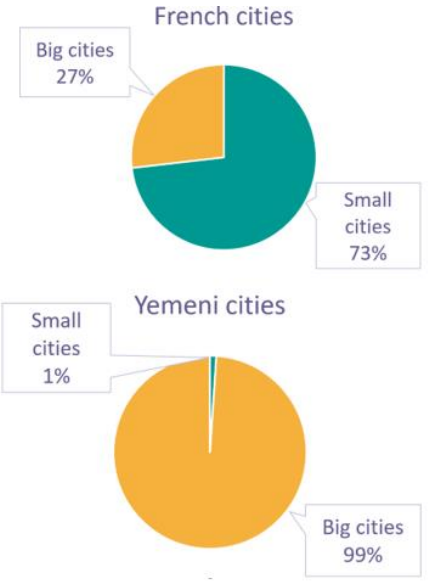
Does Yemen really not have any small cities?

Reliability of queries on Knowledge Bases

crowdsourcing



statistical query

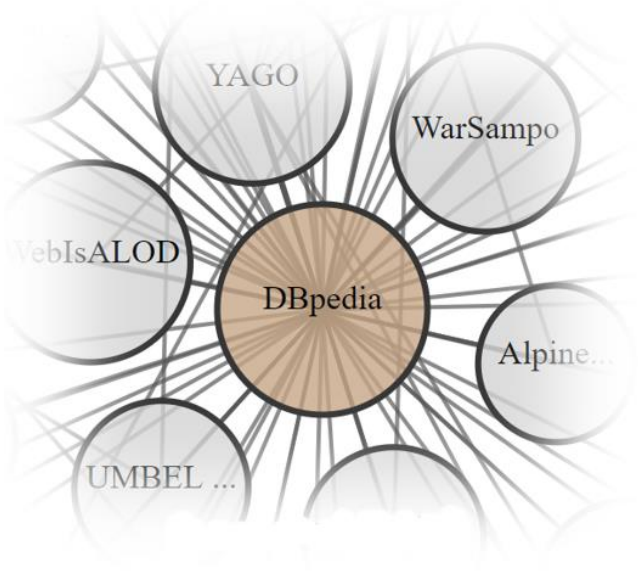


Voluntary bias

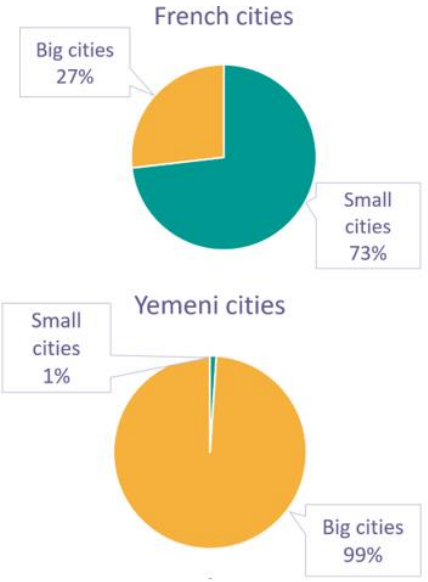
[Callahan and Herring, 2011;Wagner et al., 2015]

Reliability of queries on Knowledge Bases

crowdsourcing



statistical query



Voluntary bias

[Callahan and Herring, 2011;Wagner et al., 2015]

We do not know the KB biases, but the statistics can give us a hint?

Missing facts

Yemeni city:	Sanaa	Aden	Taiz
Population:	1,937,451	760,923	missing 615,222

[Darari et al., 2016; Galarraga et al., 2017; Lajus and Suchanek, 2018; Razniewski et al., 2015; Razniewski et al., 2016]

Several methods for estimating the completeness for facts

Missing facts ≠ Missing entities + missing facts




Yemeni city:

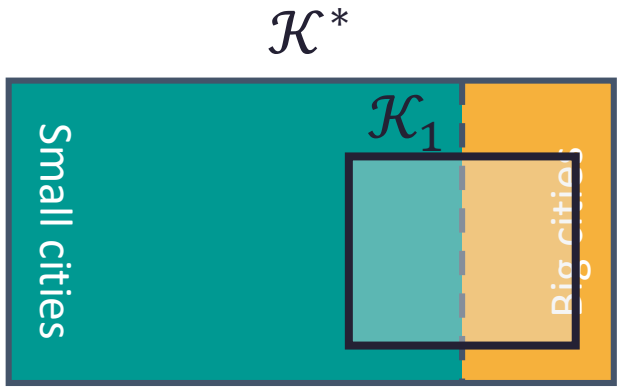
Population:

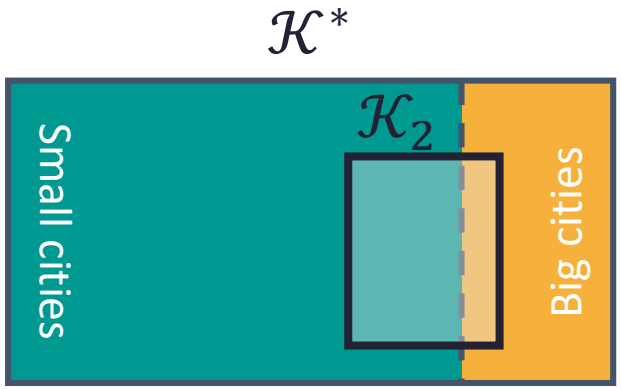
Sanaa	Aden	Taiz	Haid al-Jazil
1,937,451	760,923	missing 615,222	missing few

Missing facts due to missing entities are ignored!

Completeness
= #present facts / (#present facts + #missing facts)

Assuming that $\mathcal{K}^*$ is an ideal KB (= correct + complete):

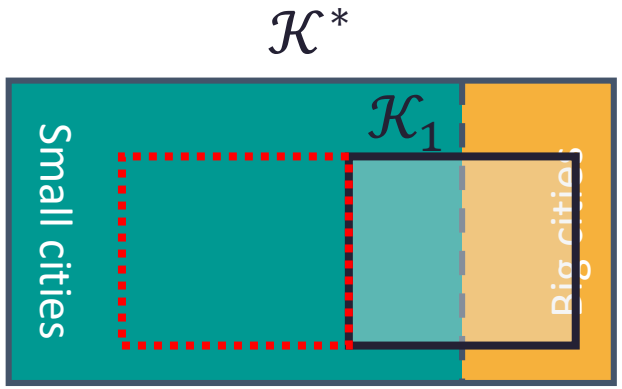




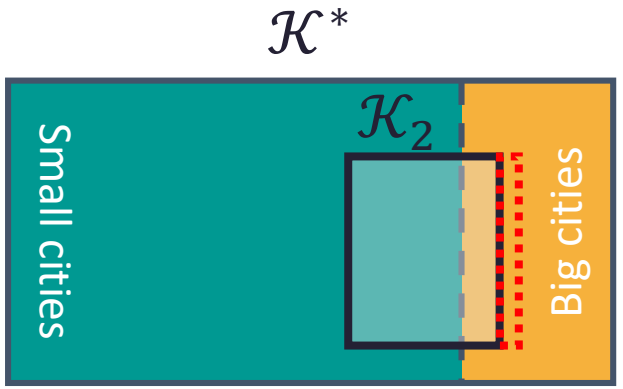
What is the best KB between $\mathcal{K}_1$ and $\mathcal{K}_2$ for statistical queries?

Completeness ≠ Representativeness

Assuming that $\mathcal{K}^*$ is an ideal KB (= correct + complete):



More complete, less representative

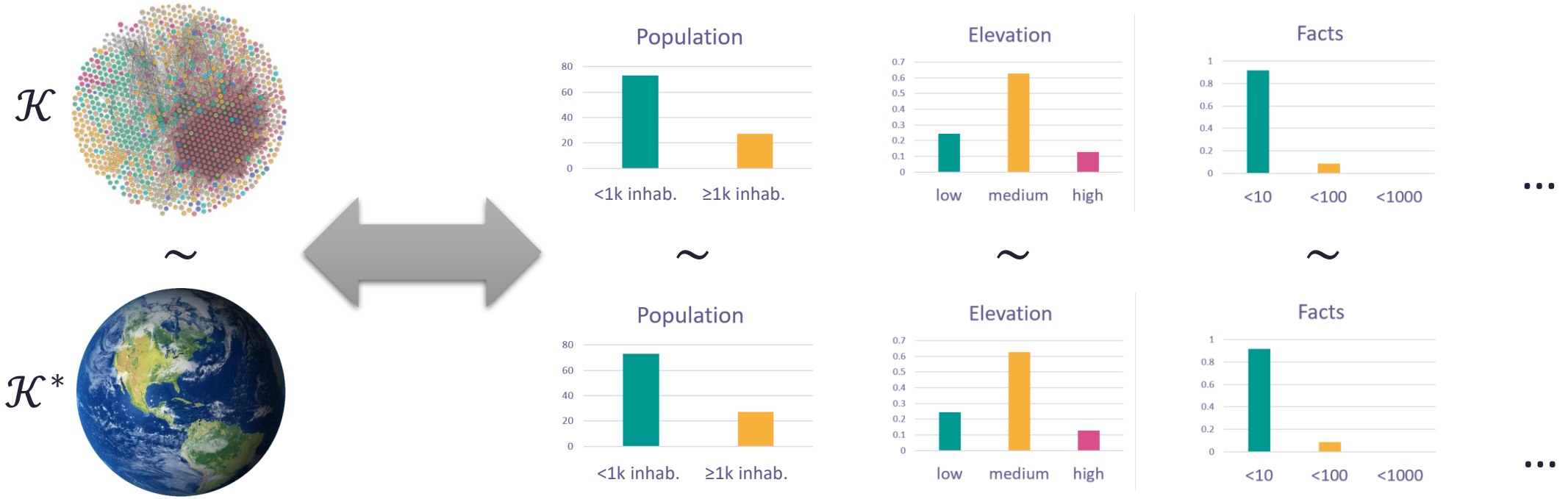


Less complete, more representative

Representativeness is more important than completeness for statistics!

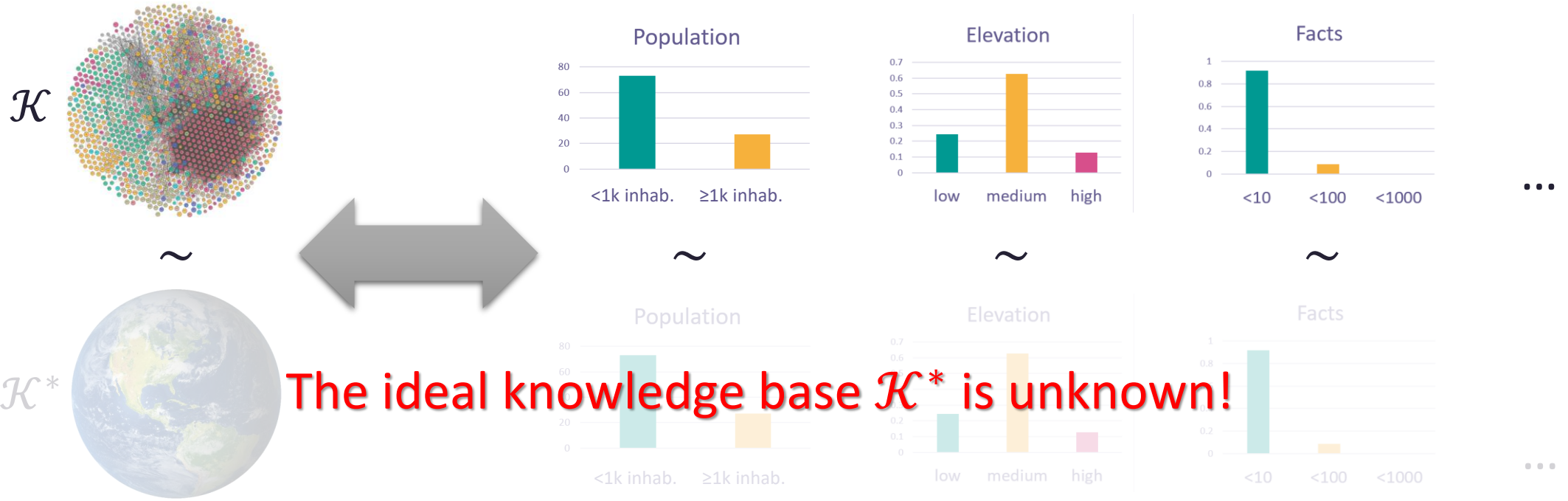
Representativeness of Knowledge Bases

A KB $\mathcal{K}$ is representative of $\mathcal{K}^*$ iff the distribution remains the same for all uniform-sampling invariant measures.



Representativeness of Knowledge Bases

A KB $\mathcal{K}$ is representative of $\mathcal{K}^*$ iff the distribution remains the same for all uniform-sampling invariant measures.



Challenge: How to estimate the representativeness?

Example: population of capitals

Abidjan	Bangkok	Conakry	Kingston	Mogadishu	Santiago
Abuja	Beijing	Dakar	Kinshasa	Montevideo	Seoul
Accra	Belgrade	Damascus	Kuala Lumpur	Nairobi	Sofia
Addis Ababa	Berlin	Dhaka	Lilongwe	Niamey	Taipei
Algiers	Bogota	Doha	Lima	Ouagadougou	Tashkent
Amman	Brasilia	Erbil	London	Paris	Tbilisi
Ankara	Brazzaville	Freetown	Luanda	Phnom Penh	Tegucigalpa
Antananarivo	Bucharest	Havana	Lusaka	Prague	Tokyo
Ashgabat	Budapest	Islamabad	Madrid	Pyongyang	Tripoli
Bahawalpur	Buenos Aires	Jakarta	Managua	Quito	Tunis
Baku	Cairo	Kabul	Maputo	Riyadh	Ulaanbaatar
Bamako	Caracas	Khartoum	Mexico City	Sana'a	Vienna

Example: population of capitals

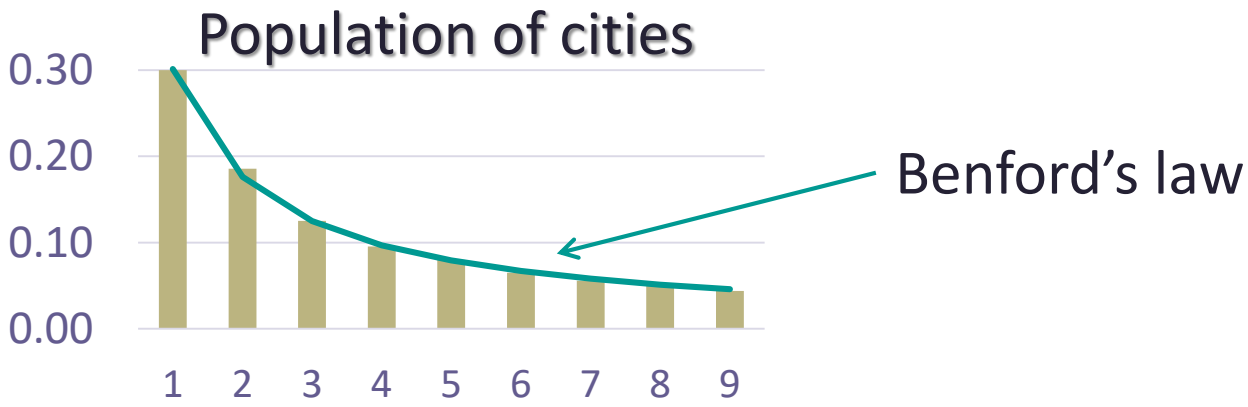
4 707 404	8 280 925	1 660 973	1 041 084	1 750 000	6 158 080
1 235 880	21 700 000	1 146 053	10 125 000	1 305 082	9 971 111
2 291 352	1 166 763	1 711 000	1 768 000	3 138 369	1 260 120
3 384 569	3 610 156	6 970 105	1 077 116	1 302 910	2 704 974
3 415 811	7 878 783	1 351 000	8 852 000	1 626 950	2 309 600
4 007 526	2 556 149	1 025 000	8 673 713	2 229 621	1 118 035
4 587 558	1 827 000	1 050 301	2 825 311	1 501 725	1 157 509
1 613 375	1 883 425	2 106 146	1 742 979	1 267 449	13 617 445
1 031 992	1 759 407	1 900 000	3 141 991	2 581 076	1 126 000
1 052 000	2 890 151	9 607 787	2 205 676	2 671 191	1 056 247
2 122 300	10 230 350	3 678 034	1 766 184	7 125 180	1 372 000
1 809 106	3 273 863	5 185 000	8 918 653	1 937 451	1 852 997

Example: population of capitals

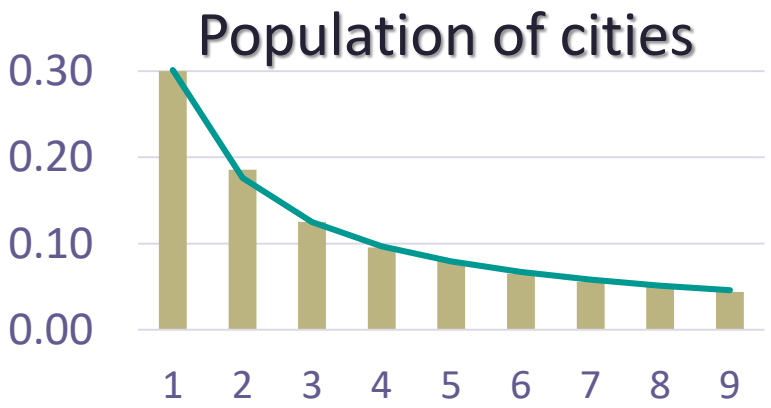
4 707 404	8 280 925	1 660 973	1 041 084	1 750 000	6 158 080
1 235 880	21 700 000	1 146 053	10 125 000	1 305 082	9 971 111
2 291 352	1 166 763	1 711 000	1 768 000	3 138 369	1 260 120
3 384 569	3 610 156	6 970 105	1 077 116	1 302 910	2 704 974
3 415 811	7 878 783	1 351 000	8 852 000	1 626 950	2 309 600
4 007 526	2 556 149	1 025 000	8 673 713	2 229 621	1 118 035
4 587 558	1 827 000	1 050 301	2 825 311	1 501 725	1 157 509
1 613 375	1 883 425	2 106 146	1 742 979	1 267 449	13 617 445
1 031 992	1 759 407	1 900 000	3 141 991	2 581 076	1 126 000
1 052 000	2 890 151	9 607 787	2 205 676	2 671 191	1 056 247
2 122 300	10 230 350	3 678 034	1 766 184	7 125 180	1 372 000
1 809 106	3 273 863	5 185 000	8 918 653	1 937 451	1 852 997

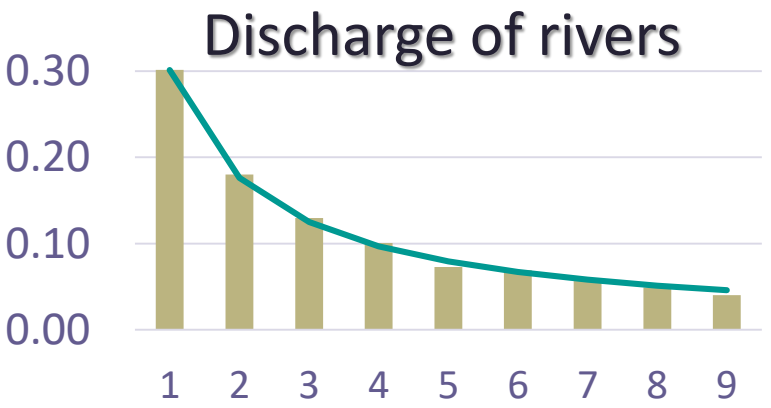
What is the distribution of the first significant digit of capital inhabitants?

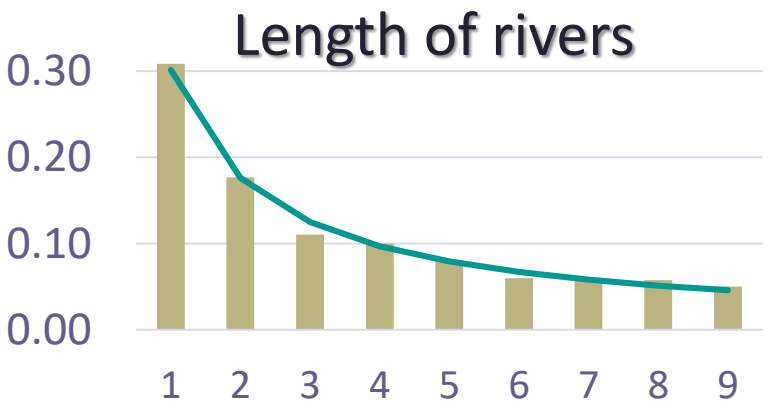
Benford's law



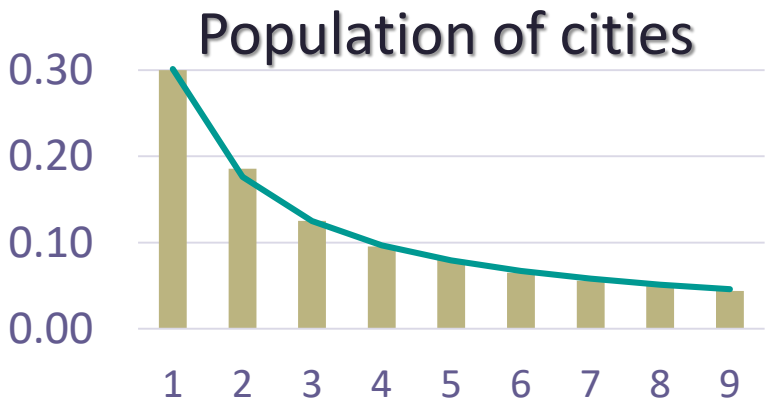
Benford's law

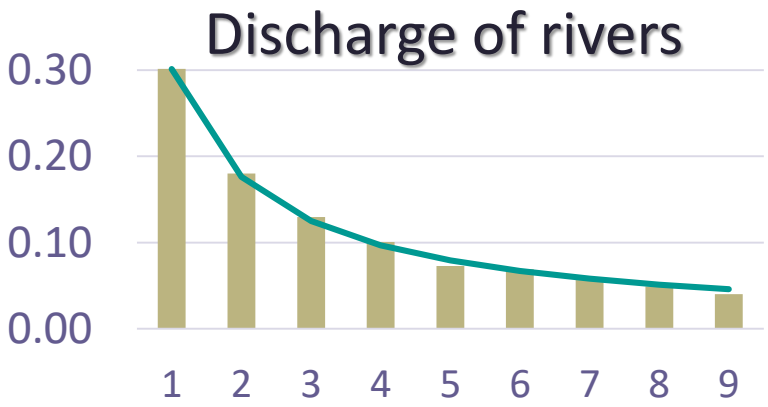


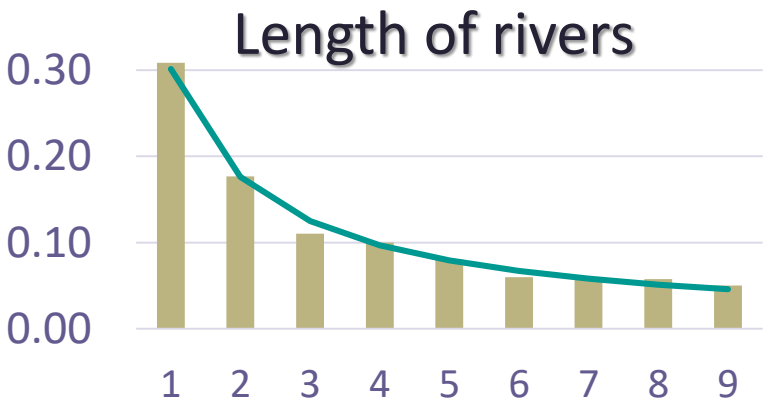




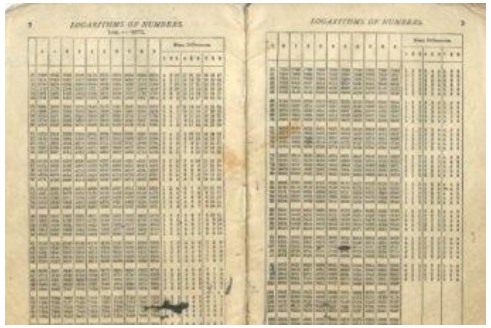
Benford's law





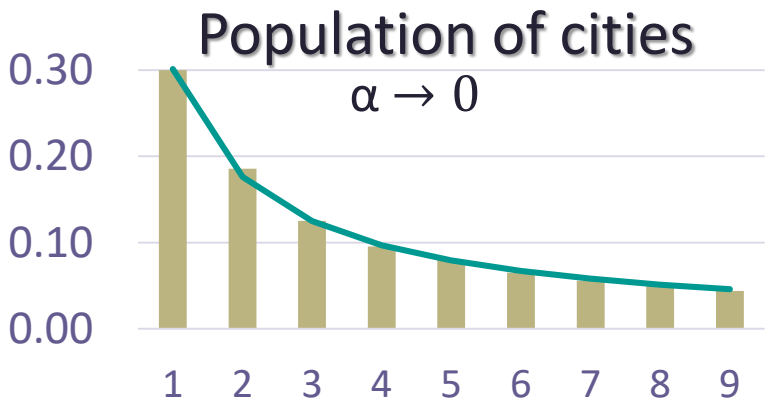


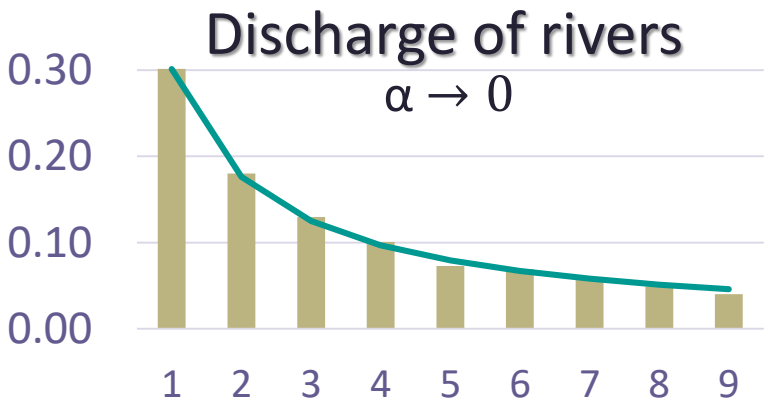
$$P(first\ digit(X) = d) = \log\left(1 + \frac{1}{d}\right)$$

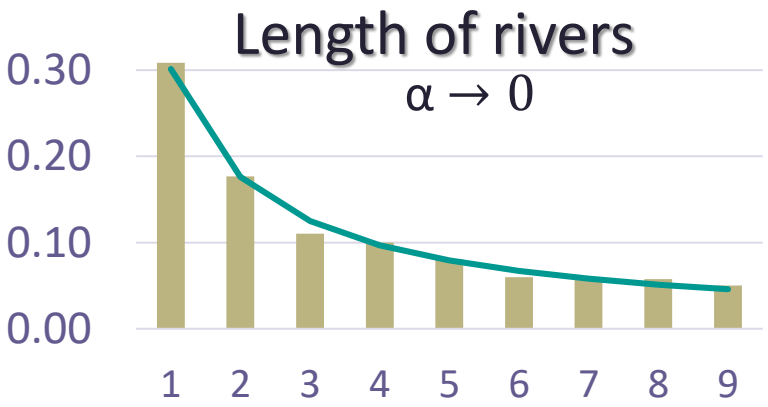


[Newcomb, 1881;Benford, 1938]

The Generalized Benford's Law



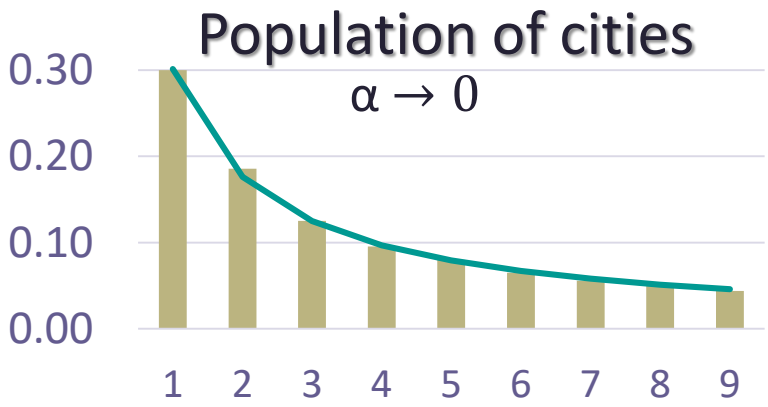


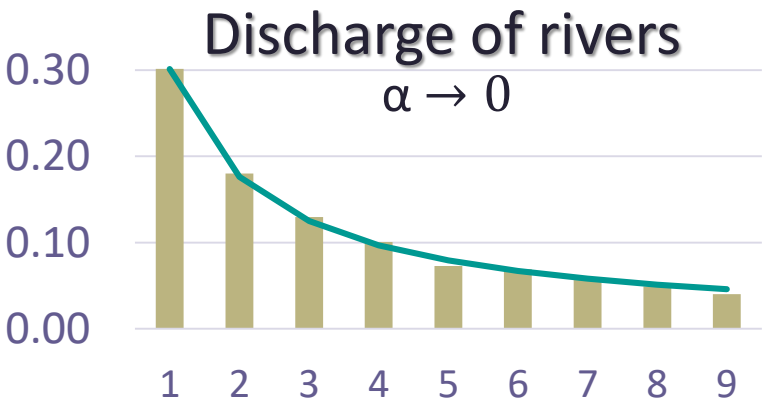


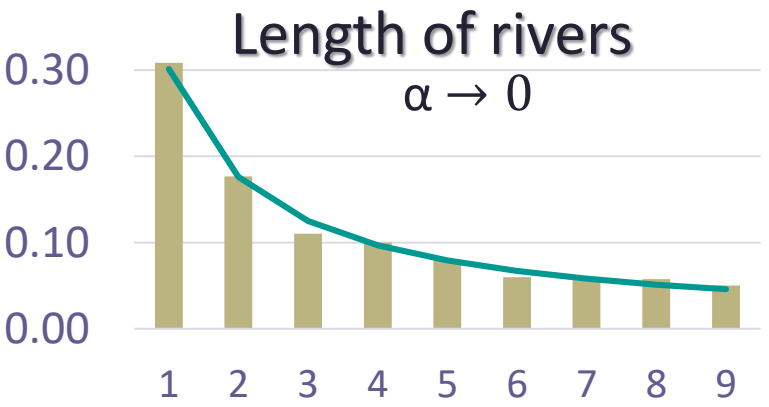
$$P(first\ digit(X) = d) = \frac{(1+d)^{\alpha} - d^{\alpha}}{10^{\alpha} - 1}$$

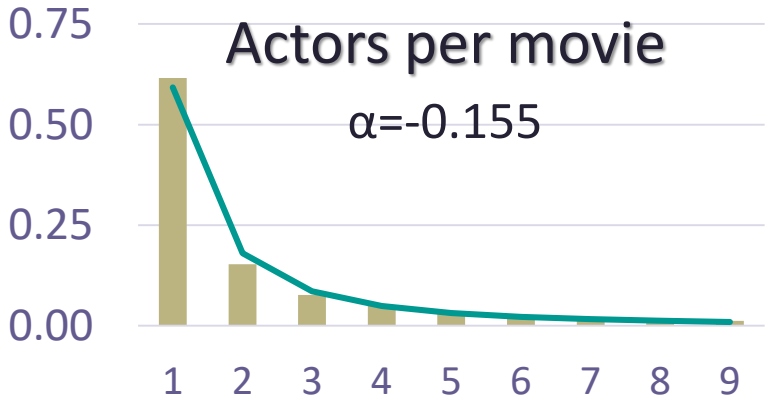
[Hürlimann, 2014]

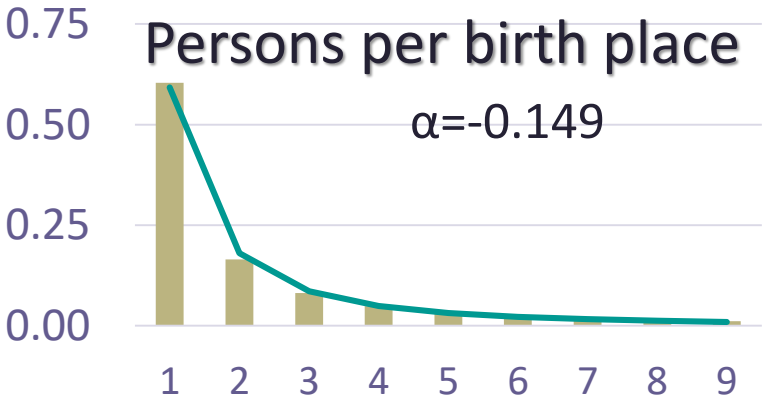
The Generalized Benford's Law

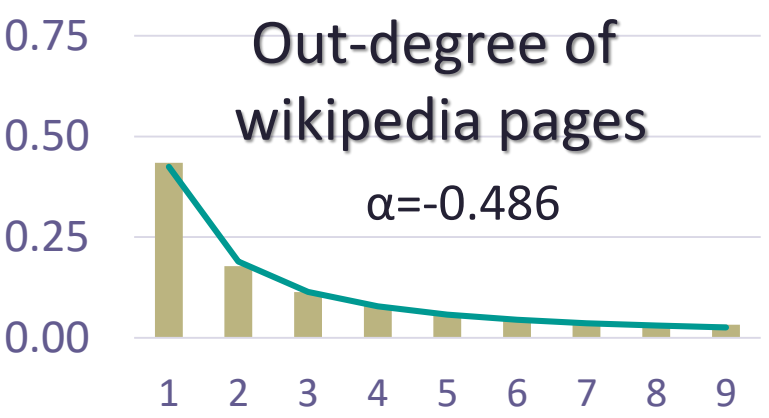








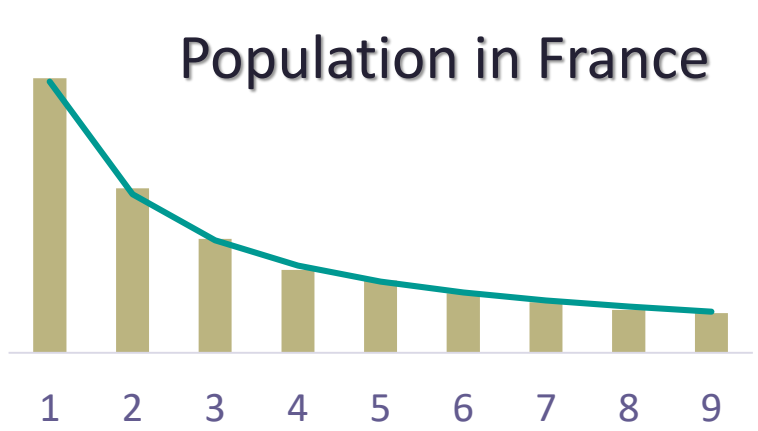




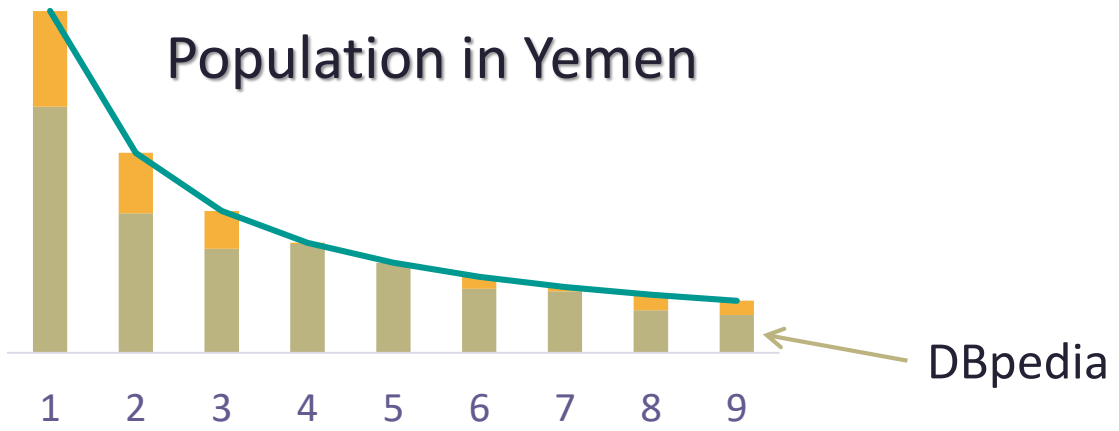
Key idea of our method

representativeness = compliance with the Generalized Benford's Law

$$= \frac{\#present\_facts}{\#present\_facts + \#missing\_facts\_for\_compliance}$$



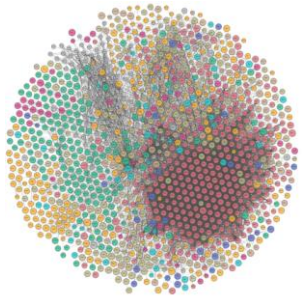
Representativeness = 97%



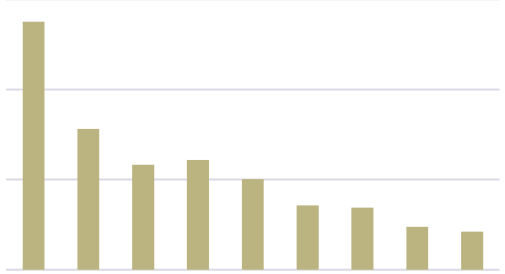
Representativeness = 79%

Our method in supervised context

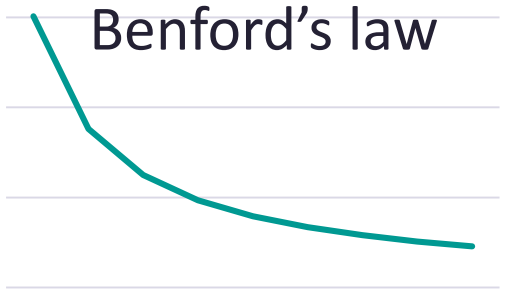
facts of r on $\mathcal{K}$



distribution of the fsd



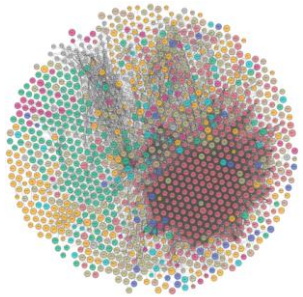
facts of r on $\mathcal{K}^*$



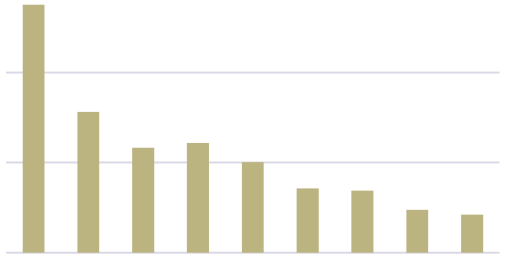
Using the known distribution of the first significant digit

Our method in supervised context

facts of r on $\mathcal{K}$

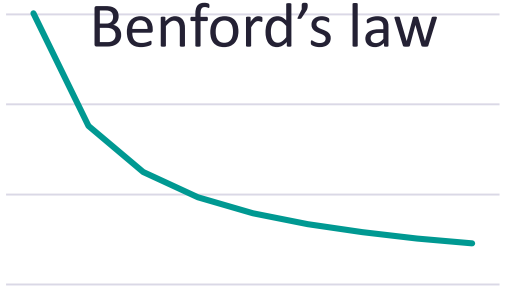


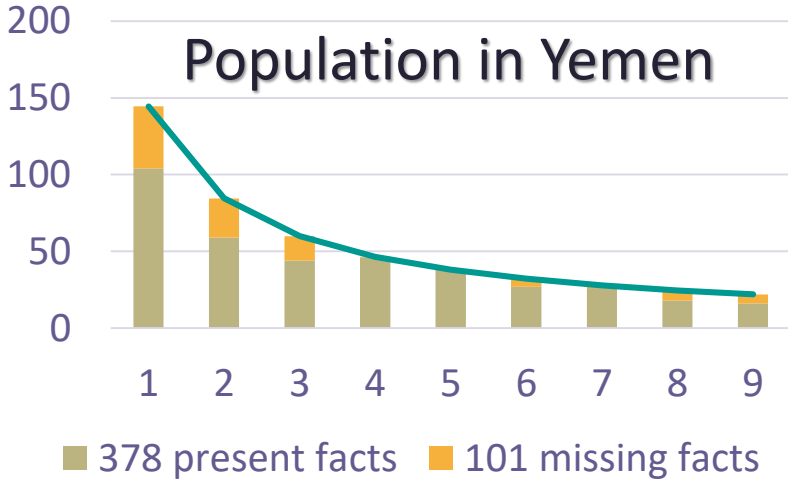
distribution of the fsd



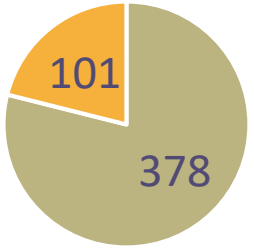
facts of r on $\mathcal{K}^*$

Benford's law





Representativeness:

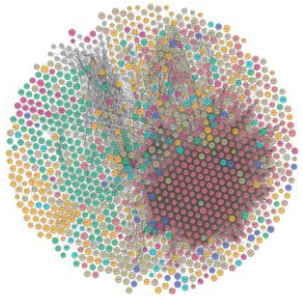


$$= \frac{378}{378 + 101} = 79\%$$

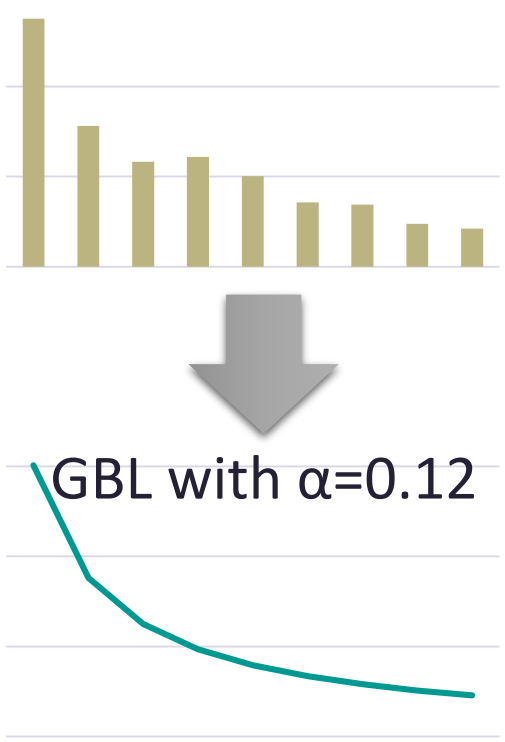
Computing the minimum number of facts for retrieving Benford's law

Our method in unsupervised context

facts of r on $\mathcal{K}$



distribution of the fsd



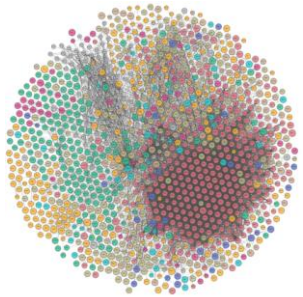
facts of r on $\mathcal{K}^*$

ideal distribution is unknown!

Learning the parameter α of the Generalized Benford's Law

Our method in unsupervised context

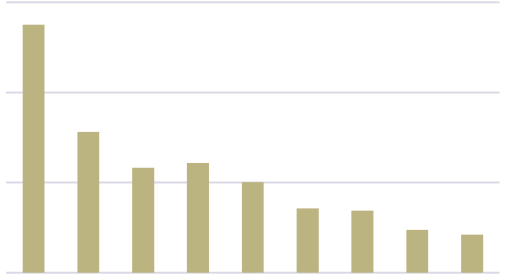
facts of r on $\mathcal{K}$



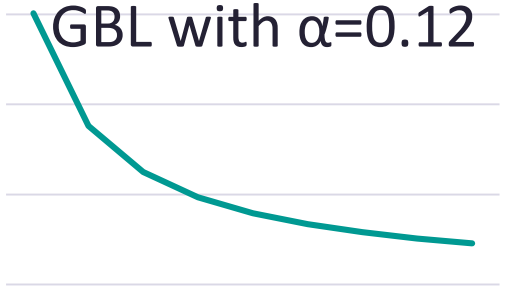
facts of r on $\mathcal{K}^*$

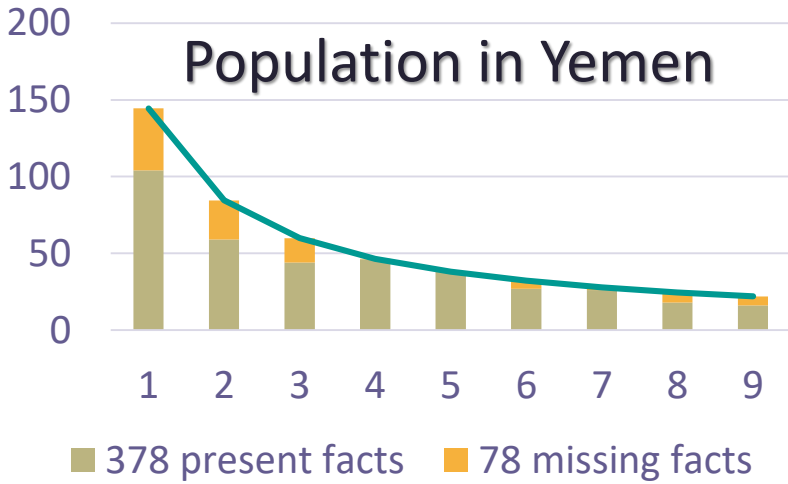
ideal distribution is unknown!

distribution of the fsd

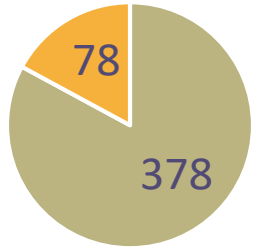


GBL with α=0.12





Representativeness:



$$= \frac{378}{378 + 78} = 82\%$$

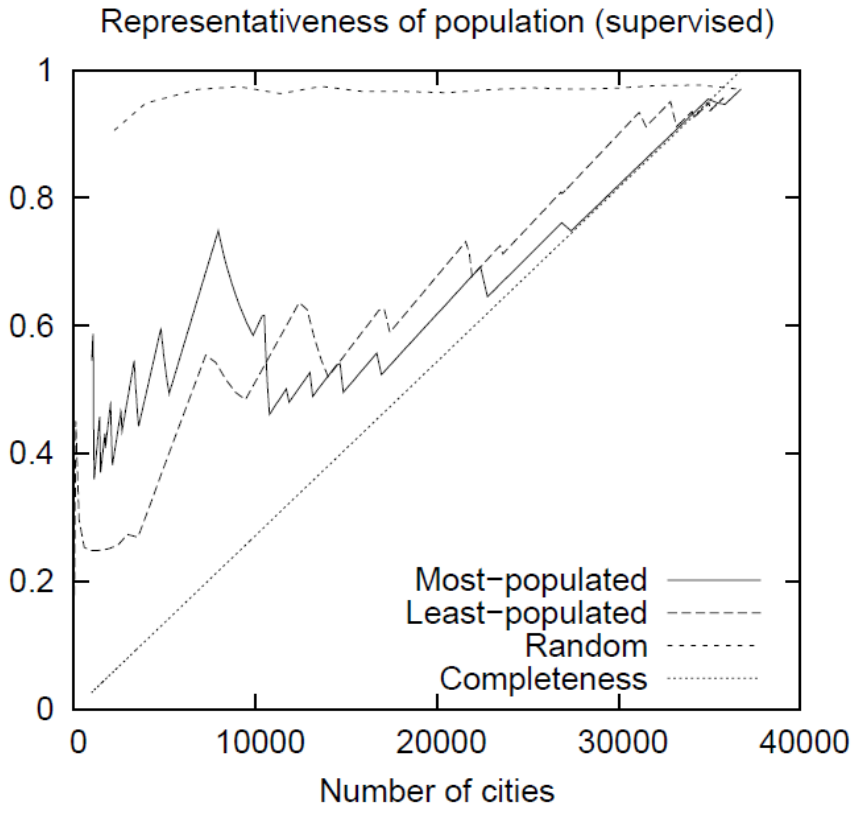
Computing the minimum number of facts for retrieving Benford's law

Experimental study

- **Evaluation protocol**
  1. Take a correct and complete relation as gold standard
  2. Degrade the completeness by discarding facts
  3. Approximate the representativeness

- **Gold standard:** population in French cities according to govt statistics
- **Degradation:**
  - Most-populated: remove the least populated cities
  - Least-populated: remove the most populated cities
  - Random: remove cities randomly

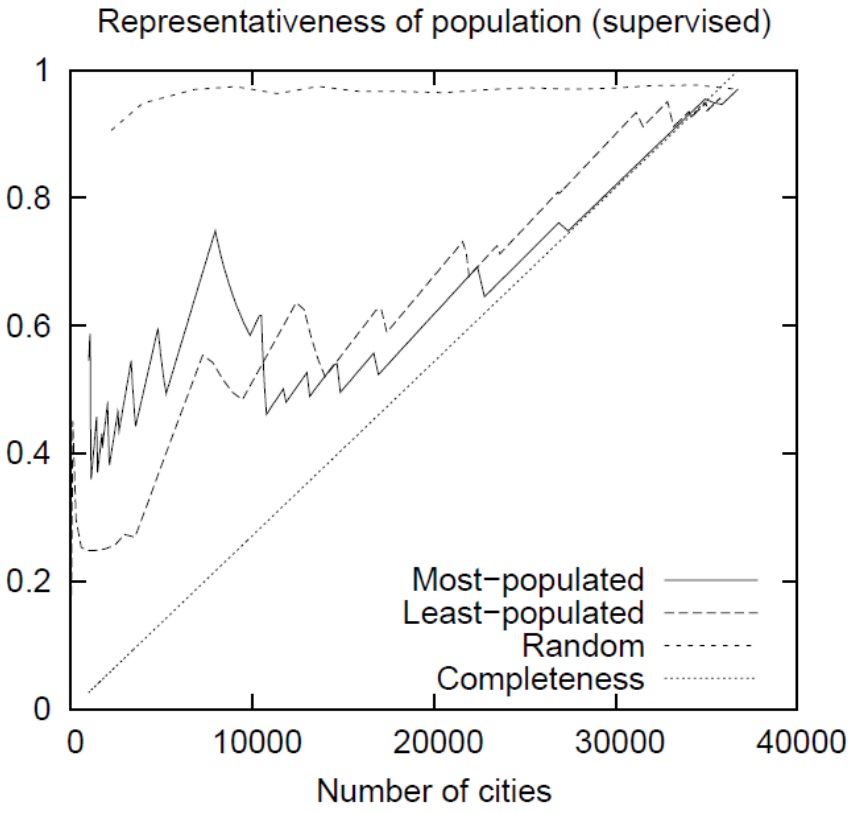
Population of French cities

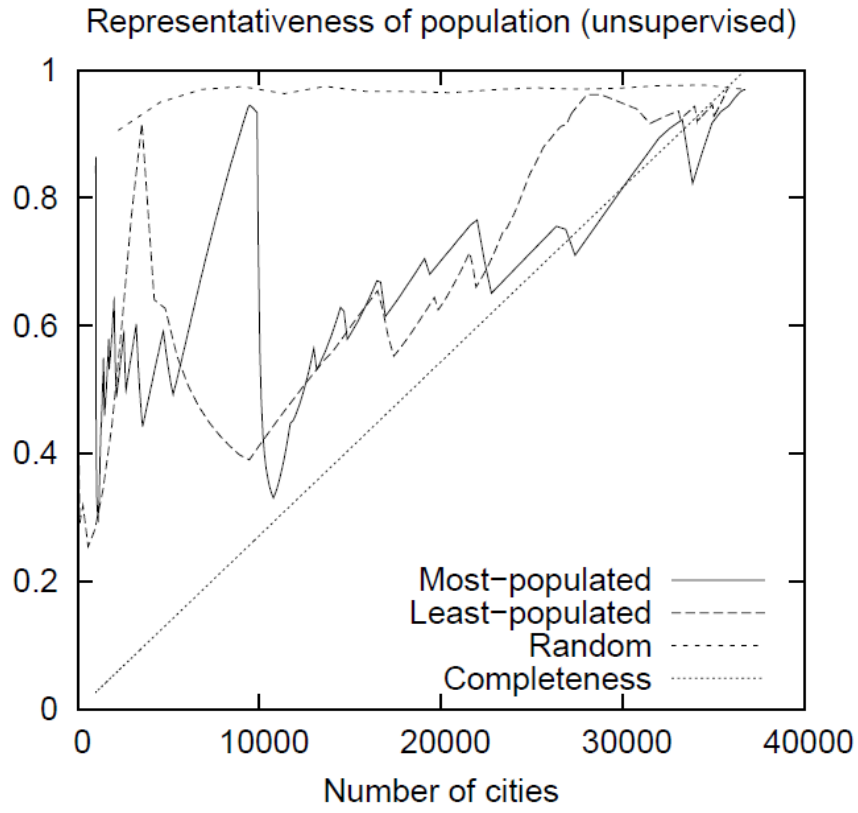


- Representativeness is an upper bound of completeness
- **Most/least-populated degradation:** tight bound if number of cities > 22k
- **Random degradation:** the representativeness is high

Representativeness approximates well the bias

Population of French cities



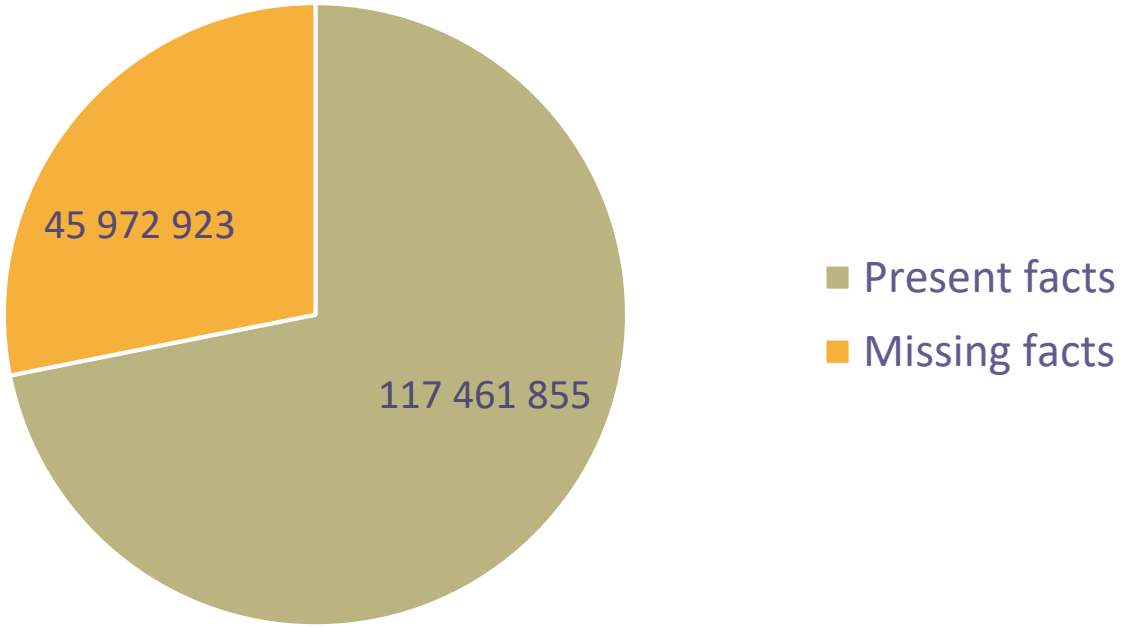


Learning the parameter α does not perturbate the approximation

Auditing DBpedia (France)

1,487 relations (out of 2,920) have a distribution statistically compliant with the GBL

Representativeness: 72%



Conclusion

- **The representativeness is more important than the completeness for achieving statistics.**
- **First use of Benford's law for approximating the proportion of missing data**
- **The approximate representativeness based on the GBL is an upper bound of the true representativeness and the true completeness.**

- **Future work:**
  - How to correct sparql queries with representativeness information?
  - How to scale up the approach to audit the LOD?