We are not alone! (at least, most of us). Homonymy in large scale social groups

Arthur Charpentier, Baptiste Coulmont

We are not alone ! (at least, most of us). Homonymy in large scale social groups

Arthur Charpentier*& Baptiste Coulmont†

December 6, 2017

1 First and last-names Homonyms

The Western system of identification is based on a first and a last-name : the first-name is a personal name, the last-name is a transmitted family name, often from the father to his children. According to Scott, Tehranian & Mathias (2002) this system is first of all a government device, to monitor individuals and to ensure the rights and duties of citizen : it surfaced with the emergence of state governments. Nowadays, the more stable the state, the stronger this system: it gives a legal civil identity to everyone under its scope.

The "first-name + last-name" couple is not, and never was, sufficient to identify someone without any ambiguity. Historians and anthropologists have often remarked that in small European villages, many individuals shared the same identity. In small settings where everyone was known to everyone, there was no "collective interest in the clear and unambiguous individuation of persons through their names" (De Piña-Cabral (2012)). In small villages, nicknames (Big John), toponyms (John from the lake) and paraphrases (the son of Jake) could be much more efficient to distinguish someone from everyone else.

If this system worked for a long period of time, it was thanks to local agents of the state who could translate a local identity (Big John) into the civil identity needed by the state or the central authorities (John Martin) and reassure the state that John Martin the conscript or John Martin the suspected tax evader was indeed Big John. With additional elements such as the precise date of birth, the place of birth, the names and profession of the parents... the first and last-names could be used to identify someone in a much larger regional or national setting (Noiriel (2001)).

And today in our "global village" the first and last-names are still the basis for worldwide identification. But without intimate knowledge or local agents in charge of the disambiguation, the collision of identities becomes problematic and more frequent. Every day in a random airport, someone sharing the identity of a known terrorist may be interrogated by customs agents or banned from flying. And every second, bibliographic databases are trying to differentiate John Lee the mathematician from John Lee the biologist in order to compute their scientific outputs (Gomide, Kling & Figueiredo (2017)).

Yesterday's homonymy was the shared sign of belonging to the same locality. There may have been hundreds of John Martins around 1700, but if they were not from the same place, they did not know they existed. Today's homonymy is shared between strangers in random places. In our interconnected societies, electronic social networks and multiple registrations enable us to "meet" or to "bump into" people with the same names as ours, often in circumstances when we have to assert a right (to vote, to travel, to buy...) based on our civil identity. From the point of view of the individual, then, homonymy is a random annoyance, a discomfort or a personal catastrophe, depending on the circumstances.

But from the point of view of the manager of any large scale register, today's homonymy seems to be a very common nuisance, if we consider the great numbers of personal identifiers that are meant to distinguish individuals without ambiguity. Personal identification numbers such as the Social Security Number in the United States, or the "*numéro d'inscription au répertoire des personnes physiques*" (NIR) in France were created to resolve this particular problem (Lévy (2000)).

These numbers are not used daily by people, who still prefer to be known by their names, and who do not gain anything by using a number instead. In the academic field, the "ORCID" promises to be "a persistent digital identifier that distinguishes you from every other researcher". It is meant to be used widely and the incentive is another promise : it "ensur[es] that your work is recognized".

But we do not know how frequent these identity collisions are. We do not know if, in a large scale society, many people have homonyms, or if only a small percentage does. An because of privacy issues, and of required anonymity of databases, estimating the proportion of homonyms in a large group is rather difficult. For instance in France, if researchers were allowed to access the NIR database, we could easily get the true proportion. But it is not the case.

*Université de Rennes 1
†Université Paris 8 and INED

In this article[1] a propose a technique to provide an estimation of the proportion of homonyms in large scale groups based, on the distribution of first-names and last-names in a (smaller) subset of these groups. The main result is that, in societies such as France or the United States, identity collisions (based on first + last-names) are frequent. The large majority of the population has at least one homonym.

2 A Birthday Paradox Problem

The birthday problem concerns the probability that, in a set of randomly chosen people, some pair of them will have the same birthday. It is usually seen as a paradox since, within a group of $n = 23$ persons, it is more likely to find than not to find two people sharing the same birthday, which seems counter-intuitive given the large number of possible birthdays and the small size of the group. Here also, we want to compute the probability that people share the same names (so called "homonymy") in a given group. The first difference between these two problems is that the number of possible birth dates is known (365 - if we exclude the 29th of February) while names belong to a much larger set. The second difference, in the computation of probabilities on birthdays, is that we usually assume that births are uniformly distributed over the year. But we cannot assume names to be uniformly distributed (and purely randomly chosen) : as suggested in Li (2012) for first-names, a Pareto/Zipf law can be considered, see Figure 1, for first and last-names distribution, in France. With Zipf's law the probability of occurrence is inversely proportional to the rank. Here we consider a mild version a power function (for instance the probability of occurrence can be inversely proportional to the square root of the rank). It is related to Pareto principle where a small number of names is accounting for a large proportion of the observations.

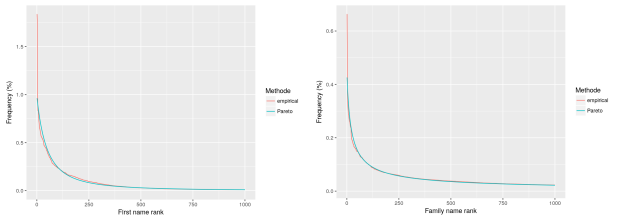


Figure 1: Empirical distribution of first (on the left) and last (on the right) names in France, with the estimate Pareto/Zipf fit.

For instance here there are more than 74,000 first-names in Paris and Marseille (1,757,895 persons on our lists). The top 10 (0.01% of all possible first-names) account for almost 9.5% of the population. And 740 first-names (1%) are shared 82% of the population. It is an extremely strong version of the Pareto principle (the popular 80/20 rule: 80% of the wealth is owned by 20% of the population). And out of 300,000 last-names, the top 1% is shared by 31% of the population. As a comparison, with a pure Zipf's law (when the probability of occurrence is proportional to the inverse of the rank) the top 1% of the first-names should be shared by 60% of the population. And for financial wealth (Bricker *et al.* (2017)) claims that in the U.S., in 2016, the richest 1% of families controlled 38.6% of the country's wealth.

Probit and Logit Transforms

Following ideas in demography, (Pearl & Reed (1922)) and (Winsor (1932)) introduced the *logistic function* in the 1920's (see (Cramer (2003)) for more details about the origin of the function and the term). This function yields the popular logistic regression, to model a probability as a (linear) function of some covariates, since it is naturally obtained when the logarithm of an odds-ratio is a linear function of the covariates, where $P = \mathbb{P}[Y = 1|X = x]$ satisfies

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = a+bx \text{ or } P = \frac{e^{a+bx}}{1+e^{a+bx}}$$

At the same time, (Bliss (1934)) suggested an alternative regression model, based on a latent representation. Assume that there is an (unobservable) variable $\tilde{Y}$, with $\tilde{Y} = a+bx+\varepsilon$, where ε is a Gaussian noise. Assume also that

$$Y = \begin{cases} 0 \text{ if } \tilde{Y} \leq 0 \\ 1 \text{ if } \tilde{Y} > 0 \end{cases}$$

Then

$$P = \Phi(a+bx) \text{ or } \Phi^{-1}(P) = a+bx,$$

where Φ is the cdf of the centered reduced $\mathcal{N}(0,1)$ distribution. Φ^{-1} was called the *probability unit transform*, or shortly the *probit transform*. Hence, statistician like to represent either the *logit* or the *probit* transformation of probabilities, hoping they will have a linear model for the later.

This non-uniformity makes our problem complex (see Munford (1977), DasGupta (2005), Inoue & Aki (2008), or Nunnikhoven (1992) for some attempts, and more recently Cortina Borja (2013)). Let $P_{n,k}$ denote the proportion of people having an homonym in a group of size n when k names are available. Before using real data, let us run simulation to visualize the evolution of the proportion $P_{n,k}$ as a function of n (or its logarithm). Let us consider first the case where names are uniformly (randomly) chosen among the k

[1] Additional material, including mathematical explanation and R codes used for computations (and to produce graphs) is available on a GitHub repository, https://github.com/freakonometrics/homonym

possible names (analogous to the birthday problem when $k = 365$). This case is on the left of Figure 2.

For different k's, different lines represent the evolution of $P_{n,k}$, which is (naturally) increasing with n : the larger the group, the more homonyms there should be. Since probabilities are constrained to lie in the interval $[0, 1]$, alternatives can be to visualize the odd-ratio or the probit transformation of those probabilities. If the graph on top show the evolution of $P_{n,k}$, the graph in the middle show the evolution of the logit transformation of that proportion (the logarithm of the odds ratio) in the middle, and the probit transformation below.

On the right of Figure 2 we can visualize the evolution of the proportion of homonyms, $P_{n,k}$, as a function of the group size n, when names have a Zipf law on the set of possible names (of size k). One can observe that either the logit or the probit transformation of proportion $P_{n,k}$ is linearly increasing in the logarithm of n. We will use that property to estimate the proportion of homonyms when n is too large (or when our sample size is too small).

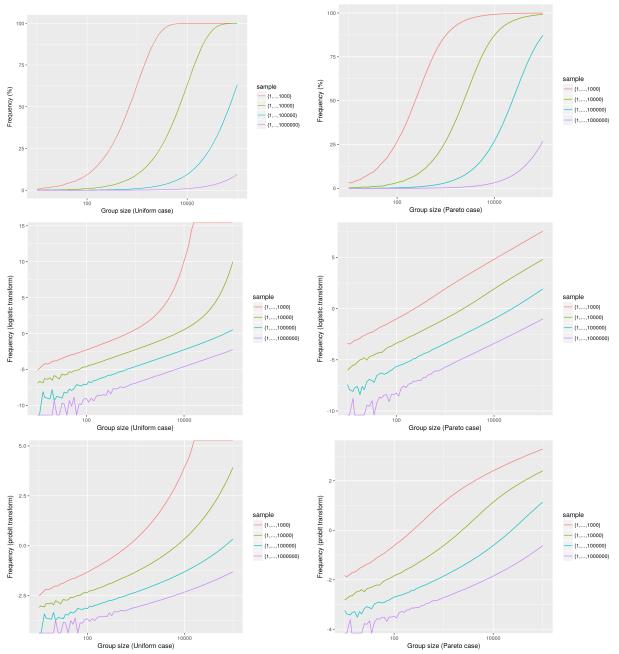


Figure 2: Evolution of $P_{n,k}$ as a function of the group size n (on a log scale) as a function of k, for different distributions $\boldsymbol{p}$ (uniform on the left and Pareto on the right).

3 First and last-names

In Höhle (2017), only the first-name was considered, but to study homonymy, we need to study pairs (first-name, last-name). We now face a bivariate problem, with - potentially - k_1 first-names and k_2 last-names. The two components cannot be considered as independent, see among recent references Chalabi & Flower (2014), which starts with the observation that "*Michael Smith*" might not be the most popular American names, even if "*Michael*" and "*Smith*" are probably the most popular first and last-names. Similarly on French data, "*Jean*"/"*Marie*" and "*Martin*" are the most popular first and last-names (for males and females respectively, on Paris and Marseille electoral lists) but the most popular full name is actually "*Thi Nguyen*", with 350 records (84 people in Paris and Marseille named "*Jean Martin*", the third highest score). To go one step further in the analysis, a chi-square test can be performed, as in Chalabi & Flower (2014). On Figure 3 are plotted (normalized) residuals from a chi-square test between first-names of the top 25, and last-names. For instance, "*Michel* " is both a popular first and last-names. But the proportion of "*Michel Michel* " is much smaller than was would be expected if the two were independent. One can also observe some strong 'communitarian' effect, with Vietnamese names ("*Than*" or "*Thi*" for the first-name and "*Nguyen*" or "*Tran*" for the last-name), African names (such as the popular Peul last-name "*Diallo*" and the first-name "*Mamadou*") or Jewish names ("*Cohen*" and "*Levy* " are positively correlated with "*David* ", but negatively correlated with "*Marie* "). Finally there are also negative correlations for names that share similar phonemes (such as "*Pierre Robert* " both ending with [ɛʁ], "*Alain Martin* " ending with [ɛ̃] and "*Laurent Durand* " [ɑ̃]).

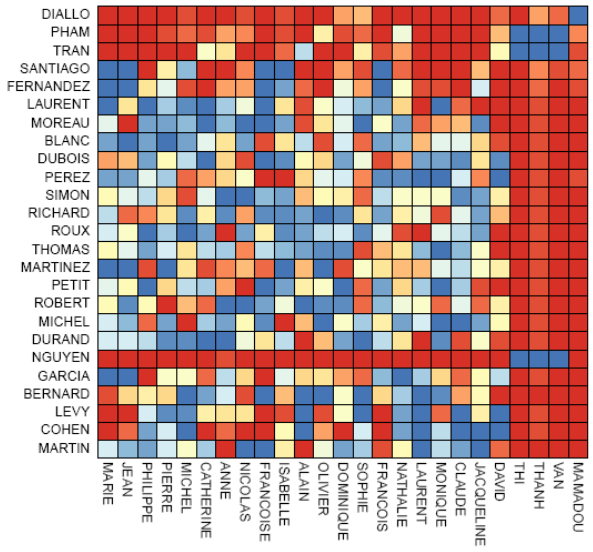


Figure 3: Pearson's residuals from a chi-square test of independence in the contingency table last vs. first-names, in France.

As explained in the introduction, our sample size is too small to estimate the proportion of people with an homononym in the country. So we will have to extrapolate P_n for n large, based on accurate estimations obtained when n is much smaller. And having a linear model to extrapolate is the most convenient

way to do it. As discussed in the previous section, having a linear relationship between the probit transform of P_n and $\log(n)$ would be interesting because that yields the popular least square problem : we fit a linear trend on observed data, and extrapolate for non-observed ones.

4 Application on French Data

In order to compute the probability P_n in the context of French names, the electoral roll of Paris and Marseille (for the year 2015) has been used. In this dataset, we have the first-name, last-name and date of birth of registered electors in Paris and Marseille (1,757,895 observations). Overall, we kept 1,542,528 observations, because of some typos in the original dataset. There where $k_1 = 74,085$ first-names in that dataset, and $k_2 = 309,907$ last-names (almost half of those appeared only once). Because of the variety of the first and last-names, our sample size ($n = 1.5$ million) was too small to estimate the proportion of people with an homonym in the entire French population (65 million). Resampling from pairs (first and last-names together) will over-estimate the proportion of homonyms in a very large group. Nevertheless, as mentioned in the previous section, it is not realistic to draw first and last-names independently, since both are correlated. On Figure 4 we can visualize the proportion of homonyms when drawing from the French population in Paris, either drawing pairs or drawing first and last-names independently.

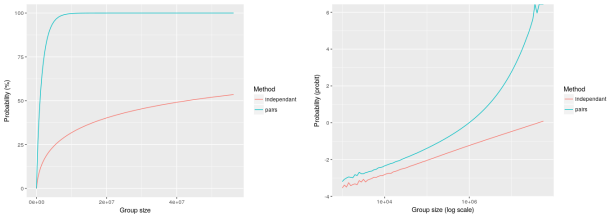


Figure 4: Proportion of homonyms when drawing from the French population in Paris, either drawing pairs (first and last-names) in blue, or drawing first and last-names independently, in red. Empirical probabilities P_n are on the left, and the probit transform of P_n is on the right.

When n is not too large, drawing pairs should yield a good approximation (the red line on Figure 4), but when n is too small we keep sampling from our too small sample, so we should over-estimate the proportion of people with an homonym. On the other hand, drawing independently first and last names will under-estimate the proportion of people with an homonym since it does not account for the dependence that exists between the two. Nevertheless, in that case, we can observe that the assumption of a linear relationship between the probit transform of the proportion and $\log(n)$ is valid, whatever the value of n. So, as we can see on Figure 5, we use a linear approximation when n is between 5,000 and 50,000, on approximations obtained drawing pairs, and then we extrapolate that linear approximation for large n's.

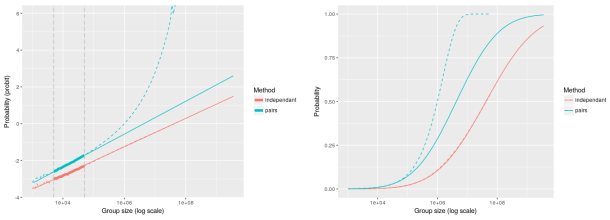


Figure 5: Proportion of homonyms with a linear extrapolation when pairs were drawn (linear on the *probit* transform as a function of $\log n$).

As we can see in the second box on another dataset in the United States, the approach we use here is rather general, and not only valid in France.

5 Temporal evolution of that Proportion

On two larger datasets[2], we can observe the evolution of first and last-names in France, see Tables 1 and 2 (those datasets contained statistics about first and last-names, respectively, but not paired).

time period	size	top 10	top 100
1916-1940	95,000	25.17%	79.05%
1941-1965	105,000	20.50%	72.61%
1966-1990	245,000	12.59%	56.98%

Table 1: first-names in France.

time period	size	top 10	top 100
1916-1940	638,000	1.83%	8.66%
1941-1965	669,000	1.76%	8.41%
1966-1990	814,000	1.57%	7.83%

Table 2: last-names in France.

The pattern observed in Table 1 for the first-names was already observed in (Li (2012)) (on U.S. first-names). Ninety years ago, most of the first-names were related to Saints or the Bible, so the list was much smaller (with a lot of "*Marie*", "*Jean*", "*David*"), while nowadays there are first-names from various place in the world (not to mention odd names that can now be given to babies nowadays). Old first-names remain, and new ones appear, but it is not the case for last-names. Last-names are more stable with

[2]The first one is the *fichier des prénoms, 2016 edition* available from https://www.data.gouv.fr/fr/datasets/fichier-des-prenoms-edition-2016/ produced by the National Institute of Statistics (INSEE) and the second one is the *fichiers des noms de famille - 1891-1990 - 1999 edition*, produced by INSEE, available from ADISP-CMH.

time, even if one could also object that there were less "*Nguyen*" in France in 1915 than in 2015. But in the case of Vietnamese names, even in Vietnam there are less names than in France (names are related to clans, and they are about 200 - 85% of Vietnamese people share 12 names as mentioned in (Krowolski & Nguyên (1999))). This might explain the (relative) stability observed on last-names.

It is then possible, assuming independence between first and last-names, to visualize the evolution of the proportion of homonyms, approximated using Monte Carlo simulations, on Figure 6, for groups of size 10,000 up to 200,0000 people (from bottom to top).

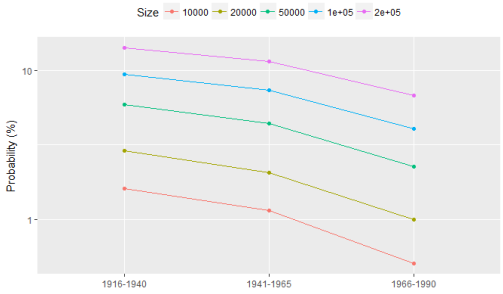
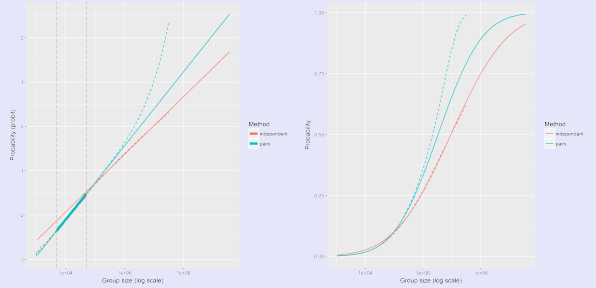


Figure 6: Evolution of the proportion of homonyms, $P_n^{\perp}$, assuming independence between first and last-names.

Application to Ohio Data

It is a dataset with 7.8 million individuals. Using independence between first and last-names to generate names, here also we obtain that the probit transform of the proportion of people with an homonym is linear in the logarithm of the population size. Further, observe that in that group, 50% of people have an homonym in that specific state. It might be interesting to extrapolate to a much higher n. As described in Figure below, in a population of $n = 320$ million people (the country size), we can estimate that 95.1% Americans have an homonym (in the United States).



6 Conclusion

As the interconnexion of our world increases and as the realm of interactions widens , we encounter a increasing number of homonyms. These collisions are annoying. But we continue to value the use of a basic identification system. Some contemporary changes reduce the chance of collisions: we increasingly choose rare names for our children, and, at least in Europe, the transmission of the father's last-name is slowly replaced by the possibility to choose to transmit the mother name or to create a combination of both parents' names.

7 Datasets

- Ohio Voter Files available at https://www6.sos.state.oh.us/ords/f?p=111:1
- Paris and Marseille Voter Files
- Fichier des prénoms, édition 2016, INSEE, available at https://www.data.gouv.fr/fr/datasets/fichier-des-prenoms-edition-2016/
- Fichiers des noms de famille - 1891-1990 - Édition 1999, INSEE [producteur], ADISP-CMH [diffuseur]

References

Bliss, C. (1934). The Method of Probits. *Science*, vol. 79-2037, 38–39.

Bricker, Jesse, Lisa J. Dettling, Alice Henriques, Joanne W. Hsu, Lindsay Jacobs, Kevin B. Moore, Sarah Pack, John Sabelhaus, Jeffrey Thompson & Richard A. Windle (2017) Changes in U.S. Family Finances from 2013 to 2016. *Federal Reserve Bulletin*, **103** (3).

Cramer, J.S. (2003). Logit models from economics and other fields. Cambridge University Press.

Chalabi, M. & Flowers, A. (2014). Dear Mona, What's The Most Common Name In America? http://53eig.ht/2yuhhbc

Chatterjee, S., Diaconis, P. & Meckes, E. (2004). Exchangeable pairs and Poisson approximation. *Electronic Encyclopedia of Probability.*

Cortina Borja, M. (2013). The strong birthday problem. *Significance*, **10**:6, 18–20.

DasGupta, A. (2005). The matching birthday and the strong birthday problem: a contemporary review. *Journal of Statistical Planning and Inference*, **130**, 377–389.

Eshel, A. (2013). On the Frequency Distribution of first-names. *Names: A Journal of Onomastics*, **49**, 55–60.

Gomide, Janaína, Hugo Kling, and Daniel Figueiredo (2017). Name Usage Pattern in the Synonym Ambiguity Problem in Bibliographic Data. *Scientometrics*: 1—20. doi:10.1007/s11192-017-2410-2

Höhle, M. (2017). Why the "most popular" baby names might not be the most popular. *Significance*, **14**:3, 30-33. doi: 10.1111/j.1740-9713.2017.01037.x

Inoue, K. & Aki, S. (2008). Methods for studying generalized and coupon collection problem. *Communications in Statistics - Simulation and Computation*, **37**, 844-862.

Krowolski, N. & Nguyên, T. (1999). Nom et appelation au Viêt-Nam. *in* D'un nom à l'autre en Asie du Sud-Est, Pauwels & Massard-Vincent Eds. Karthala.

Michel-Louis Lévy (2000), Le numéro INSEE : de la mobilisation clandestine (1940) au projet Safari (1974), *Dossiers & Recherche*, Paris (France) : Ined, no 86, septembre 2000, 23–34

Li, W. (2012). Analyses of baby name popularity distribution in U.S. for the last 131 years. *Complexity*, **18**:1, 44–50 .

Mase, S. (1992). Approximation to the birthday problem with unequal occurrence probabilities and application to the surname problem in Japan. *Annals of the Institute of Statistical Mathematics*, **44**:3, 479–499.

Munford, A.G. (1977). A note on the uniformity assumption in the birthday problem. *American Statistical*, **31**, 119.

Noiriel, Gérard (2001). The Identification of the Citizen: The Birth of Republican Civil Status in France. In *Documenting Individual Identity: The Development of State Practices in the Modern World*. Jane Caplan and John Torpey, eds. Pp. 28–48. Princeton (New Jersey): Princeton University Press.

Nunnikhoven, T.S. (1992). A birthday problem solution with non-uniform birth frequency. *American Statistical*, **46**, 270–274.

Pearl, R. & L. J. Reed (1922). A further note on the mathematical theory of population growth. *Proceedings of the National Academy of Sciences* **8**, 365–368.

De Piña-Cabral, João (2012). The Functional Fallacy: On the Supposed Dangers of Name Repetition. *History and Anthropology* **23**:1, 17-–36.

Scott, James C, John Tehranian, and Jeremy Mathias (2002). The Production of Legal Identities Proper to States: The Case of the Permanent Family Surname. *Comparative Studies in Society and History*, **44**:1, 4-–44.

Winsor, C. P. (1932). A comparison of certain symmetrical growth curves. *Proceeding of Washington Academy of Sciences*, **22**, 73–84.