Last name analysis of mobility, gender imbalance, and nepotism across academic systems

Jacopo Grilli[a,1] and Stefano Allesina[a,b,c,1,2]

[a]Department of Ecology & Evolution, University of Chicago, Chicago, IL 60637; [b]Computation Institute, University of Chicago, Chicago, IL 60637; and [c]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved June 1, 2017 (received for review March 1, 2017)

In biology, last names have been used as proxy for genetic relatedness in pioneering studies of neutral theory and human migrations. More recently, analyzing the last name distribution of Italian academics has raised the suspicion of nepotism, with faculty hiring their relatives for academic posts. Here, we analyze three large datasets containing the last names of all academics in Italy, researchers from France, and those working at top public institutions in the United States. Through simple randomizations, we show that the US academic system is geographically well-mixed, whereas Italian academics tend to work in their native region. By contrasting maiden and married names, we can detect academic couples in France. Finally, we detect the signature of nepotism in the Italian system, with a declining trend. The claim that our tests detect nepotism as opposed to other effects is supported by the fact that we obtain different results for the researchers hired after 2010, when an antinepotism law was in effect.

academic systems | isonomy | gender imbalance | nepotism

. . . [S]tat rosa pristina nomine, nomina nuda tenemus.

Umberto Eco, *The Name of the Rose*

Since its inception, science has been a worldwide endeavor, with scholarly publications and conferences connecting researchers across the globe. Despite the many similarities (for example, the organization of scholars into departments and the ubiquitous academic ranks), academic systems around the world are, however, quite distinct in their goals and practices. In many European countries, for example, professors are civil servants, and therefore, their hiring procedures are subject to special regulations. In contrast, American universities have more freedom in choosing their faculty. Salaries, duties, and resources also vary widely both within and between systems.

Here, we examine differences in academic systems using a very simple form of data: a list of names of professors working at a given institution along with their rank, field of study, and geographic location. These data are easy to obtain and can be used to unveil patterns in mobility and immigration (are researchers employed in the region where they were born and raised?), gender imbalance (are women underrepresented in certain fields?), and even nepotism (do professors hire their relatives for academic posts?).

The use of last names as a form of data has a long history in biology, starting with George Darwin (son of Charles), who used the distribution of last names in England to estimate the prevalence of marriages by first cousins (like his parents) (1). Soon dubbed the "poor's man population genetics" (2), the study of isonymies (occurrences of people with the same name) provided a cheap source of (large) data, with the advantage that last names would well-approximate neutral alleles (2, 3), allowing for the study of human migrations (4). With the advent of modern molecular methods, last names have been associated with Y-chromosome haplotypes (5). More recently, the association of ethnic-specific first and last names has been shown to be predictive of occupational success (6). Closer to the spirit of this work, the distribution of last names in Italian academics has been used to test the hypothesis of nepotistic hires (7, 8): these studies have highlighted a significant scarcity of last names in certain fields and regions, raising the suspicion of nepotistic hires, in which professors recruit relatives for academic positions.

Here, we expand on these results by presenting an international comparison and by introducing specific randomizations that probe different aspects of each academic system. Although our focus is on academia, the same approach could be used in a variety of contexts [for example, in studies of social mobility (9) or health disparities (10)] and even to test whether longevity is related to inbreeding (11).

We analyze last names in three datasets of unprecedented quality and size: all Italian academics in four different years (2000, 2005, 2010, and 2015), researchers currently working at the CNRS in France, and academics working at research-intensive public institutions in the United States. These datasets allow us to track the evolution of last names in time (Italy) and the geographic variability both within and between countries. Special features of the data allow us to detect the presence of academic couples in France and probe the effects of antinepotism legislation in Italy.

Results show that the Italian academic system tends to attract researchers mostly at the local level—many researchers have last names that are typical of the region or even the city in which they work—whereas the American system is geographically well-mixed, with a strong influence of immigration. Moreover, in the United States, certain last names are typical of specific scientific fields—meaning that immigration and researchers of given ethnic/cultural backgrounds tend to target preponderantly specific

Significance

In the age of Big Data and high-throughput sequencing, a list of names might seem like a meager source of data. However, here we show that, by analyzing last name distributions, one can highlight distinctive patterns in academic systems around the world. By collecting data on academics in Italy, France, and the United States, we show that, in the Italian system, professors tend to work in their native region, whereas the US system is geographically well-mixed. We can detect the effect of field-specific immigration in the United States and highlight patterns of gender imbalance in the sciences. Finally, we show that, in Italy, the plague of nepotism—professors hiring their relatives—is slowly declining.

Author contributions: J.G. and S.A. designed research; J.G. and S.A. performed research; J.G. and S.A. analyzed data; and S.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data and the code needed to generate the results are publicly available on GitHub at github.com/StefanoAllesina/namepairs.

[1]J.G. and S.A. contributed equally to this work.

[2]To who correspondence should be addressed. Email: sallesina@uchicago.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1703513114/-/DCSupplemental.

areas of research. Using the distribution of first names, we show strong gender imbalance in science, technology, engineering, and mathematics (STEM) disciplines in all systems. Finally, we show that nepotism is present (but declining) in Italy.

Results

Data. We collected four datasets for the Italian academic system, including the names of all professors holding permanent (or since they were introduced in 2010, temporary "tenure-track") positions along with their institution, academic field (area; 14 coarse-grained fields), rank (which we coarse-grained into assistant, associate, and full professor), and gender. We enriched the data by adding a city and region to each record. The number of professors is 52,004 for year 2000, 60,288 for 2005, 58,692 for 2010, and 54,102 for 2015.

For France, we collected the names, unit, and region for all of the researchers affiliated with CNRS (Chercheurs CNRS) or working at a mixed CNRS–university research unit (Chercheurs non-CNRS). Each unit is associated with a scientific field and a location. Whenever available, we stored the self-reported maiden names. The database contains 44,860 researchers.

For the United States, we collected from state records the names of professors at selected R1 institutions (research universities—highest research activity according to the Carnegie Classification of Institutions of Higher Education). We collected data on 38 institutions, privileging the states in which more than one R1 operate. Because the data do not contain a disciplinary field, we associated professors with a discipline using the Scopus database. We were able to successfully match 36,308 professors in this way.

Details on data collection and processing are reported in *SI Appendix*. The data are publicly available.

Isonymous Pairs. Each researcher is associated with an institution and field. Two researchers with the same last name working at the same institution and in the same field form an isonymous pair (IP). As a shorthand, we define the "department" d as the set of all researchers working in a certain field at a given institution. For each last name i, n_{id} measures how many researchers with that name work in department d. The number of IPs in a given department is $p_d = \sum_i \binom{n_{id}}{2}$. For example, if in department d, we find three researchers whose last name is Hopper and four called Pollock, we have that $p_d = 3 + 6 = 9$ IPs. This measure can be interpreted as the number of edges connecting researchers with the same name in a network where the nodes are the researchers working in the same department (*SI Appendix*, Fig. S1), and it has excellent statistical properties compared with other quantities (*SI Appendix*, Fig. S2).

Given that each department belongs to a geographic region and a discipline, we can sum the number of IPs by region ($p_r = \sum_{d \in r} p_d$) or field ($p_f = \sum_{d \in f} p_d$). Using randomizations, we probe whether the observed p_r (or p_f) is significantly different from what we would expect at random.

Three Randomizations. For each dataset, we calculate p_r and p_f for each region and field. We then repeatedly randomize the data in three different ways, each time recording the values of p_r and p_f for the randomized data. In this way, we obtain an approximate P value measuring the probability of finding a number of IPs greater than or equal to what was observed empirically in a given region or field. Importantly, each randomization provides us with a different angle to probe the data, unveiling distinctive patterns of mobility and immigration.

In the first randomization (by nation), we simply shuffle 10^6 times the last names in the database, each time tracking p_r and p_f. This randomization tells us whether the empirical data contain more IPs at the regional or field level than we would expect when resampling all academics at random.

In the second randomization (by city), we shuffle the last names of academics within each city. That is, for each department, we assign researchers at random from those working in the same city. As such, names that are common at the city level but rare nationwide (reflecting, for example, geographic, linguistic, or cultural barriers) will be sampled with high probability, increasing the expected number of IPs.

In the third randomization (by field), last names are shuffled within field. This procedure allows us to test the existence of field-specific names (for instance, as a consequence of immigration targeting a specific field). For example, a recent National Science Foundation survey (12) found that, of 5.2 million immigrant scientists and engineers in the United States, 57% were born in Asia and that immigrants targeted disproportionately computer science, mathematics, and engineering.

Randomizing by nation, we find that, in all systems, at least a few sectors (Fig. 1) and regions (Fig. 2) have a significant excess of IPs (with stronger deviations in Italy and France).

This excess of IPs could be caused by region-specific distributions of last names, in which case the difference between local and national distributions would drive the results. Randomizing by city, we observe a large drop in the ratio between observed and expected IPs in Italy and France (i.e., blue vs. red bars in Fig. 1), meaning that, in these systems, the excess of IPs for many fields and regions is likely due to the geographic

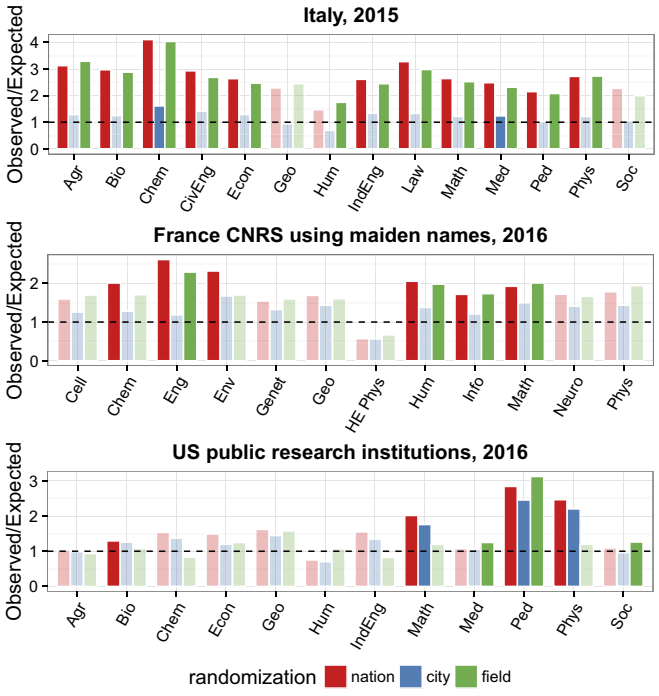

Fig. 1. Ratio between observed and expected numbers of IPs for each academic system and field. Different colors stand for three randomizations explained in the text; saturated colors mark fields in which the probability of finding a higher or equal number of IPs by chance is $\leq$ 0.05 per number of fields (i.e., significant after applying a Bonferroni correction for multiple hypothesis testing). Agr, agriculture; Bio, biological sciences; Cell, cell and molecular biology; Chem, chemistry and pharmaceutical sciences; CivEng, civil engineering and architecture; Econ, economics and statistics; Eng, engineering; Env, environmental sciences; Genet, genetics; Geo, geology and Earth sciences; HE Phys, high-energy physics; Hum, philology, literature, archeology; IndEng, industrial, electronic, and electric engineering; Info, information and communications sciences; Law, law; Math, mathematics and computer science; Med, medical sciences; Neuro, neuroscience; Ped, pedagogy, psychology, history, philosophy; Phys, physics and astrophysics; Soc, social and political sciences.

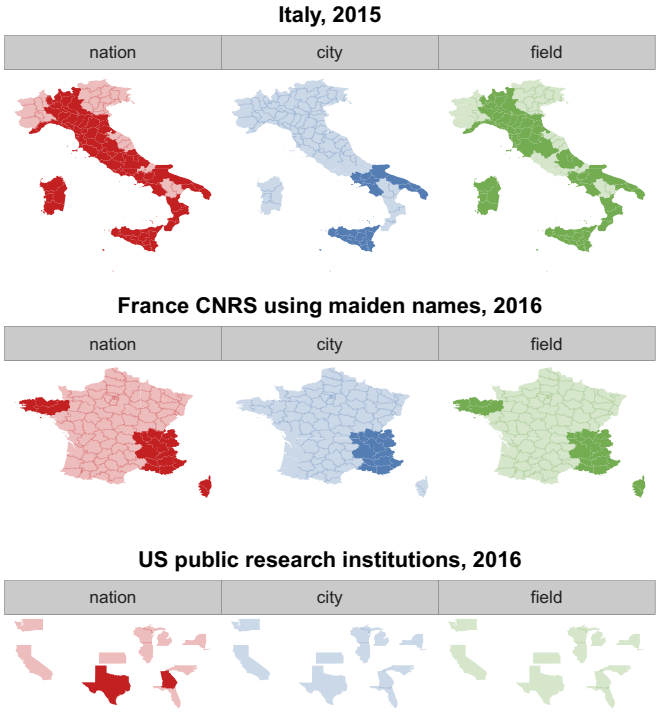

Fig. 2. The same randomizations as in Fig. 1 but summing IPs by region. Saturated colors stand for significantly higher numbers of IPs than expected at random (i.e., P value ≤ 0.05 per number of regions).

distribution of last names (i.e., the national pool of names is much more diverse than the local one). For Italy, this hypothesis is confirmed by plotting the similarity between last name distributions and geographic distance (*SI Appendix*, Fig. S3). The second randomization yields no significant results for fields in France, whereas two fields test significantly in Italy, and three fields test significantly in the United States. Three regions test significantly in Italy (Campania, Puglia, and Sicilia), and two regions test significantly in France (Provence-Alpes-Côte d'Azur and Rhône-Alpes). No state in the United States yields significant results. Note that, in the US academic system, accounting for regional names has very little effect compared to Italy and France. Therefore, the regional distribution of last names is not much different from the national one: there are no last names that are typical of a state or city.

The fact that physics and mathematics yield significant results in the United States suggests that the explanation for the excess IPs could be found analyzing immigration. For example, in our US dataset, the name Zhang is the most common in chemistry and mathematics and the 3rd most common in agriculture, geology, and physics but only the 41st most common name in sociology and the 115th most common name in humanities. Smith, however, is among the top three names in humanities, sociology, medicine, and agriculture but only the 20th in chemistry and the 47th in geology. Randomizing by field, we observe a large decline in the ratio between observed and expected IPs for mathematics and physics, whereas for fields in which immigration is less preponderant (pedagogy, medicine, and sociology), the effect is reversed. Note that, in Italy and France, randomizations by field yield about the same results as those at the national level, meaning that immigration is either very scarce or evenly distributed among fields.

Academic Couples. In Italy, women keep their maiden name when they marry—in our datasets, spouses have distinct last names. For the French dataset, whenever provided, we used self-reported maiden names (nom de jeune fille) for the analysis to compare the results with the Italian ones more directly. In the United States, more and more frequently, women are retaining their maiden names—especially women holding advanced degrees (13). However, given that changing one's name was customary until recently and that maiden names are not reported, we cannot measure how much of an effect married couples have on the results.

We can, however, experiment with the French dataset to see whether we can detect the fact that many married couples work in the same department. In the dataset, 2,933 women list different maiden and married names. We can "force" them to assume their husband's name: in case of double-barrel last names, we "subtract" the maiden name to obtain the husband's name (e.g., Magritte-Duchamp, listing Duchamp as maiden name, would yield Magritte); when the married name does not contain the maiden name, it is assumed to be the husband's name. Having modified the data in this way, we rerun the analysis, finding that now all fields and many regions become significantly enriched in IPs (Fig. 3). Thus, accounting for married couples sharing the same name produces highly significant results, meaning that our method can highlight genuine family ties when they are present.

First Names and Gender Imbalance. Repeating the same types of randomization for first names instead of last names shows that, in certain fields, there are more couples sharing the same first names working in the same department than expected (Fig. 4 and *SI Appendix*, Fig. S7). This fact, used to criticize previous studies (14), has, however, a very simple explanation (15): women are underrepresented in certain scientific areas as shown plotting the ratio between observed and expected IPs vs. the proportion of women for each field (Fig. 4). Note that, accordingly, randomizing by city has little effect, whereas randomizing by field considerably lowers the ratio in fields where women are scarce (e.g., industrial engineering and physics) and increases the ratio in those where women are more represented (humanities, pedagogy, and biology). In a way, the effect is similar to that of immigration but with women playing the role of immigrants.

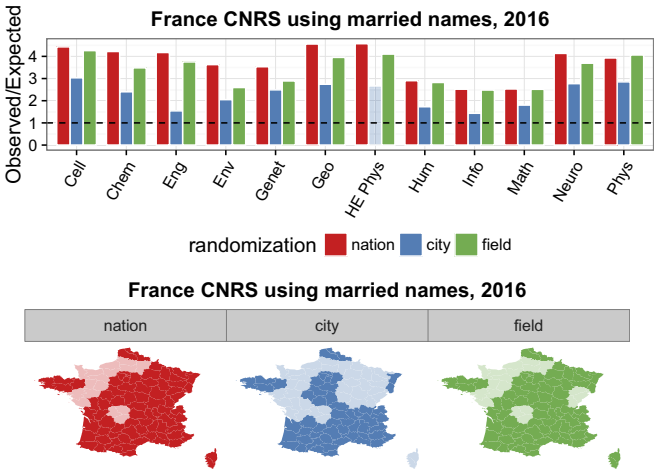

Fig. 3. The same as in Figs. 1 and 2 but using married names instead of maiden names. The large difference in the results is caused by married couples working in the same department. Saturated colors mark significant results once accounted for multiple hypothesis testing. Cell, cell and molecular biology; Chem, chemistry and pharmaceutical sciences; Eng, engineering; Env, environmental sciences; Genet, genetics; Geo, geology and Earth sciences; HE Phys, high-energy physics; Hum, philology, literature, archeology; Info, information and communications sciences; Math, mathematics and computer science; Neuro, neuroscience; Phys, physics and astrophysics.

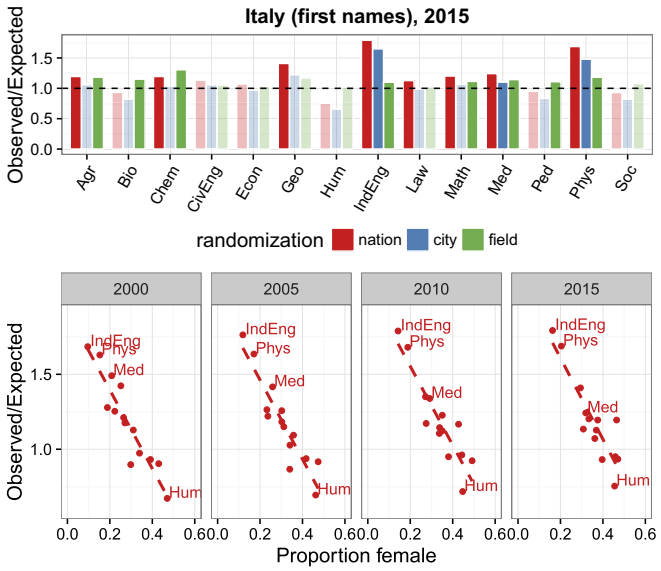

Fig. 4. (*Upper*) The same as in Fig. 1 but using first names instead of last names. (*Lower*) Ratio between observed and expected number of IPs vs. proportion of women for all years (national randomization). Some of the fields are highlighted for reference. Saturated colors mark significant results once accounted for multiple hypothesis testing. Agr, agriculture; Bio, biological sciences; Chem, chemistry and pharmaceutical sciences; CivEng, civil engineering and architecture; Econ, economics and statistics; Geo, geology and Earth sciences; Hum, philology, literature, archeology; IndEng, industrial, electronic, and electric engineering; Law, law; Math, mathematics and computer science; Med, medical sciences; Ped, pedagogy, psychology, history, philosophy; Phys, physics and astrophysics; Soc, social and political sciences.

One caveat on the analysis of first names is that, contrary to last names, first names can fluctuate widely from year to year, sometimes following specific events (16). For example, in *SI Appendix*, Fig. S6, we show that the frequency of newborns named Francesco (the most common first name among Italian boys born in the last decade) increased of about 40% after the election of Pope Francis. Because of these idiosyncratic trends, researchers of the same age would be more likely to share first names than those of different ages—a problem that is absent in the study of last names.

Time Evolution. For the Italian system, we have collected four snapshots between 2000 and 2015 in intervals of 5 years. We can, therefore, repeat the randomizations for all datasets and track the evolution of the system in time. Earlier years yield a higher number of significant results, with one-half of the fields testing significantly (randomization by city) in 2000 and 2005; there were five significant fields in 2010 and only two significant fields in 2015 (Fig. 5). The results by region follow a similar pattern (*SI Appendix*, Figs. S8 and S9).

Is Italian Academia Nepotistic? As shown above, the geographic distribution of last names as well as field-specific immigration can greatly affect the number of IPs within departments. In Italy, even when accounting for these factors, we do observe significant results. Previous studies (7, 8) have suggested that the excess IPs observed in Italian academia could be caused by nepotistic hires, with fathers hiring their children and siblings for academic posts (mothers hiring their children would be undetectable, because they would have different last names). Although proving this hypothesis would require access to data on actual family ties, which are not available, in this section, we present four statistical tests probing whether our results are compatible with the hypothesis of nepotism. All tests have the same structure. First, a category is assigned to all of the researchers (e.g., academic rank, gender, hired, or retired). Second, IPs for a certain combination of categories are computed (e.g., IPs of the type male–female or retired–not retired). Third, the categories are repeatedly scrambled within each department to estimate a P value.

For example, if the excess number of IPs was caused by nepotism, we would expect many of the pairs of isonyms within the same department to have different ranks because of the age difference between fathers and children. We thus measure the number of IPs of the kind full professor↔not full professor and compute the probability of observing a higher or equal number of IPs of this kind when shuffling the ranks within departments. In all four Italian datasets, we find a significant excess of IPs of this type (P value <0.01 for all years, computed out of 10^4 randomizations).

Similarly, given that last names are inherited by line of father, in the case of nepotistic hires, we would expect an excess of male↔male IPs (or equivalently, fewer IPs involving a woman). Measuring the number of IPs of this kind, we find that, in all cases, the number of male↔male IPs is higher than expected by chance, with 2 years yielding significant results (2005: P value $<$ 0.01; 2010: P value $<$ 0.03) and two differences that are not significant (2000: P value = 0.13; 2015: P value = 0.07).

If nepotistic hires were orchestrated by senior faculty members, we would expect retirees to be more likely to share names with the remaining faculty than expected by chance. Take two consecutive periods (for example, 2000 and 2005). Some names appear in the 2000 database but do not appear in the 2005 database: these faculty members have retired or exited the system in the meantime—we mark these as "retired." All of those who did not retire are marked as "remained." Measuring the number of IPs of the type retired↔remained and computing the probability of observing a larger or equal number of IPs of this kind when shuffling the labels retired/remained within each department, we see that, in all years, the number of IPs of this type is significantly higher than expected (2000 and 2005: P value $<$ 0.01; 2010: P value $<$ 0.02).

Similarly, we can find new hires for the years 2005–2015 and test whether new hires are more or less likely to share names with the professors already in the system. This test is interesting, because a "natural experiment" was carried out during these years: the Italian law 240 of 2010, which reformed the university system, included a provision (article 18) preventing departments from hiring relatives of their faculty, with the explicit intent of curbing nepotism. Our results show that the effects of this law can be detected in the data. Measuring the number of IPs of the kind hired↔already present, we find that, in 2005 and 2010,

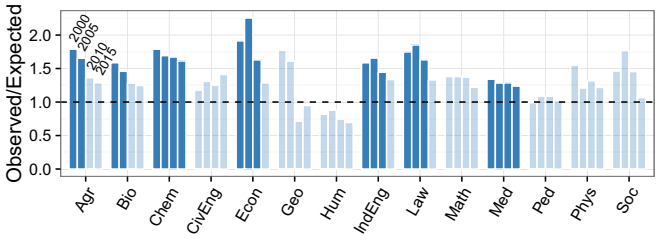

Fig. 5. Evolution of the ratio between observed and expected number of IPs in Italy between 2000 and 2015. Saturated colors mark significant results once accounted for multiple hypothesis testing. Agr, agriculture; Bio, biological sciences; Chem, chemistry and pharmaceutical sciences; CivEng, civil engineering and architecture; Econ, economics and statistics; Geo, geology and Earth sciences; Hum, philology, literature, archeology; IndEng, industrial, electronic, and electric engineering; Law, law; Math, mathematics and computer science; Med, medical sciences; Ped, pedagogy, psychology, history, philosophy; Phys, physics and astrophysics; Soc, social and political sciences.

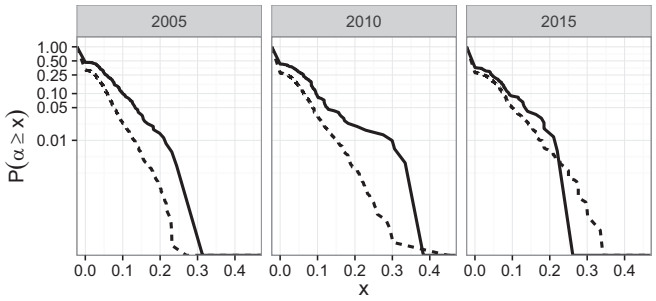

Fig. 6. Cumulative distribution of the maximum likelihood estimates $\hat{\alpha}$. For each department, $\hat{\alpha}$ is the maximum likelihood estimate of the probability of sampling new hires from the names already present in the department as opposed to the rest of the city. The solid lines show the distribution of the data, whereas the dashed lines are obtained repeatedly by randomizing the last names of all new hires in a 5-year period. For example, for the hires between 2000 and 2005, we find that 10% of the departments yield a $\hat{\alpha} \geq 0.1$, whereas in the randomizations, we find that only 2.4% of the departments should have such elevated values of $\hat{\alpha}$.

the observed value is not significantly smaller than expected by chance (2005: P value = 0.29; 2010: P value = 0.13), but the faculty members hired between 2010 and 2015 are less likely to share names with those already in the system than expected by chance (P value = 0.04).

A Model for Nepotism. Given that our results are consistent with the hypothesis of nepotistic hires, we attempt to quantify the phenomenon using a simple statistical model. Suppose a department d has to decide on a new hire: with probability α, they pick among the relatives of their faculty; with probability $1-\alpha$, they pick from the general population. Under this model, the probability of picking name j would be $q_j = \alpha\pi_j^{(d)} + (1-\alpha)\pi_j^{(c)}$, where $\pi_j^{(d)}$ is the proportion of professors with name j in the department, and $\pi_j^{(c)}$ is the proportion of professors with name j in the general population (that we estimated as the frequency in the city excluding the department). We want to find the maximum likelihood estimate of α for each department d and year. Large values of the maximum likelihood $\hat{\alpha}$ mean that departments tend to hire disproportionately faculty whose name is already present in the department, whereas low values mean that departments tend to pick names from the city at random. Details of the model are in *Materials and Methods*.

Computing $\hat{\alpha}$ for all departments that hired more than 10 faculty members for a given period (to have more accurate estimates) and recomputing this value after scrambling the last names of all new hires in each city, we find differences between researchers hired before 2010 and those hired under the new law (Fig. 6). For the faculty hired between 2000 and 2005 and those hired between 2005 and 2010, the distribution of $\hat{\alpha}$ is significantly different from what was expected (Kolmogorov–Smirnov test: 2005 $D_n = 0.19$, P value ≤ 0.001; 2010 $D_n = 0.16$, P value ≤ 0.001). For the hires between 2010 and 2015 (when the new antinepotism law was in effect), the distributions of $\hat{\alpha}$ are pulled closer together, yielding nonsignificant differences (Kolmogorov–Smirnov test: $D_n = 0.087$, P value = 0.083).

Discussion

Here, we have taken an ostensibly meager source of data—a list of names of professors along with their field of research and geographic information—and used elementary randomizations to investigate differences in academic systems. Importantly, we produced a specific randomization for each angle that we wanted to probe, showing that even extremely simple methods can shed light on subtle patterns in the data.

In Italy, names cluster by city (*SI Appendix*, Fig. S3), showing that professors tend to work where they were born. The American system, however, is geographically well-mixed (*SI Appendix*, Fig. S5). The strong signal of immigration is highlighted by the US randomizations, where, for example, physics and mathematics test significantly when randomizing by city but not when randomizing by field: certain names are associated with specific fields, consistent with field-specific immigration and the fact that American researchers of certain heritages tend to target preponderantly science and engineering.

The analysis of married vs. maiden names for the French system shows that our methods can detect the signal of family ties when they are present. Note that, in the Italian system, all women keep their maiden name, whereas in the United States, an unspecified fraction of married women takes their husbands' names—possibly explaining the excess of IPs in pedagogy and other fields. The analysis of first names highlights strong gender imbalance in STEM fields.

Even when accounting for geographical and field-specific distribution of last names, Italian academics display an excess of last name sharing within departments. The results of our additional analysis are consistent with the hypothesis of nepotism as testified by the fact that we can detect the effects of an antinepotism law in effect for the period 2010–2015. Importantly, our analysis shows that nepotism is field- and region-specific and likely driven by a handful of departments. For example, when measuring $\hat{\alpha}$ for the hires in 2005, we found that 10% of departments had an $\hat{\alpha} \geq 0.1$ (we would expect 2.4% at random), whereas the vast majority of departments had $\hat{\alpha} \approx 0$. Similarly, the randomizations in Figs. 1, 2 and 5 show that specific regions and fields drive the results.

For the Italian system, evidence of the efficacy of antinepotism laws and the fact that the phenomenon seems to be declining should be greeted as good news, with two caveats. First, the decrease in IPs is largely because of retirements: we showed that retirees are more likely to share last names than new hires. Moreover, after a large increase in the number of faculty between 2000 and 2005, the size of Italian academia has been steadily declining, with a staggering 10% overall loss during the last decade. The numbers look even worse when examined at the level of regions, fields, or single institutions (*SI Appendix*): Toscana and Liguria lost one-quarter of their faculty (Siena, −30.2%; Florence, −29.3%; Genoa, −24.3%), and geology (−21.4%) and the humanities (−18.9%) have lost a large fraction of their professors. Solving the problem of nepotism by disbanding the university system would be throwing the baby out with the bathwater. Second, antinepotism laws can have negative side effects, especially when targeting spousal hires. For example, in the first half of the 20th century, antinepotism laws in the United States created the phenomenon of the "vanishing wives" (17): because spouses could not be hired in the same department as their husbands, many women worked as unpaid guests, slowing down the process leading to equal gender representation.

The examples of France, which has hiring procedures that are quite close to those of the Italian system, and the United States, where practices are, however, very different, show that one can build a fair academic system without the need for especially harsh measures. Indeed, many US institutions welcome couples (spousal hires; often extended to domestic partners), although antinepotism provisions are in place, so that one partner cannot be responsible for the other partner's career advancements.

Materials and Methods

Data. The data were collected from publicly available websites, checked for quality, and organized as detailed in *SI Appendix*. After collection, the data were anonymized by using a numeric identifier for each last name. The data and the code needed to generate the results are publicly available at github.com/StefanoAllesina/namepairs.

Randomizations. In all datasets, for each researcher, we have information on first name, last name, institution, and field of study as well a geographic information (city and region). The department is obtained by combining institution and field. For each department, we count the number of pairs of researchers with the same last name (IPs). We then sum the IPs by region or field. The three randomizations are obtained by (*i*) randomizing all last names (randomized by nation), (*ii*) randomizing last names within each city (by city), and (*iii*) randomizing last names within each field (by field).

Modeling Nepotism. We consider a simple mixture model, in which the probability of choosing to hire a researcher with name j in the department d and city c is $q_j = \alpha\pi_j^{(d)} + (1-\alpha)\pi_j^{(c)}$, where $\pi_j^{(d)}$ is the frequency of name j in the department, and $\pi_j^{(c)}$ is the frequency in the general population from which the new researcher is sampled. The parameter α can be interpreted as the probability of a nepotistic hire. For instance, in a perfectly nonnepotistic system, α would be equal to zero, and all of the last names of new hires are random samples from the general population. Note that, even if $\alpha = 0$, it is possible to hire a person with a last name that is already present in the department. The parameter α quantifies, therefore, the probability of a nepotistic hire using the excess of IPs.

For a given period (e.g., 2000–2005), we compile a list of all new hires for each department, obtaining m_j^d: the number of people hired in department d with last name j. Under our model, the probability of observing a set of new hires with last names $\{m^d\}$ is given by the multinomial distribution

$$P(\{m^d\}) = \frac{\left(\sum_j m_j^d\right)!}{\prod_j m_j^d!} \prod_j \left(q_j^{(d)}\right)^{m_j^d}. \quad [1]$$

The maximum likelihood estimate $\hat{\alpha}$ can be found by maximizing this quantity. By taking the logarithm of the likelihood and neglecting the terms independent of α, one obtains

$$\sum_j m_j^d \log\left(\alpha\pi_j^{(d)} + (1-\alpha)\pi_j^{(c)}\right). \quad [2]$$

We determined $\pi_j^{(d)}$ as the frequency of last name j in department d at the beginning of the period (e.g., in 2000 if the period 2000–2005 is considered) and $\pi_j^{(c)}$ as the frequency in the city (removing the department) at the beginning of the period. One special case that needs to considered is that in which the name of the new hire is not present in the department or the city, in which case $\pi_j^{(d)} = \pi_j^{(c)} = 0$. In such cases, we postulate that the name is present in the city at an unknown (low) frequency. This assumption is quite convenient, because for $\pi_j^{(d)} = 0$, the exact value of $\pi_j^{(c)}$ does not impact the maximum likelihood estimate of α. In fact, the term $\log((1-\alpha)\pi_j^{(c)})$ appearing in Eq. **2** in the case of $\pi_j^{(d)} = 0$ can be written as $\log(1-\alpha) + \log\pi_j^{(c)}$, and therefore, the second additive term does not impact the maximum likelihood estimate.

Note that, given the finiteness of the data, a maximum likelihood estimate $\hat{\alpha} > 0$ could be a consequence of fluctuations and not nepotistic hires. To assess the importance of these fluctuations, we compared the maximum likelihood estimate of $\hat{\alpha}$ with the one obtained by randomizing the names of new hires within a department.

ACKNOWLEDGMENTS. We thank M. J. Michalska-Smith for comments. Data were provided by Scopus.com. J.G. was supported by the Human Frontier Science Program; S.A. was supported by National Science Foundation Grant DEB 1148867.

1. Darwin GH (1875) Marriages between first cousins in England and their effects. *J Stat Soc Lond* 38:153–184.
2. Crow JF (1983) Surnames as markers of inbreeding and migration. *Discuss Hum Biol* 55:383–397.
3. Zei G, Guglielmino CR, Siri E, Moroni A, Cavalli-Sforza LL (1983) Surnames as neutral alleles: Observations in Sardinia. *Hum Biol* 55:357–365.
4. Piazza A, Rendine S, Zei G, Moroni A, Cavalli-Sforza LL (1987) Migration rates of human populations from surname distributions. *Nature* 329:714–716.
5. Jobling MA (2001) In the name of the father: Surnames and genetics. *Trends Genet* 17:353–357.
6. Goldstein JR, Stecklov G (2016) From Patrick to John F.: Ethnic names and occupational success in the last Era of mass migration. *Am Sociol Rev* 81:85–106.
7. Allesina S (2011) Measuring nepotism through shared last names: The case of Italian academia. *PLoS One* 6:e21160.
8. Durante R, Labartino G, Perotti R (2011) *Academic Dynasties: Decentralization, Civic Capital and Familism in Italian Universities*. Working paper 17572 (National Bureau of Economic Research, Cambridge, MA). Available at www.nber.org/papers/w17572. Accessed May 22, 2017.
9. Clark G, Cummins N, Hao Y, Vidal DD (2015) Surnames: A new source for the history of social mobility. *Explor Econ Hist* 55:3–24.
10. Elliott MN, et al. (2009) Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcome Res Methodol* 9: 69–83.
11. Montesanto A, Passarino G, Senatore A, Carotenuto L, De Benedictis G (2008) Spatial analysis and surname analysis: Complementary tools for shedding light on human longevity patterns. *Ann Hum Genet* 72:253–260.
12. Lan F, Hale K, Rivers E (2015) *Immigrants' Growing Presence in the U.S. Science and Engineering Workforce: Education and Employment Characteristics in 2013* (NSF InfoBriefs, Arlington, VA), pp 15–328.
13. Goldin C, Shim M (2004) Making a name: Women's surnames at marriage and beyond. *J Econ Perspect* 18:143–160.
14. Ferlazzo F, Sdoia S (2012) Measuring nepotism through shared last names: Are we really moving from opinions to facts? *PLoS One* 7:e43574.
15. Allesina S (2012) Measuring nepotism through shared last names: Response to Ferlazzo and Sdoia. arXiv:1208.5792.
16. Kessler DA, Maruvka YE, Ouren J, Shnerb NM (2012) You name it–how memory and delay govern first name dynamics. *PLoS One* 7:e38790.
17. Lykknes A, Opitz DL, Van Tiggelen B, eds (2012) *For Better or for Worse? Collaborative Couples in the Sciences* (Springer Science & Business Media, Basel), Vol 44.

Last name analysis of mobility, gender imbalance, and nepotism across academic systems Supporting Information

Jacopo Grilli[a] **and Stefano Allesina**[a,b,c]

[a]Department of Ecology & Evolution, University of Chicago, 1101 E. 57th Chicago, IL 60637, USA.; [b]Computation Institute, University of Chicago.; [c]Northwestern Institute on Complex Systems, Northwestern University.

S1. Data

For our analysis, we collected a large database containing information on university professors working in Italy, France, and the United States of America.

Italy The data were downloaded from the website `cercauniversita.cineca.it`, maintained by the Italian Ministry of Education, University and Research, in September 2016. The website provides data on all Italian university professors from year 2000 onward. For the years 2000, 2005, 2010, and 2015, we downloaded data on all the disciplinary fields (*Area*—a coarse-grained division into 14 fields). The number of professors in the database ranged from 52,004 (year 2000) to 60,288 (year 2010). Data include the first and last name, the institution, information on the disciplinary field, rank, and gender of all Italian professors. We enriched the data by adding a region and city to every institution. In case of institutions with multiple campuses, we chose the main one. All names were transliterated into ASCII, and made into lowercase, for better handling of accents and apostrophes. Finally, we numbered all last names and first names, and used this anonymized version of the data for our analysis (the same was done for all data sets).

The labels in the figures and tables refer to the following disciplinary fields (Area): `Agr`, agriculture and veterinary sciences (07 Scienze agrarie e veterinarie); `Bio`, biological sciences (05 Scienze biologiche); `Chem`, chemistry and pharmaceutical sciences (03 Scienze chimiche); `CivEng`, civil engineering and architecture (08 Ingegneria civile e architettura); `Econ`, economics and statistics (13 Scienze economiche e statistiche); `Geo`, geology and Earth sciences (04 Scienze della terra); `Hum`, philology, literature, archeology (10 Scienze dell'antichità, filologico-letterarie e storico artistiche); `IndEng`, industrial, electronic, and electric engineering (09 Ingegneria industriale e dell'informazione); `Law`, law (12 Scienze giuridiche); `Math`, mathematics and computer science (01 Scienze matematiche e informatiche); `Med`, medical sciences (06 Scienze mediche); `Ped`, pedagogy, psychology, history, philosophy (11 Scienze storiche, filosofiche, pedagogiche, psicologiche); `Phys`, physics and astrophysics (02 Scienze fisiche); `Soc`, social and political sciences (14 Scienze politiche e sociali).

France The data were downloaded from the official website of the CNRS `web-ast.dsi.cnrs.fr/l3c/owa/annuaire.recherche` in September 2016. Queries targeted each and every one of the research units (by specifying a *Code unité*). The list of personnel of each of the *Unité mixte de recherche* and *Unité propre de recherche* was downloaded. Units operating principally outside of continental France and Corse were excluded (e.g., units working in Martinique and Guyane). For each unit, we assigned a field by selecting the most represented *Groupe(s) de discipline*. For example, a unit listing *SC - Chimie (70%)* and *SDE - Sciences de l'Environnement (30%)* would be assigned to Chemistry. In case of ties, we pick the first listed field. Similarly, for assigning the region and city to each unit, in case of multiple listings we chose the first city and region. Among the personnel, we extracted the names of all *Chercheurs CNRS* as well as *Chercheurs non CNRS* (i.e., researchers working in a CNRS laboratory, but officially affiliated with another research institution). Last names are listed in all capitals, while first name(s) are capitalized. We separated the two using regular expressions, and transliterated the strings to ASCII. Maiden names were gathered by matching the records with those obtained searching the website `annuaire.cnrs.fr/l3c/owa/personnel.frame_recherche`.

The labels in the figures refer to the following disciplinary fields (*Groupe de discipline*): `Math`, Mathematics (MATH Mathématiques); `Phys`, Physics (PHY Physique); `HE Phys`, High Energy Physics (PNHE - Physique Nucléaire et des Hautes Energies); `Chem`, Chemistry (SC - Chimie); `Env`, Environmental Sciences (SDE - Sciences de l'Environnement); `Cell`, Cell and Molecular Biology (SDV1 - Biologie cellulaire et moléculaire); `Neuro`, Neuroscience (SDV2 - Biologie intégrative et neurosciences); `Genet`, Genetics (SDV3 - Génétique); `Hum`, Humanities and Social Sciences (SHS - Sciences de l'Homme et de la Société); `Eng`, Engineering (SPI - Sciences pour l'Ingénieur); `Geo`, Earth Science and Astronomy (SPU - Sciences de la Planète et Univers); `Info`, Information and Communications Sciences (STIC - Sciences et Technologies de l'Information et de la Communication).

United States For institutions in the US, a ready-made database of all professors does not exist. We therefore took a different route, and downloaded data on public salaries. Many states list the salaries of all the state employees, often including university personnel. We searched for this type of data, privileging the states in which more than one R1 operates (to have multiple institutions within a region/state). Table S1 lists the institutions, state, and website from which the information was downloaded. For each institution, we downloaded the most recent year available (typically, 2015).

State	Institution(s)	Website(s)
California	UC Berkeley; UC Davis; UC Irvine; UC Los Angeles; UC Riverside; UC San Diego; UC Santa Barbara; UC Santa Cruz	ucannualwage.ucop.edu
Florida	F International U; F State U; U Central F; UF	floridahasarighttoknow.myflorida.com
Georgia	G Institute Technology; G State U; UG Athens	open.georgia.gov/sta/search.aud
Illinois	UI Chicago; UI Urbana-Champaign	salarysearch.ibhe.org/search.aspx
Michigan	M State U; UM Ann Arbor; Wayne State U	www.msusalaries.info (M State U); www.umsalary.info (UM Ann Arbor); www.waynestatesalaries.info (Wayne State U)
New York	City U NY; State U NY (SUNY) Albany; SUNY Buffalo State College; SUNY Stony Brook	seethroughny.net/payrolls
North Carolina	NC State U; UNC Chapel Hill	www.newsobserver.com/news/databases/public-salaries/
Texas	T A & M U; U Houston; U North T; UT Arlington; UT Austin; UT Dallas	salaries.texastribune.org/agencies/university
Washington	UW; W State U	fiscal.wa.gov/WaStEmployeeHistSalary.txt
Wisconsin	UW Madison; UW Milwaukee	host.madison.com/ (search for University Salaries)

Table S1. List of US R1 Public Universities considered in this study, along of the website from which the list of professors was downloaded.

We then filtered all the data in order to select only Assistant, Associate or Full professors. Adjunct and Visiting professors, as well as research assistants and associates were removed. Note that, contrary to the case of Italy and France, the data does not contain a disciplinary field. We therefore attempted matching researchers and fields by searching for their last name, first name (and when available, middle initials or middle name) in Scopus. For each researcher, a pairing was considered valid if it matched the institution, the last name, and the first name (initials were also matched when available).

For each researcher, Scopus returns the number of articles published in a given "subject-area". We took the most represented subject-area for each researcher and coarse grained into the 12 labels displayed in the figures and tables. In particular, the mapping between the Scopus subject-areas* and our labels is as follows: `Agr`, AGRI; `Bio`, BIOC, ENVI, MULT, NEUR; `Chem`, CENG, CHEM, ENER; `Econ`, BUSI, DECI, ECON; `Geo`, EART; `Hum`, ARTS; `IndEng`, ENGI; `Math`, COMP, MATH; `Med`, DENT, HEAL, IMMU, MEDI, NURS, PHAR; `Ped`, PSYC; `Phys`, MATE, PHYS; `Soc`, SOCI.

Data availability All the data are available at `http://github.com/StefanoAllesina/namepairs` for download in anonymized form: the first and last names have been replaced with numerical identifiers. All the analysis presented in this study have been performed on this database.

S2. Isonymous pairs

As explained in the main text, throughout the article we use the number of isonymous pairs (IPs) as our main observable. For a given institution and scientific field, we take the *department* to be the set of professors working in that institution and field. For a department d, the number of professors having last name i is n_{id}. The number of isonymous pairs is therefore $p_d = \sum_i \binom{n_{id}}{2}$. This quantity can be interpreted as the number of edges in a graph in which the nodes are the researchers working in the department, and edges connect researchers with the same last name (Figure S1).

We chose this observable because it has excellent statistical properties. In particular, take a list of names (for example, all the researchers working in Sardinia), and randomly extract a sample of k researchers without replacement. Then, the expected number of IPs in the set is approximately $p\binom{k}{2}$, where p is the proportion of isonymous pairs in the list of names:

$$p = \frac{\sum_i \binom{n_i}{2}}{\binom{\sum_i n_i}{2}} = \frac{\sum_i n_i(n_i-1)}{\left(\sum_j n_j\right)\left(\sum_j n_j - 1\right)} \qquad [S1]$$

Figure S2 shows that the variance around this expectation is very modest, guaranteeing that we can detect small but significant deviations, and that the number of IPs has better statistical properties than other measures—for example the number of unique last names in the sample. The choice of measuring IPs in this way differentiates our work from previous attempts at measuring the level of familism in academia. For example, Allesina (1) counted the number of distinct names in each discipline, while Durante *et al.* (2) constructed two indices of "homonymy" by counting how many members of a department have one (or more) namesakes as colleagues.

* For a list of all subject-areas, see `api.elsevier.com/documentation/search/SCOPUSSearchTips.htm`.

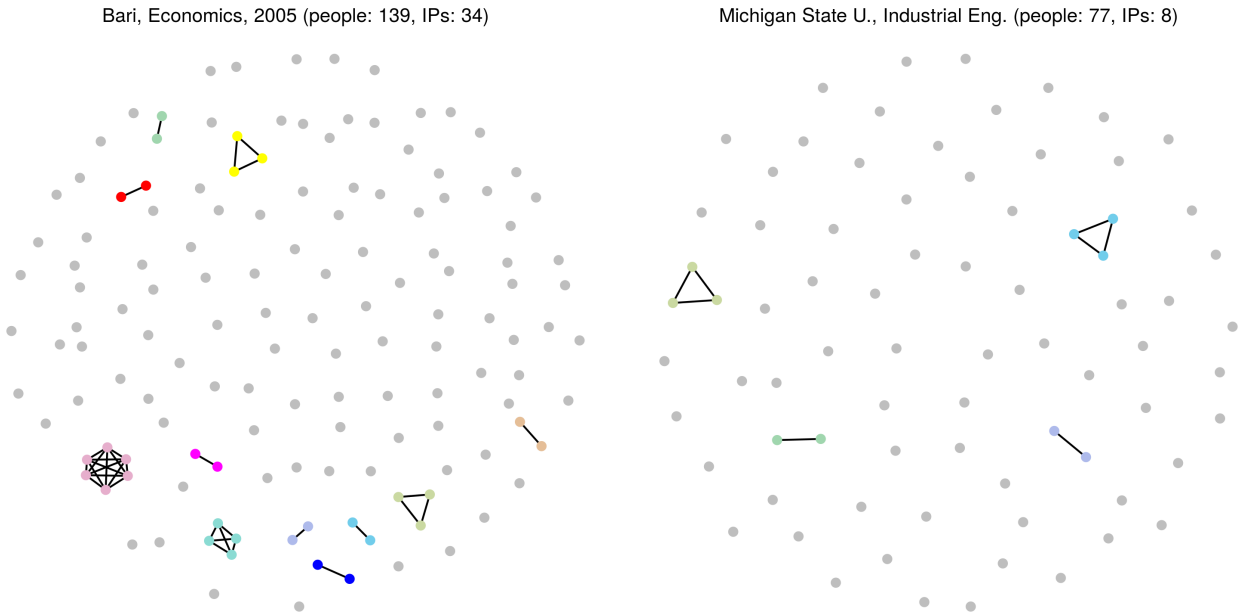

Fig. S1. Number of isonymous pairs as the number of edges in a graph. We can take all the researches in a department (left, University of Bari, Economics, 2005; right, Michigan State University, Industrial Engineering), and connect any two researchers that have the same last name. The total number of edges in the graph is the number of IPs, p_d. In the figure, researchers with unique last names are colored in gray.

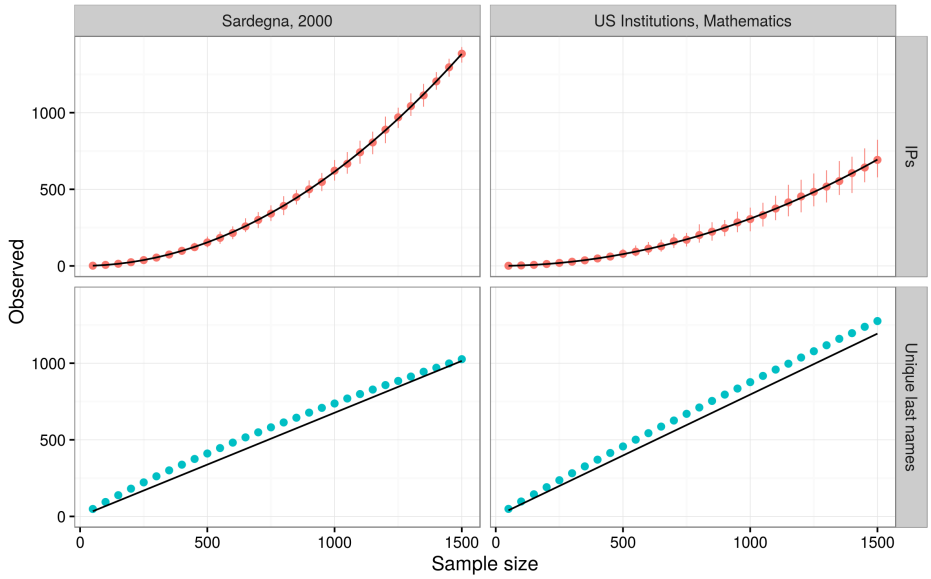

Fig. S2. Statistical properties of IPs. We took two data sets: the researchers working in Sardegna in 2000 (1580 researchers) and the mathematicians working at the US public institutions we included in the analysis (3259 people). For each sample size k (from 50 to 1500 in steps of 50), we sampled k researchers at random and computed the number of IPs (red), or the number of unique last names in the sample (blue). We repeated the sampling 100 times for each value of k and data set (the error bars mark the 5^{th} and 95^{th} percentile of the distribution). The black line shows the expectation: while the number of IPs scales with $\binom{k}{2}$, the number of unique last names in the sample scales in a non-trivial, non-linear way with k.

S3. Geographical distance and similarity

As discussed in the main text, Fig. 1 suggests the existence of a clustering by city of Italian last names, which is instead not present for US institutions. In this section, we quantify the similarity between the last names distribution of two cities and correlate this quantity with their geographical distance. The results support the conclusion of the main text, showing that the similarity between the last names of two Italian cities is generally negatively correlated with distance and that this pattern is not present for cities in the US.

Let n_{ic} be the number of people with last name i in city c and $n_c = \sum_i n_{ic}$ the number of people in city c. The fraction of people with last name i in city c is then $\pi_i^{(c)}$ defined as

$$\pi_i^{(c)} = \frac{n_{ic}}{n_c} \ . \qquad [S2]$$

We define the similarity between city c and city c' as

$$S_{c|c'} = \frac{\sum_i \pi_i^{(c)} \pi_i^{(c')}}{\sum_i \left(\pi_i^{(c')}\right)^2} \ . \qquad [S3]$$

This choice is motivated as follows. The numerator is (proportional to) the covariance between $\pi_i^{(c)}$ and $\pi_i^{(c')}$. The covariance between these frequency has been used, under the name of kinship, to compare last names of different locations (3). The covariance is, on the other hand, strongly influenced by the sample size of the two different cities. Our strategy is to normalize the covariance to obtain a quantity $S_{c|c'}$ that is independent of the sample size of city c'.

Assuming that the last names of city c were sampled from the same distribution of city c', one would expect to have, on average under this null hypothesis, $n_{ic} = n_c \pi_i^{(c')}$. The expected value of the covariance is therefore (proportional to) $\sum_i \left(\pi_i^{(c')}\right)^2$. Dividing the covariance by its expected values under the null hypothesis, one obtains Eq. S3. Under this definition, the similarity $S_{c|c'}$ is not symmetric and does not depend on the sample size of city c' (while it depends on the size of city c). Figures S3, S4 and S5 show, for each city c, the similarity between c and another city c' vs. the distance between c and c'.

Italian last names S3 are characterized by a strong geographical pattern. For many cities, the similarity clearly decreases with distance, indicating a geographical signal of last names in the Italian universities. In France and in the US this pattern is not present. This could be caused by different factors. One possibility is that the typical length-scale of spatial correlation of last names is smaller than the resolution at which we are observing the system. Alternatively, the last names of the whole population could lack a geographical signal (i.e., names are not associated to a specific geographical location). Finally, the absence of a relation between similarity and distance could indicate that the university system is effectively well-mixed and researcher move within the nation (or that their movement is not determined by geographical distance from their place of birth). Immigration from abroad does also play an important role in reducing the geographical signal.

S4. Last names by region and sector

Tables S2, S3 report the results for the three randomizations when summing IPs by field or by region for the Italian data set (2015). Tables S4 and S5 for the CNRS data, and Tables S7 and S6 for the US institutions. The tables correspond to Figures 1 and 2 in the main text.

field	observed	by country	by city	by field
Agr	86	**27.5 (5.8)** $p < 0.001$	66.8 (9.7) $p = 0.033$	**26.1 (5)** $p < 0.001$
Bio	125	**42 (7.2)** $p < 0.001$	100.2 (11.8) $p = 0.025$	**43.4 (6.8)** $p < 0.001$
Chem	70	**17 (4.4)** $p < 0.001$	**43.4 (7.4)** $p = 0.001$	**17.3 (4.2)** $p < 0.001$
CivEng	104	**35.5 (7)** $p < 0.001$	73.7 (11.3) $p = 0.011$	**38.7 (6.6)** $p < 0.001$
Econ	91	**34.5 (6.5)** $p < 0.001$	70.7 (9.7) $p = 0.026$	**36.8 (6.2)** $p < 0.001$
Geo	5	2.2 (1.5) $p = 0.075$	5.3 (2.4) $p = 0.596$	2 (1.4) $p = 0.054$
Hum	61	41.4 (7.2) $p = 0.007$	87.7 (10.6) $p = 0.997$	**34.9 (5.9)** $p < 0.001$
IndEng	205	**78.4 (10.9)** $p < 0.001$	153.4 (17.1) $p = 0.005$	**83.8 (10)** $p < 0.001$
Law	102	**31.1 (6)** $p < 0.001$	76.6 (9.8) $p = 0.009$	**34.2 (5.9)** $p < 0.001$
Math	46	**17.4 (4.5)** $p < 0.001$	37.6 (6.8) $p = 0.125$	**18.2 (4.3)** $p < 0.001$
Med	579	**232.9 (20.5)** $p < 0.001$	**467.7 (28.8)** $p < 0.001$	**249.9 (17.7)** $p < 0.001$
Ped	75	**34.9 (6.6)** $p < 0.001$	73.3 (9.7) $p = 0.437$	**36.1 (6.2)** $p < 0.001$
Phys	25	**9.2 (3.2)** $p < 0.001$	20.5 (4.9) $p = 0.201$	**9.1 (3)** $p < 0.001$
Soc	12	5.3 (2.4) $p = 0.013$	11.3 (3.6) $p = 0.447$	6 (2.5) $p = 0.024$

Table S2. Observed and expected number of IPs for the dataset Italy, 2015. For each field, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

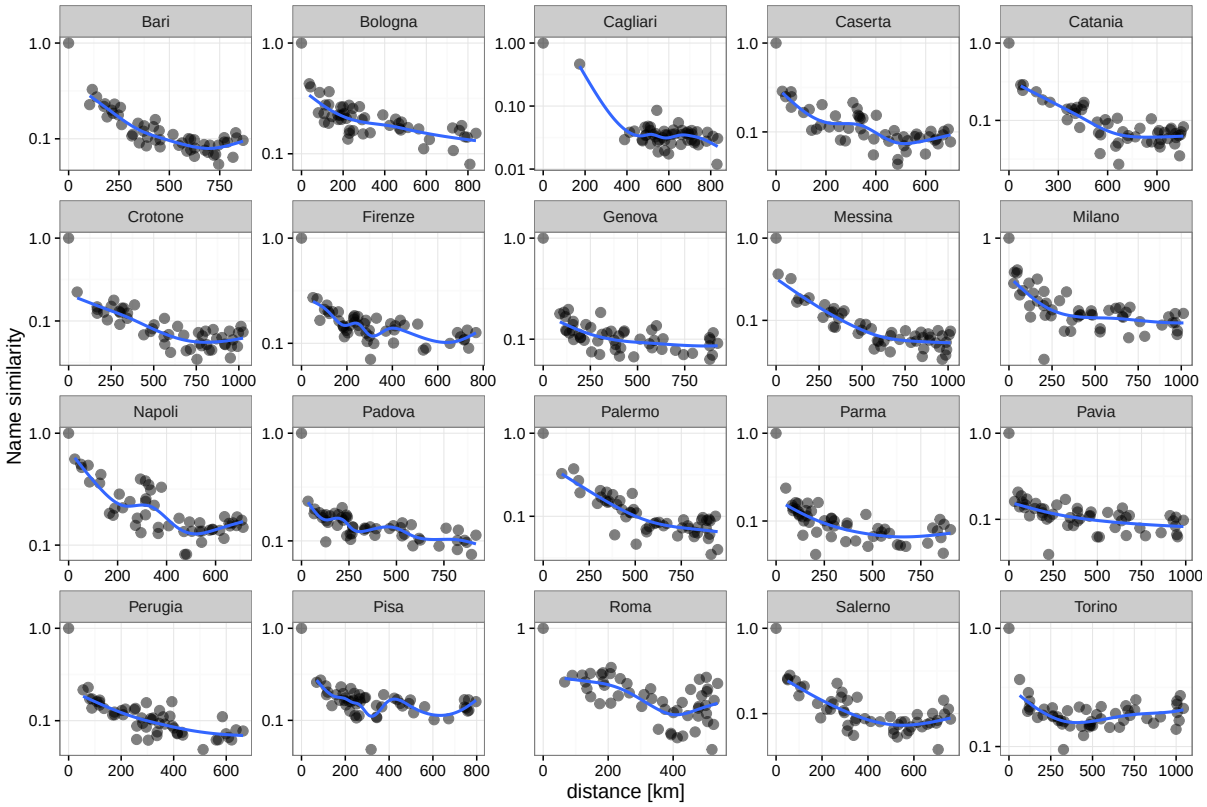

Fig. S3. Correlation between similarity of the last names of two cities and their geographical distance for the 20 cities in Italy with the largest number of researchers. Each figure shows the similarity $S_{c|c'}$, as defined in Eq. S3, between a city c (indicated in the title of the panel) and all the other cities c' vs the geographical distance between c and c'. Each city c is compared only to cities c' with more than 50 researchers. The blue line is the best fitting spline, once removed the point at distance 0 (corresponding to $S_{c|c}$, which is equal to 1 by definition).

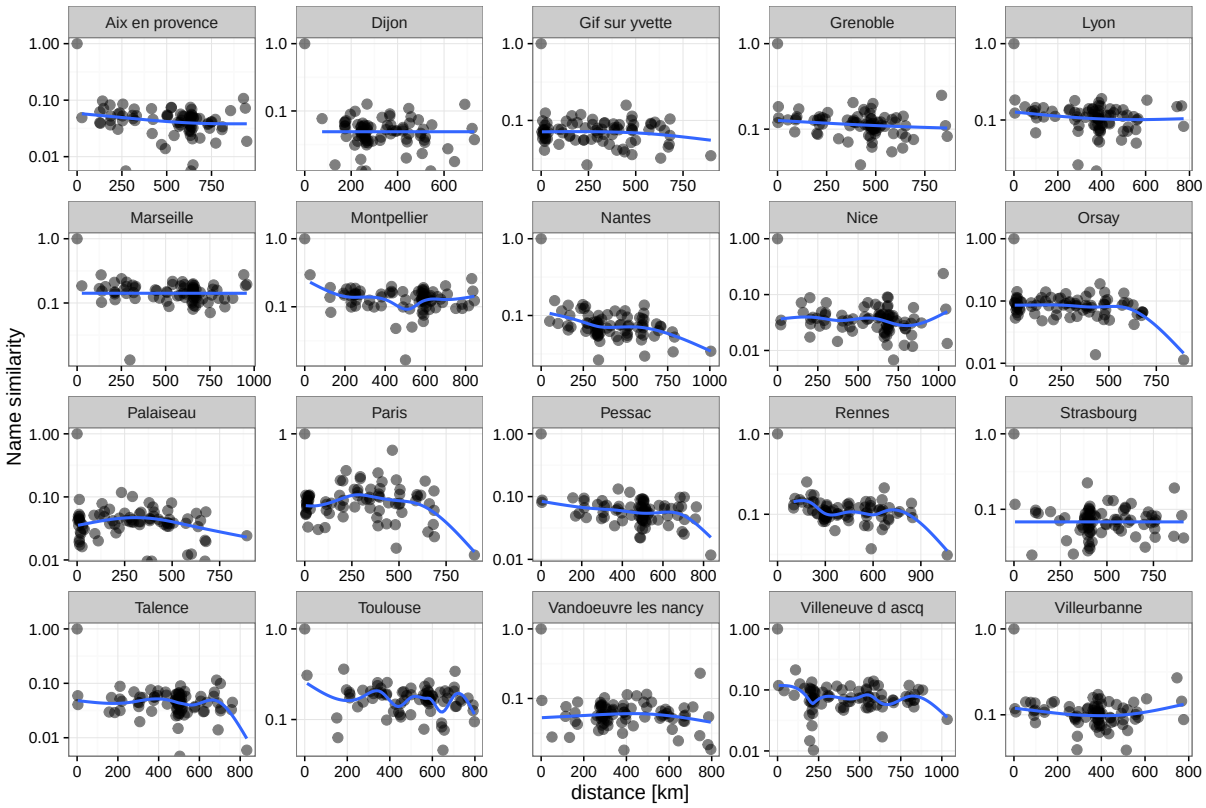

Fig. S4. Same as Figure S3, for the 20 cities in France with the largest number of researchers.

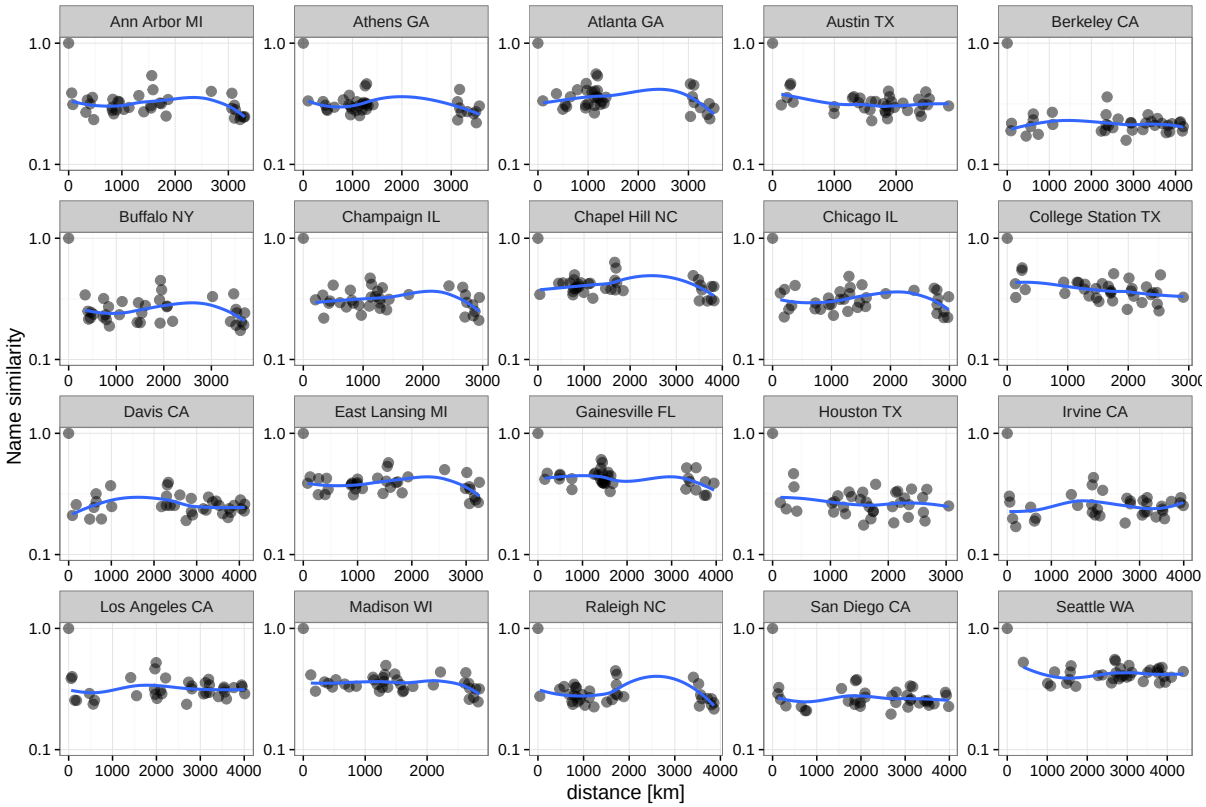

Fig. S5. Same as Figure S3, for the 20 cities in the US with the largest number of researchers.

region	observed	by country	by city	by field
Abruzzo	21	**7.3 (2.9)** $p < 0.001$	14.7 (3.6) $p = 0.062$	**7.6 (2.9)** $p < 0.001$
Basilicata	4	0.8 (0.9) $p = 0.011$	1.8 (1.3) $p = 0.1$	0.8 (0.9) $p = 0.01$
Calabria	33	**6.2 (2.6)** $p < 0.001$	31.9 (5.4) $p = 0.444$	**6.4 (2.7)** $p < 0.001$
Campania	241	**53.8 (8.4)** $p < 0.001$	**177.3 (14.1)** $p < 0.001$	**56.3 (8.5)** $p < 0.001$
Emilia-Romagna	138	**56.6 (8.6)** $p < 0.001$	112.1 (10.4) $p = 0.009$	**58.3 (8.6)** $p < 0.001$
Friuli-Venezia Giulia	7	5.6 (2.5) $p = 0.334$	5.5 (2.3) $p = 0.309$	5.7 (2.5) $p = 0.343$
Lazio	209	**148.9 (17.3)** $p = 0.002$	180.8 (16.4) $p = 0.052$	157.1 (16.7) $p = 0.003$
Liguria	27	**11.6 (3.8)** $p < 0.001$	26.1 (5.1) $p = 0.447$	**12.1 (3.8)** $p = 0.001$
Lombardia	244	**120.9 (13.5)** $p < 0.001$	211.7 (17.7) $p = 0.043$	**127.7 (13.6)** $p < 0.001$
Marche	13	5.2 (2.4) $p = 0.005$	11.3 (3.1) $p = 0.334$	5.4 (2.4) $p = 0.006$
Molise	1	0.5 (0.7) $p = 0.407$	1 (1) $p = 0.664$	0.5 (0.8) $p = 0.418$
Piemonte	69	43.6 (7.8) $p = 0.003$	56.8 (7.5) $p = 0.063$	45.4 (7.9) $p = 0.005$
Puglia	82	**22.1 (5.2)** $p < 0.001$	**55.4 (7.2)** $p < 0.001$	**22.8 (5.2)** $p < 0.001$
Sardegna	99	**9.2 (3.2)** $p < 0.001$	83.6 (9.1) $p = 0.056$	**9.4 (3.2)** $p < 0.001$
Sicilia	236	**37 (6.8)** $p < 0.001$	**181.4 (13.5)** $p < 0.001$	**38.7 (6.9)** $p < 0.001$
Toscana	79	**33.5 (6.4)** $p < 0.001$	62.5 (7.8) $p = 0.024$	**34.6 (6.4)** $p < 0.001$
Trentino-Alto Adige	4	2.6 (1.7) $p = 0.261$	3.1 (1.7) $p = 0.37$	2.7 (1.7) $p = 0.285$
Umbria	26	**7.7 (3)** $p < 0.001$	16.7 (4) $p = 0.02$	**7.9 (3)** $p < 0.001$
Valle D'Aosta	0	0 (0.2) $p = 1$	0 (0) $-$	0 (0.2) $p = 1$
Veneto	53	36.3 (6.8) $p = 0.014$	54.4 (7.2) $p = 0.595$	37.3 (6.8) $p = 0.019$

Table S3. Observed and expected number of IPs for the dataset Italy, 2015. For each region, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Cell	13	8.1 (3) $p = 0.082$	10.3 (3.4) $p = 0.244$	7.7 (2.8) $p = 0.054$
Chem	31	**15.4 (4.1)** $p < 0.001$	24.2 (4.9) $p = 0.104$	18.2 (4.3) $p = 0.005$
Eng	27	**10.3 (3.4)** $p < 0.001$	22.8 (4) $p = 0.172$	**11.8 (3.5)** $p < 0.001$
Env	16	**6.9 (2.7)** $p = 0.004$	9.5 (3) $p = 0.031$	9.4 (3.1) $p = 0.031$
Genet	7	4.5 (2.3) $p = 0.185$	5.3 (2.4) $p = 0.281$	4.4 (2.1) $p = 0.153$
Geo	13	7.7 (2.9) $p = 0.058$	9.1 (3.2) $p = 0.139$	8.1 (2.8) $p = 0.069$
HE Phys	1	1.8 (1.4) $p = 0.82$	1.8 (1.4) $p = 0.824$	1.5 (1.2) $p = 0.791$
Hum	56	**27.3 (5.4)** $p < 0.001$	40.7 (6.2) $p = 0.012$	**28.3 (5.4)** $p < 0.001$
Info	73	**42.6 (7.1)** $p < 0.001$	60.4 (7.9) $p = 0.066$	**42.1 (6.5)** $p < 0.001$
Math	32	**16.6 (4.4)** $p = 0.002$	21.4 (5.1) $p = 0.031$	**15.9 (4)** $p < 0.001$
Neuro	10	5.8 (2.5) $p = 0.08$	7.1 (2.7) $p = 0.178$	6 (2.4) $p = 0.084$
Phys	15	8.4 (3.1) $p = 0.033$	10.4 (3.4) $p = 0.12$	7.7 (2.8) $p = 0.014$

Table S4. Observed and expected number of IPs for the dataset France CNRS using maiden names, 2016. For each field, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

region	observed	by country	by city	by field
Alsace	16	6.5 (2.8) $p = 0.004$	12.2 (3.1) $p = 0.142$	6.7 (2.8) $p = 0.004$
Aquitaine	9	6.9 (2.8) $p = 0.267$	8.1 (2.6) $p = 0.414$	7.2 (2.8) $p = 0.303$
Auvergne	4	2.5 (1.7) $p = 0.238$	2.9 (1.5) $p = 0.323$	2.6 (1.7) $p = 0.253$
Basse-Normandie	2	1 (1) $p = 0.265$	2.1 (1.4) $p = 0.633$	1 (1) $p = 0.267$
Bourgogne	2	1.7 (1.4) $p = 0.507$	2.2 (1.4) $p = 0.667$	1.9 (1.4) $p = 0.552$
Bretagne	32	**12.5 (3.9)** $p < 0.001$	26.4 (4.8) $p = 0.142$	**12.9 (3.9)** $p < 0.001$
Centre	1	1.6 (1.3) $p = 0.787$	1.4 (1.1) $p = 0.786$	1.6 (1.3) $p = 0.798$
Champagne-Ardenne	1	0.8 (1) $p = 0.553$	0.4 (0.5) $p = 0.355$	0.8 (1) $p = 0.556$
Corse	11	**0.5 (0.8)** $p < 0.001$	7 (1.9) $p = 0.031$	**0.6 (0.8)** $p < 0.001$
Franche-Comté	8	2.9 (1.9) $p = 0.022$	5.6 (2) $p = 0.177$	3.1 (1.9) $p = 0.024$
Haute-Normandie	2	0.7 (0.9) $p = 0.152$	1.4 (1) $p = 0.436$	0.8 (0.9) $p = 0.171$
Ile-de-France	51	40 (6.6) $p = 0.062$	41.8 (6.1) $p = 0.082$	40.9 (6.7) $p = 0.078$
Languedoc-Roussillon	15	7.5 (2.9) $p = 0.015$	11 (3) $p = 0.122$	8.2 (3) $p = 0.029$
Limousin	3	1.9 (1.5) $p = 0.283$	3.2 (1.2) $p = 0.715$	1.9 (1.5) $p = 0.29$
Lorraine	10	6.2 (2.7) $p = 0.11$	9.3 (2.8) $p = 0.45$	6.6 (2.7) $p = 0.14$
Midi-Pyrénées	23	13.9 (4) $p = 0.025$	17.3 (4.2) $p = 0.109$	14.2 (4) $p = 0.029$
Nord-Pas-de-Calais	15	8.4 (3.1) $p = 0.037$	11.6 (3.1) $p = 0.174$	8.6 (3.1) $p = 0.04$
PaysdelaLoire	3	3.3 (1.9) $p = 0.637$	3.6 (1.8) $p = 0.72$	3.6 (2) $p = 0.692$
Picardie	0	0.6 (0.8) $p = 1$	0.6 (0.7) $p = 1$	0.6 (0.8) $p = 1$
Poitou-Charentes	11	3.1 (1.9) $p = 0.002$	11 (2) $p = 0.598$	3.6 (2.1) $p = 0.005$
Provence-Alpes-Côted'Azur	31	**11.8 (3.6)** $p < 0.001$	**16.6 (3.9)** $p = 0.001$	**12.3 (3.6)** $p < 0.001$
Rhône-Alpes	44	**21.2 (4.9)** $p < 0.001$	**27.3 (5)** $p = 0.001$	**21.5 (4.9)** $p < 0.001$

Table S5. Observed and expected number of IPs for the dataset France CNRS using maiden names, 2016. For each region, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Agr	35	33.8 (7.3) $p = 0.434$	35.2 (7.1) $p = 0.519$	37.4 (6.3) $p = 0.663$
Bio	279	**215.6 (21.1)** $p = 0.003$	222 (20.2) $p = 0.005$	258.4 (18.3) $p = 0.137$
Chem	17	11.1 (3.7) $p = 0.079$	12.3 (3.9) $p = 0.14$	20.5 (4.6) $p = 0.805$
Econ	40	26.9 (6) $p = 0.025$	33.2 (6.7) $p = 0.174$	32.1 (5.7) $p = 0.1$
Geo	9	5.5 (2.6) $p = 0.128$	6.2 (2.7) $p = 0.189$	5.7 (2.4) $p = 0.124$
Hum	44	58.6 (9.3) $p = 0.956$	62.2 (9.3) $p = 0.984$	41.2 (6.5) $p = 0.346$
IndEng	40	25.8 (5.9) $p = 0.018$	29.6 (6.4) $p = 0.067$	48.6 (7.4) $p = 0.897$
Math	124	**61.5 (9.5)** $p < 0.001$	**70.5 (10.2)** $p < 0.001$	103.8 (10.6) $p = 0.038$
Med	634	589.9 (43.3) $p = 0.157$	597 (37.5) $p = 0.165$	**507.7 (28)** $p < 0.001$
Ped	29	**10.2 (3.5)** $p < 0.001$	**11.8 (3.8)** $p < 0.001$	**9.3 (3)** $p < 0.001$
Phys	64	**26 (5.9)** $p < 0.001$	**29 (6.2)** $p < 0.001$	53.6 (7.5) $p = 0.097$
Soc	164	150.1 (15.9) $p = 0.196$	169.8 (16.7) $p = 0.638$	**129.8 (11.6)** $p = 0.003$

Table S6. Observed and expected number of IPs for the dataset US public research institutions, 2016. For each field, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

region	observed	by country	by city	by field
California	203	176.2 (18.3) $p = 0.08$	167.1 (13.4) $p = 0.006$	178.4 (17.6) $p = 0.089$
Florida	199	183.8 (23.7) $p = 0.255$	183.5 (15.5) $p = 0.164$	184.1 (22.1) $p = 0.245$
Georgia	86	**53.5 (9.3)** $p = 0.001$	67.4 (8.2) $p = 0.018$	60.7 (10.3) $p = 0.014$
Illinois	104	82.6 (14) $p = 0.075$	86.3 (10.3) $p = 0.055$	84.2 (13.3) $p = 0.08$
Kansas	16	11.3 (3.9) $p = 0.14$	11.3 (3.2) $p = 0.102$	12.8 (4.2) $p = 0.245$
Michigan	137	112.8 (14.6) $p = 0.059$	133.3 (12.6) $p = 0.387$	121.3 (15.1) $p = 0.157$
New York	77	56.1 (9.4) $p = 0.022$	67.3 (8.4) $p = 0.136$	58.4 (9.6) $p = 0.037$
North Carolina	173	154.9 (21.9) $p = 0.201$	146.6 (14.8) $p = 0.048$	154.4 (20.1) $p = 0.178$
Texas	161	**104 (12.9)** $p < 0.001$	132.4 (11.4) $p = 0.009$	120.7 (14.4) $p = 0.005$
Washington	250	200.1 (28.1) $p = 0.049$	224.1 (20) $p = 0.105$	193 (24.8) $p = 0.019$
Wisconsin	73	79.5 (13.4) $p = 0.684$	59.8 (7.9) $p = 0.06$	80 (13.1) $p = 0.705$

Table S7. Observed and expected number of IPs for the dataset US public research institutions, 2016. For each region, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

S5. Married vs. maiden names

As detailed in the main text, for the French data set, we repeated the randomization when forcing married women to take their husband's last name. In particular, 2,933 researchers listed a different maiden and married names (out of 44,860 researchers). Tables S8 and S9 detail the results shown in Figure 3 of the main text.

field	observed	by country	by city	by field
Cell	36	**8.1 (3)** $p < 0.001$	**11.9 (3.6)** $p < 0.001$	**8.4 (2.9)** $p < 0.001$
Chem	65	**15.4 (4.1)** $p < 0.001$	**27 (5.2)** $p < 0.001$	**18.7 (4.4)** $p < 0.001$
Eng	43	**10.3 (3.4)** $p < 0.001$	**27.8 (4.3)** $p < 0.001$	**11.5 (3.4)** $p < 0.001$
Env	25	**6.9 (2.7)** $p < 0.001$	**12.2 (3.1)** $p < 0.001$	**9.6 (3.1)** $p < 0.001$
Genet	16	**4.5 (2.3)** $p < 0.001$	**6.4 (2.6)** $p = 0.002$	**5.5 (2.4)** $p < 0.001$
Geo	35	**7.7 (2.9)** $p < 0.001$	**12.8 (3.4)** $p < 0.001$	**8.9 (3)** $p < 0.001$
HE Phys	8	**1.7 (1.4)** $p = 0.001$	3 (1.6) $p = 0.008$	**1.9 (1.3)** $p < 0.001$
Hum	79	**27.2 (5.4)** $p < 0.001$	**45.7 (6.5)** $p < 0.001$	**28 (5.3)** $p < 0.001$
Info	107	**42.5 (7.1)** $p < 0.001$	**74.6 (8.3)** $p < 0.001$	**43.1 (6.6)** $p < 0.001$
Math	42	**16.6 (4.4)** $p < 0.001$	**23.3 (5.2)** $p = 0.001$	**16.7 (4.1)** $p < 0.001$
Neuro	24	**5.8 (2.5)** $p < 0.001$	**8.7 (2.9)** $p < 0.001$	**6.5 (2.5)** $p < 0.001$
Phys	33	**8.4 (3)** $p < 0.001$	**11.6 (3.5)** $p < 0.001$	**8.1 (2.8)** $p < 0.001$

Table S8. Observed and expected number of IPs for the dataset France CNRS using married names, 2016. For each field, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

region	observed	by country	by city	by field
Alsace	26	**6.5 (2.8)** $p < 0.001$	**13.4 (3.3)** $p < 0.001$	**7 (2.8)** $p < 0.001$
Aquitaine	20	**6.9 (2.8)** $p < 0.001$	**10.1 (2.9)** $p = 0.002$	**7.4 (2.9)** $p < 0.001$
Auvergne	14	**2.5 (1.7)** $p < 0.001$	**5.3 (2.1)** $p < 0.001$	**2.7 (1.7)** $p < 0.001$
Basse-Normandie	5	1 (1) $p = 0.007$	2.1 (1.4) $p = 0.054$	1.1 (1.1) $p = 0.007$
Bourgogne	9	**1.7 (1.4)** $p < 0.001$	3.2 (1.7) $p = 0.003$	**1.9 (1.4)** $p < 0.001$
Bretagne	45	**12.4 (3.9)** $p < 0.001$	**26.8 (4.8)** $p < 0.001$	**13.2 (3.9)** $p < 0.001$
Centre	11	**1.6 (1.3)** $p < 0.001$	**3 (1.6)** $p < 0.001$	**1.7 (1.3)** $p < 0.001$
Champagne-Ardenne	6	**0.8 (1)** $p = 0.001$	4.1 (0.8) $p = 0.045$	**0.9 (1)** $p = 0.001$
Corse	13	**0.5 (0.8)** $p < 0.001$	9 (2.1) $p = 0.051$	**0.6 (0.8)** $p < 0.001$
Franche-Comté	11	**2.9 (1.9)** $p = 0.002$	6.6 (2.2) $p = 0.046$	3.2 (1.9) $p = 0.002$
Haute-Normandie	3	0.7 (0.9) $p = 0.037$	1.8 (1.1) $p = 0.239$	0.8 (0.9) $p = 0.044$
Ile-de-France	73	**39.9 (6.6)** $p < 0.001$	**45.5 (6.3)** $p < 0.001$	**43 (6.8)** $p < 0.001$
Languedoc-Roussillon	27	**7.4 (2.9)** $p < 0.001$	**12.2 (3)** $p < 0.001$	**8.6 (3.1)** $p < 0.001$
Limousin	8	1.9 (1.5) $p = 0.004$	5.9 (1.6) $p = 0.156$	1.9 (1.5) $p = 0.004$
Lorraine	20	**6.1 (2.7)** $p < 0.001$	10.9 (3) $p = 0.004$	**6.7 (2.8)** $p < 0.001$
Midi-Pyrénées	42	**13.8 (4)** $p < 0.001$	**18.8 (4.3)** $p < 0.001$	**14.6 (4.1)** $p < 0.001$
Nord-Pas-de-Calais	26	**8.4 (3.1)** $p < 0.001$	**14.3 (3.3)** $p = 0.002$	**8.8 (3.2)** $p < 0.001$
PaysdelaLoire	8	3.3 (1.9) $p = 0.026$	4.7 (2) $p = 0.09$	3.7 (2) $p = 0.042$
Picardie	1	0.6 (0.8) $p = 0.423$	1.3 (0.9) $p = 0.79$	0.6 (0.8) $p = 0.444$
Poitou-Charentes	14	**3.1 (1.9)** $p < 0.001$	13.4 (2.1) $p = 0.489$	**3.6 (2)** $p < 0.001$
Provence-Alpes-Côted'Azur	60	**11.8 (3.6)** $p < 0.001$	**20 (4.2)** $p < 0.001$	**12.8 (3.7)** $p < 0.001$
Rhône-Alpes	71	**21.2 (4.8)** $p < 0.001$	**32.6 (5.3)** $p < 0.001$	**22.4 (4.9)** $p < 0.001$

Table S9. Observed and expected number of IPs for the dataset France CNRS using married names, 2016. For each region, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

S6. Analysis of first names

Before presenting the figures and tables for the analysis of first names, we want to highlight an important difference between first and last names. Last names are passed down from generation to generation, and behave like neutral alleles (with diversity being maintained by either immigration or "mutation"—the creation of new last names). First names frequencies, on the other hand, fluctuate from year to year—certain names become fashionable and increase in frequency, while others decrease (4). These fluctuations can be quite large, and can be caused by well-defined events. For example, in Figure S6 we show the frequency of the most common names for boys and girls in Italy[†], where it is apparent that the election of Pope Francis coincided with a large increase in the number of boys named Francesco, and a more modest increase in the number of girls being named Francesca.

[†] Data taken from `http://www.istat.it/en/products/interactive-contents/baby-names`.

[‡] Interestingly, the name Giulia grew considerably in popularity in 2003 (and then more modestly in 2004). While it is difficult to pinpoint the cause of this increase, we note that the song "Dedicato a te" by the Italian band Le Vibrazioni—with its powerful chorus "Sei immensamente Giulia!"—was the top-selling single in March 2003.

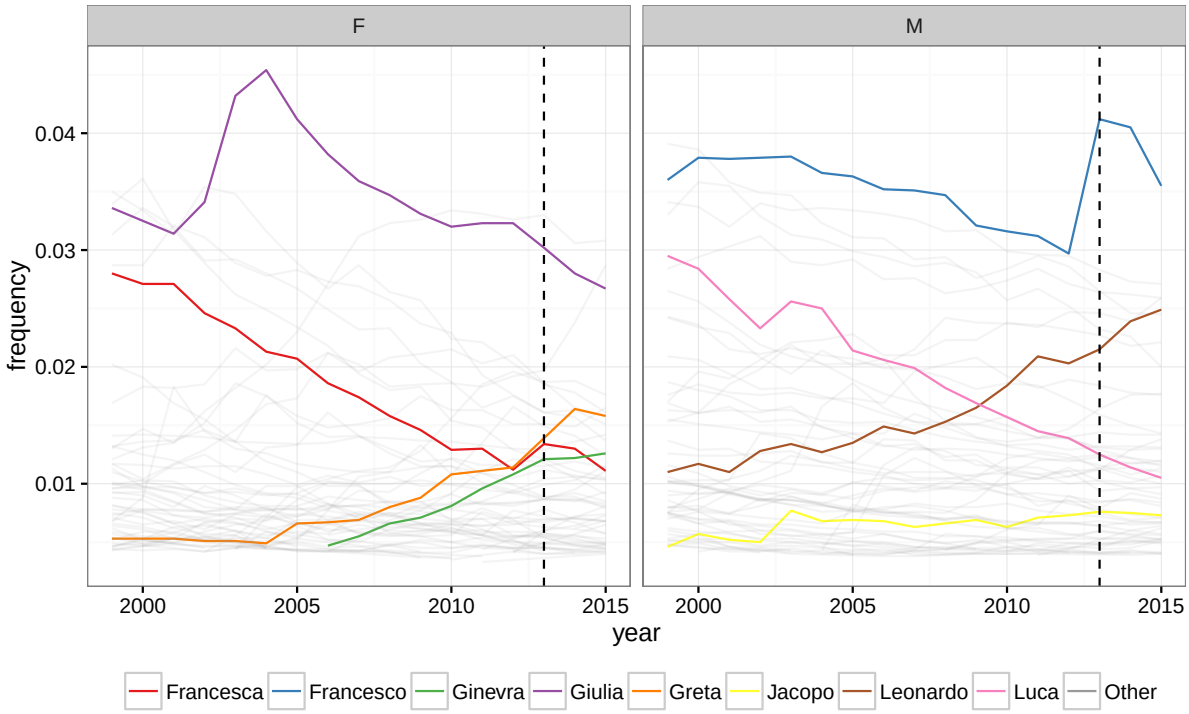

Fig. S6. Frequency of the 100 most common names for boys and girls in Italy from 1999 to 2015. Some of the time series are highlighted in color. Note that in this short span of time names can triple in frequency. Interestingly, the election of Pope Francis in 2013 (dashed line) coincides with a large increase in the number of boys being named Francesco[‡].

This fact can introduce complications in the analysis of first-name isonymies: departments hiring many researchers in a short span of time could result in elevated number of IPs that are simply due to the fact that the new hires have about the same age, and therefore are likely to share first-names that were popular when they were born. For example, it was about 6 times more likely to sample two newborn boys named Leonardo in 2015 than it was in 1999, while the probability of sampling two boys named Luca decreased 9-fold during the same period.

In order to test this effect we measured the number of IPs of first names of the kind full-professor↔not-full-professor. As explained in the main text, Italian last names show a significant excess of pairs of this kind, consistently with the hypothesis of nepotism. If the large temporal fluctuations of first names affect their occurrences, we would expect people of similar age to have more similar names and, therefore, to have significantly fewer pairs of the kind full-professor↔not-full-professor. In order to remove the effects of different gender imbalance of departments, we analyzed male and female researchers separately. Consistently with our hypothesis, we obtained significant results for all the years and both genders (2000: p-value $< 10^{-4}$, 2005: 0.018, 2010: 0.018, 2015: 0.019 for males and 2000: p-value $< 10^{-4}$, 2005: 0.018, 2010: 0.022, 2015: 0.015 for females).

With this caveat in mind, we report the results for first-name IPs analysis in Figure S7 and Tables S10, S11, S12, S13, S14 and S15. As shown in Fig. 4 of the main text, scarcity of first names is associated with fields in which women (or men) are under-represented. Note however that fields in which immigration is preponderant (e.g., Mathematics and Physics in the US) would show the opposite trend — the fields are much richer in first names than expected by chance.

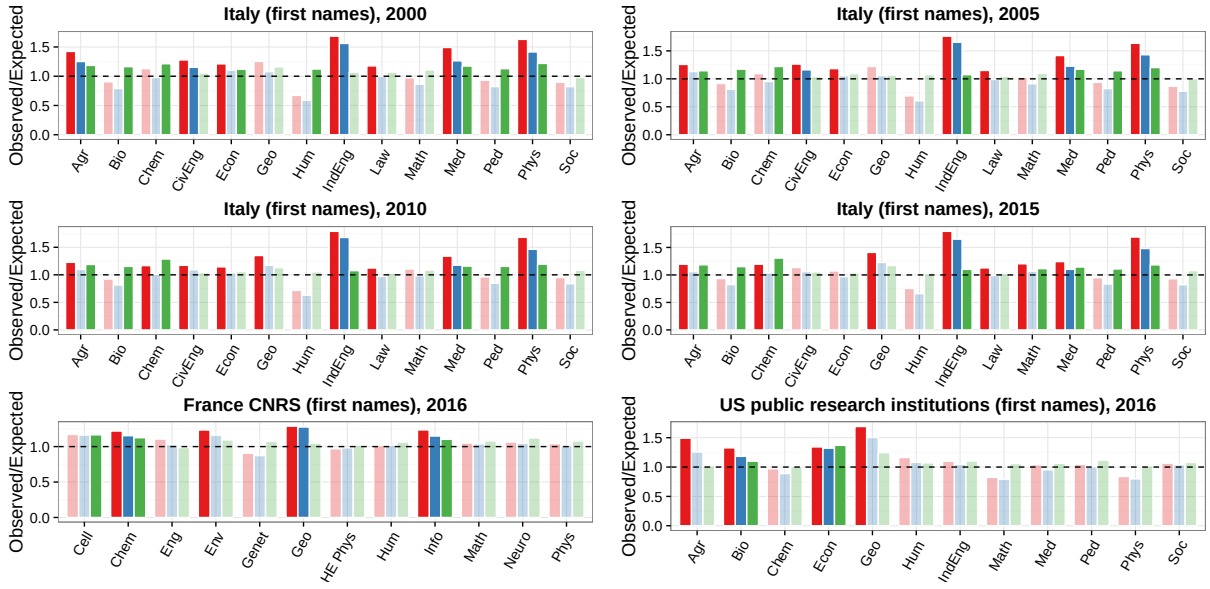

Fig. S7. The same randomizations as in Fig. 1 of the main text, but using first names rather than last names.

field	observed	by country	by city	by field
Agr	1840	**1292.4 (70.8)** $p < 0.001$	**1471 (80.7)** $p < 0.001$	**1552.7 (50.2)** $p < 0.001$
Bio	1992	2203.2 (95.9) $p = 0.988$	2521 (108.9) $p = 1$	**1712.4 (52.1)** $p < 0.001$
Chem	1317	1167.1 (64) $p = 0.013$	1345.1 (71.4) $p = 0.647$	**1085.3 (40.6)** $p < 0.001$
CivEng	2664	**2083.5 (107.9)** $p < 0.001$	**2312.3 (114.3)** $p = 0.002$	2528.9 (80) $p = 0.049$
Econ	1440	**1188.9 (66.2)** $p < 0.001$	1304.4 (69.8) $p = 0.032$	**1288 (48.6)** $p = 0.002$
Geo	243	193.8 (19.9) $p = 0.01$	225.4 (22.5) $p = 0.22$	209.4 (15.4) $p = 0.021$
Hum	1910	2840.8 (115.2) $p = 1$	3234.5 (125.2) $p = 1$	**1701 (49.7)** $p < 0.001$
IndEng	4591	**2725.1 (125.8)** $p < 0.001$	**2942.2 (127.1)** $p < 0.001$	4299.7 (110.3) $p = 0.007$
Law	1600	**1360.6 (72)** $p = 0.001$	1603 (83.9) $p = 0.5$	1494.3 (55.6) $p = 0.034$
Math	934	958.8 (57.1) $p = 0.661$	1080.2 (61.8) $p = 0.993$	842.4 (35.6) $p = 0.007$
Med	17983	**12061.3 (328.9)** $p < 0.001$	**14239.9 (373.4)** $p < 0.001$	**15320.8 (261)** $p < 0.001$
Ped	1882	2019.7 (97.3) $p = 0.923$	2286.4 (103.6) $p = 1$	**1667.1 (58.8)** $p < 0.001$
Phys	1047	**642.6 (43.1)** $p < 0.001$	**740.2 (48.7)** $p < 0.001$	**859 (36.2)** $p < 0.001$
Soc	191	212.8 (22.1) $p = 0.844$	232.3 (23.1) $p = 0.974$	194.8 (16.3) $p = 0.588$

Table S10. Observed and expected number of first-name IPs for the dataset Italy (first names), 2000. For each field, we report the observed number of pairs, as well as the expectation when randomizing first names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Agr	1984	**1577.7 (81)** $p < 0.001$	1756 (90.5) $p = 0.009$	**1733.9 (53.9)** $p < 0.001$
Bio	2432	2652.6 (107.1) $p = 0.986$	2986.2 (120.5) $p = 1$	**2073.1 (56.6)** $p < 0.001$
Chem	1249	1141.4 (61) $p = 0.044$	1309.7 (68.9) $p = 0.812$	**1023.2 (37.8)** $p < 0.001$
CivEng	2954	**2336.8 (115)** $p < 0.001$	**2543.4 (118.5)** $p < 0.001$	2837.7 (91.7) $p = 0.105$
Econ	1811	**1531.1 (73)** $p < 0.001$	1713.5 (80.1) $p = 0.116$	1658.6 (53.4) $p = 0.004$
Geo	215	176.1 (18.2) $p = 0.024$	203.2 (20.6) $p = 0.279$	201.6 (15.3) $p = 0.197$
Hum	2139	3079.2 (117) $p = 1$	3525.4 (129.1) $p = 1$	1989.1 (53.5) $p = 0.004$
IndEng	5869	**3327.9 (142.7)** $p < 0.001$	**3547 (142.5)** $p < 0.001$	**5462.4 (136.6)** $p = 0.003$
Law	2013	**1750.1 (80)** $p = 0.001$	2035.6 (94.4) $p = 0.583$	1924.7 (58.2) $p = 0.07$
Math	1093	1063.6 (58.3) $p = 0.302$	1197.6 (64.1) $p = 0.954$	995.9 (38.1) $p = 0.007$
Med	23755	**16770.5 (447.4)** $p < 0.001$	**19369.8 (487.8)** $p < 0.001$	**20291.7 (385.8)** $p < 0.001$
Ped	2143	2285.4 (97.8) $p = 0.93$	2593.5 (109.5) $p = 1$	**1872.8 (58.9)** $p < 0.001$
Phys	1078	**659.1 (42)** $p < 0.001$	**753.5 (46.9)** $p < 0.001$	**896.9 (36.2)** $p < 0.001$
Soc	234	269.9 (24.7) $p = 0.938$	300.6 (26.6) $p = 0.998$	233.7 (17.5) $p = 0.487$

Table S11. Observed and expected number of first-name IPs for the dataset Italy (first names), 2005. For each field, we report the observed number of pairs, as well as the expectation when randomizing first names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Agr	1658	**1351.3 (70.6)** $p < 0.001$	1509.5 (78) $p = 0.034$	**1395.4 (46.1)** $p < 0.001$
Bio	2041	2209.7 (91.2) $p = 0.972$	2500.8 (105.3) $p = 1$	**1770.4 (49.5)** $p < 0.001$
Chem	1047	**896.8 (50.7)** $p = 0.002$	1037.7 (58.4) $p = 0.435$	**813.6 (32.6)** $p < 0.001$
CivEng	2320	**1979 (104.6)** $p < 0.001$	2123.9 (105.1) $p = 0.035$	2232.6 (83.2) $p = 0.148$
Econ	1890	**1651.2 (72.7)** $p < 0.001$	1822.3 (78.6) $p = 0.197$	1784.6 (54.3) $p = 0.031$
Geo	162	**120 (14)** $p = 0.003$	138.1 (15.9) $p = 0.074$	143.3 (12.6) $p = 0.078$
Hum	1724	2401.1 (95.1) $p = 1$	2742 (107.8) $p = 1$	1635.5 (47.7) $p = 0.034$
IndEng	6259	**3497.6 (147.8)** $p < 0.001$	**3730.5 (146.2)** $p < 0.001$	**5816.5 (150.1)** $p = 0.003$
Law	1864	**1659.5 (71.3)** $p = 0.004$	1919.8 (86.2) $p = 0.74$	1813.7 (53.8) $p = 0.177$
Math	1042	942.2 (50.9) $p = 0.029$	1058.8 (57.1) $p = 0.61$	958.6 (36) $p = 0.013$
Med	18336	**13694.3 (377.3)** $p < 0.001$	**15603.8 (408.8)** $p < 0.001$	**15836 (316.6)** $p < 0.001$
Ped	2002	2080.2 (91.3) $p = 0.807$	2365.1 (102.4) $p = 1$	**1735.2 (56.3)** $p < 0.001$
Phys	843	**501.6 (34.4)** $p < 0.001$	**576.2 (38.8)** $p < 0.001$	**706.5 (31.5)** $p < 0.001$
Soc	253	266.2 (24.2) $p = 0.707$	301.5 (26.7) $p = 0.974$	233.7 (17.3) $p = 0.137$

Table S12. Observed and expected number of first-name IPs for the dataset Italy (first names), 2010. For each field, we report the observed number of pairs, as well as the expectation when randomizing first names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Agr	1467	**1227.7 (65)** $p < 0.001$	1384.1 (72.8) $p = 0.129$	**1238.7 (43)** $p < 0.001$
Bio	1754	1877.1 (80.6) $p = 0.943$	2127 (91.8) $p = 1$	**1524.1 (46.4)** $p < 0.001$
Chem	909	**760.5 (43.4)** $p < 0.001$	877.2 (51.8) $p = 0.254$	**696.1 (30)** $p < 0.001$
CivEng	1794	1580.9 (86.1) $p = 0.01$	1690 (90.6) $p = 0.124$	1699.2 (70.6) $p = 0.095$
Econ	1647	1536.5 (69.2) $p = 0.064$	1698.5 (74) $p = 0.748$	1589.3 (50.5) $p = 0.128$
Geo	137	**97.2 (12.3)** $p = 0.003$	111.7 (13.3) $p = 0.04$	116.9 (11.6) $p = 0.051$
Hum	1392	1845.4 (78.9) $p = 1$	2109.1 (89.5) $p = 1$	1360.4 (43.3) $p = 0.228$
IndEng	6275	**3500.1 (148.9)** $p < 0.001$	**3802.5 (149.2)** $p < 0.001$	**5695.9 (153.3)** $p < 0.001$
Law	1566	**1388.5 (62)** $p = 0.003$	1581.7 (72.4) $p = 0.58$	1525.1 (48.4) $p = 0.206$
Math	935	**777.4 (43.9)** $p < 0.001$	876.9 (50.5) $p = 0.133$	**838.7 (33.7)** $p = 0.002$
Med	12908	**10390.5 (308.9)** $p < 0.001$	**11715.8 (331.2)** $p = 0.001$	**11283.6 (251.2)** $p < 0.001$
Ped	1477	1553.5 (73.5) $p = 0.853$	1767.7 (82.6) $p = 1$	**1332.2 (46.7)** $p < 0.001$
Phys	692	**409.8 (29.1)** $p < 0.001$	**467.5 (33.7)** $p < 0.001$	**585 (27.5)** $p < 0.001$
Soc	219	235.1 (21.6) $p = 0.772$	266.6 (23.7) $p = 0.982$	203.1 (16) $p = 0.165$

Table S13. Observed and expected number of first-name IPs for the dataset Italy (first names), 2015. For each field, we report the observed number of pairs, as well as the expectation when randomizing first names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Cell	521	444.7 (29.5) $p = 0.008$	448 (28.4) $p = 0.009$	**447 (23.2)** $p = 0.001$
Chem	1029	**844.1 (40.4)** $p < 0.001$	**891.9 (38.6)** $p < 0.001$	**914.5 (33.8)** $p < 0.001$
Eng	624	564.5 (33.4) $p = 0.044$	607 (27.7) $p = 0.265$	631.9 (28.7) $p = 0.604$
Env	466	**377.7 (25.1)** $p < 0.001$	401.1 (23.7) $p = 0.005$	425.8 (21.5) $p = 0.037$
Genet	225	248 (25.8) $p = 0.816$	257.9 (26.1) $p = 0.905$	209.5 (17.4) $p = 0.196$
Geo	542	**420.7 (27.2)** $p < 0.001$	**424.7 (25.9)** $p < 0.001$	516.9 (23.9) $p = 0.152$
HE Phys	93	95.9 (12.3) $p = 0.597$	94.6 (11.6) $p = 0.563$	91.5 (9.5) $p = 0.438$
Hum	1510	1490.3 (50.4) $p = 0.344$	1485.3 (45.8) $p = 0.298$	1420.2 (40.6) $p = 0.013$
Info	2875	**2327.3 (76.6)** $p < 0.001$	**2502.2 (71.7)** $p < 0.001$	**2607.5 (61)** $p < 0.001$
Math	955	910 (47.7) $p = 0.173$	919.2 (45.9) $p = 0.218$	884.5 (35.3) $p = 0.029$
Neuro	338	317.7 (23.1) $p = 0.193$	323.1 (22.3) $p = 0.261$	301.3 (18.1) $p = 0.025$
Phys	479	460.2 (29.5) $p = 0.265$	472.7 (29.7) $p = 0.416$	443.8 (23.7) $p = 0.074$

Table S14. Observed and expected number of first-name IPs for the dataset France CNRS (first names), 2016. For each field, we report the observed number of pairs, as well as the expectation when randomizing first names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Agr	542	**363.5 (36.6)** $p < 0.001$	431.5 (40.3) $p = 0.007$	530.6 (31) $p = 0.355$
Bio	3083	**2323.6 (117.7)** $p < 0.001$	**2606 (124.5)** $p < 0.001$	**2804.6 (81.2)** $p < 0.001$
Chem	116	119.3 (15.7) $p = 0.581$	130.4 (16.8) $p = 0.812$	115 (11.6) $p = 0.464$
Econ	389	**289.7 (26.7)** $p < 0.001$	**294.3 (27.5)** $p = 0.001$	**284 (18.1)** $p < 0.001$
Geo	101	**59.8 (11)** $p < 0.001$	67.2 (12) $p = 0.01$	81.1 (10.1) $p = 0.035$
Hum	734	631.2 (45.5) $p = 0.018$	679.7 (49) $p = 0.138$	685.7 (30) $p = 0.058$
IndEng	306	278.2 (27.4) $p = 0.158$	292.5 (28.5) $p = 0.309$	277.3 (19.4) $p = 0.079$
Math	545	661.2 (45.3) $p = 0.997$	687.5 (47) $p = 0.999$	514.7 (25.1) $p = 0.123$
Med	6612	6357.3 (268.2) $p = 0.17$	6934.1 (246) $p = 0.908$	6214.2 (157.8) $p = 0.008$
Ped	115	110 (14.5) $p = 0.357$	114.6 (15.2) $p = 0.48$	102.9 (10.4) $p = 0.135$
Phys	235	279.9 (26.9) $p = 0.96$	294.6 (27.7) $p = 0.991$	229.9 (16.3) $p = 0.381$
Soc	1723	1616.7 (81.4) $p = 0.098$	1658.8 (83.8) $p = 0.22$	1598.2 (45.1) $p = 0.004$

Table S15. Observed and expected number of first-name IPs for the dataset US public research institutions (first names), 2016. For each field, we report the observed number of pairs, as well as the expectation when randomizing first names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

S7. Italy: time series

Finally, we present the complete results for the Italian time series in Figures S8 and S9, with the associated Tables S16, S17, S18, S19, S20, and S21. The decrease in number of regions testing significant after a peak in 2005 is evident; similarly, the number of fields testing significant is smaller for 2010 and 2015.

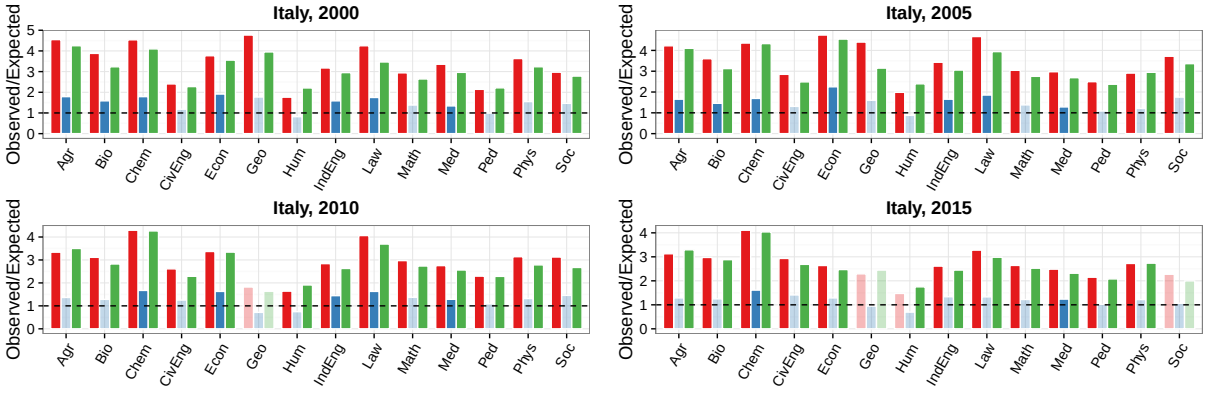

Fig. S8. The same randomizations as in Fig. 1 of the main text, but using first names rather than last names.

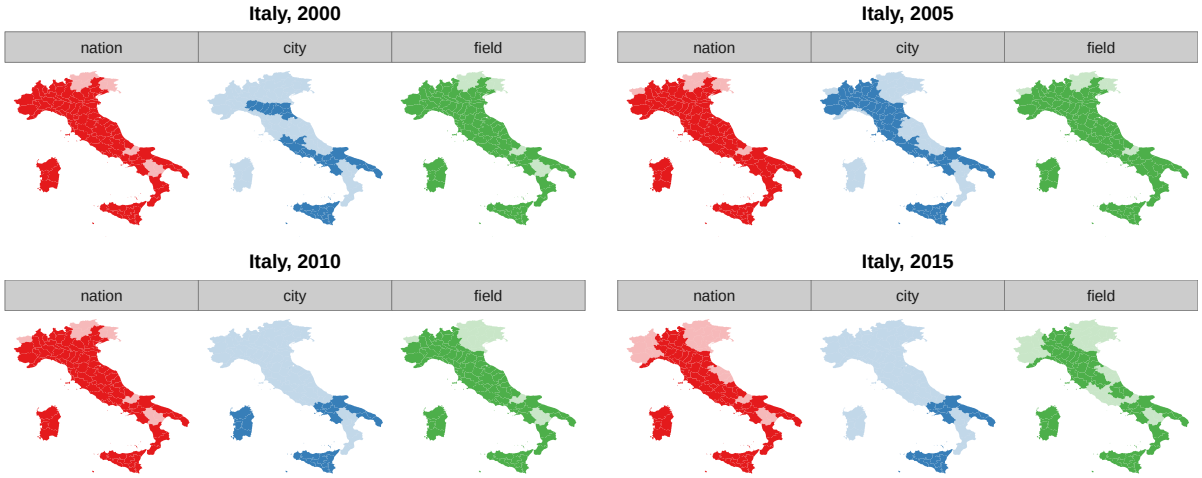

Fig. S9. The same randomizations as in Fig. 1 of the main text, but using first names rather than last names.

While the decrease in significance is a positive sign, as we explain in the main text the main driver of this pattern has been the retirements of professors that were not replaced by new hires. Some regions (e.g., Tuscany, Liguria, Table S22), fields (Geology, Humanities, Table S23) and institutions (Siena, Firenze, Trieste, Pisa, Table S24) have suffered especially dramatic losses.

field	observed	by country	by city	by field
Agr	121	**26.6 (5.7)** $p < 0.001$	**67.6 (9.6)** $p < 0.001$	**28.5 (5.3)** $p < 0.001$
Bio	176	**45.3 (7.6)** $p < 0.001$	**110.9 (12.4)** $p < 0.001$	**54.4 (7.6)** $p < 0.001$
Chem	109	**24 (5.4)** $p < 0.001$	**60.9 (8.9)** $p < 0.001$	**26.6 (5.2)** $p < 0.001$
CivEng	103	**42.9 (7.7)** $p < 0.001$	87.2 (11.5) $p = 0.095$	**45.3 (6.8)** $p < 0.001$
Econ	92	**24.4 (5.5)** $p < 0.001$	**48.1 (7.8)** $p < 0.001$	**25.9 (5.2)** $p < 0.001$
Geo	19	**4 (2.1)** $p < 0.001$	10.7 (3.5) $p = 0.023$	**4.8 (2.2)** $p < 0.001$
Hum	103	**58.4 (8.7)** $p < 0.001$	125.5 (13.1) $p = 0.967$	**46.6 (6.9)** $p < 0.001$
IndEng	178	**56.1 (8.9)** $p < 0.001$	**112.2 (13.2)** $p < 0.001$	**60.3 (7.9)** $p < 0.001$
Law	119	**28 (5.8)** $p < 0.001$	**68.1 (9.4)** $p < 0.001$	**34.3 (6)** $p < 0.001$
Math	58	**19.7 (4.9)** $p < 0.001$	41.9 (7.3) $p = 0.023$	**21.9 (4.7)** $p < 0.001$
Med	832	**247.9 (20.3)** $p < 0.001$	**620.4 (33.2)** $p < 0.001$	**280.7 (17.9)** $p < 0.001$
Ped	89	**41.5 (7.3)** $p < 0.001$	89.6 (11) $p = 0.525$	**40.2 (6.6)** $p < 0.001$
Phys	48	**13.2 (3.9)** $p < 0.001$	31 (6.2) $p = 0.008$	**14.8 (3.9)** $p < 0.001$
Soc	13	**4.4 (2.2)** $p = 0.002$	8.9 (3.2) $p = 0.127$	**4.7 (2.2)** $p = 0.002$

Table S16. Observed and expected number of IPs for the dataset Italy, 2000. For each field, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

region	observed	by country	by city	by field
Abruzzo	17	**5.9 (2.6)** $p < 0.001$	12.2 (3.3) $p = 0.103$	**6.5 (2.7)** $p = 0.001$
Basilicata	5	1.1 (1.1) $p = 0.01$	2.3 (1.4) $p = 0.065$	1.2 (1.1) $p = 0.011$
Calabria	13	**3.2 (1.9)** $p < 0.001$	10.4 (2.9) $p = 0.225$	**3.4 (1.9)** $p < 0.001$
Campania	297	**62.1 (9.3)** $p < 0.001$	**209.1 (15.2)** $p < 0.001$	**67.7 (9.6)** $p < 0.001$
Emilia-Romagna	192	**60.1 (8.9)** $p < 0.001$	**135.5 (11.6)** $p < 0.001$	**64.2 (9.1)** $p < 0.001$
Friuli-Venezia Giulia	16	8.4 (3.1) $p = 0.019$	9.7 (3) $p = 0.036$	8.6 (3.1) $p = 0.021$
Lazio	277	**144.4 (16)** $p < 0.001$	**189.5 (16.1)** $p < 0.001$	**156.9 (16.3)** $p < 0.001$
Liguria	67	**20.4 (5.2)** $p < 0.001$	45.3 (6.9) $p = 0.003$	**22.2 (5.3)** $p < 0.001$
Lombardia	204	**101 (12.3)** $p < 0.001$	165.9 (14.5) $p = 0.008$	**111.2 (12.6)** $p < 0.001$
Marche	13	**4.4 (2.2)** $p = 0.002$	10.5 (3.1) $p = 0.243$	**4.7 (2.3)** $p = 0.003$
Molise	2	0.3 (0.6) $p = 0.046$	0.9 (0.9) $p = 0.239$	0.4 (0.6) $p = 0.053$
Piemonte	82	**41.4 (7.5)** $p < 0.001$	62.8 (8) $p = 0.013$	**44.2 (7.6)** $p < 0.001$
Puglia	102	**19.5 (4.9)** $p < 0.001$	**52.4 (7.1)** $p < 0.001$	**20.7 (5)** $p < 0.001$
Sardegna	130	**9.8 (3.4)** $p < 0.001$	98.4 (10.2) $p = 0.003$	**10.7 (3.5)** $p < 0.001$
Sicilia	425	**53.6 (8.5)** $p < 0.001$	**304.1 (18.7)** $p < 0.001$	**59 (8.8)** $p < 0.001$
Toscana	126	**53.2 (8.2)** $p < 0.001$	101.7 (9.8) $p = 0.01$	**56.7 (8.3)** $p < 0.001$
Trentino-Alto Adige	3	1.2 (1.1) $p = 0.118$	1.6 (1.2) $p = 0.22$	1.3 (1.2) $p = 0.134$
Umbria	23	**8.2 (3.1)** $p < 0.001$	15.2 (3.8) $p = 0.035$	**8.8 (3.2)** $p < 0.001$
Veneto	66	**38.6 (7.1)** $p < 0.001$	55.2 (7.2) $p = 0.081$	**41 (7.3)** $p = 0.002$

Table S17. Observed and expected number of IPs for the dataset Italy, 2000. For each region, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Agr	146	**34.6 (6.6)** $p < 0.001$	**88.2 (11.2)** $p < 0.001$	**35.6 (6)** $p < 0.001$
Bio	209	**58.1 (8.6)** $p < 0.001$	**143.3 (14.5)** $p < 0.001$	**66.9 (8.6)** $p < 0.001$
Chem	109	**25 (5.5)** $p < 0.001$	**64.4 (9.2)** $p < 0.001$	**25.2 (5)** $p < 0.001$
CivEng	146	**51.2 (8.5)** $p < 0.001$	111.5 (13.4) $p = 0.01$	**58.7 (8)** $p < 0.001$
Econ	159	**33.6 (6.4)** $p < 0.001$	**70.6 (9.5)** $p < 0.001$	**35 (6)** $p < 0.001$
Geo	17	**3.9 (2)** $p < 0.001$	10.6 (3.5) $p = 0.056$	**5.4 (2.3)** $p < 0.001$
Hum	134	**67.5 (9.3)** $p < 0.001$	152.5 (14.4) $p = 0.911$	**56 (7.6)** $p < 0.001$
IndEng	250	**72.9 (10.2)** $p < 0.001$	**151.1 (15.8)** $p < 0.001$	**81.8 (9.4)** $p < 0.001$
Law	179	**38.4 (6.8)** $p < 0.001$	**96.6 (11.4)** $p < 0.001$	**45.4 (6.8)** $p < 0.001$
Math	71	**23.3 (5.3)** $p < 0.001$	51.4 (8.1) $p = 0.015$	**25.8 (5.1)** $p < 0.001$
Med	1093	**367.5 (26.8)** $p < 0.001$	**851.4 (41.5)** $p < 0.001$	**406.5 (22.9)** $p < 0.001$
Ped	125	**50.1 (8)** $p < 0.001$	114.9 (12.5) $p = 0.216$	**52.7 (7.5)** $p < 0.001$
Phys	42	**14.4 (4.1)** $p < 0.001$	34.7 (6.6) $p = 0.149$	**14.2 (3.8)** $p < 0.001$
Soc	22	**5.9 (2.6)** $p < 0.001$	12.5 (3.8) $p = 0.016$	**6.5 (2.6)** $p < 0.001$

Table S18. Observed and expected number of IPs for the dataset Italy, 2005. For each field, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

region	observed	by country	by city	by field
Abruzzo	26	**8.9 (3.2)** $p < 0.001$	17.9 (4) $p = 0.036$	**9.6 (3.3)** $p < 0.001$
Basilicata	6	**1 (1)** $p = 0.002$	2 (1.3) $p = 0.01$	**1 (1)** $p = 0.002$
Calabria	30	**5.4 (2.5)** $p < 0.001$	23.5 (4.5) $p = 0.095$	**5.9 (2.6)** $p < 0.001$
Campania	405	**83 (11)** $p < 0.001$	**296.9 (19.2)** $p < 0.001$	**90.3 (11.2)** $p < 0.001$
Emilia-Romagna	213	**71.7 (9.8)** $p < 0.001$	**152.9 (12.3)** $p < 0.001$	**75.8 (9.9)** $p < 0.001$
Friuli-Venezia Giulia	15	9 (3.2) $p = 0.05$	9.5 (2.9) $p = 0.051$	9.2 (3.2) $p = 0.058$
Lazio	379	**224.9 (22.4)** $p < 0.001$	**290.6 (23.2)** $p = 0.001$	**245 (21.8)** $p < 0.001$
Liguria	59	**19.8 (5.1)** $p < 0.001$	46.1 (6.9) $p = 0.043$	**21.6 (5.2)** $p < 0.001$
Lombardia	277	**136.8 (14.4)** $p < 0.001$	**223.9 (17.2)** $p = 0.002$	**150.4 (14.7)** $p < 0.001$
Marche	19	**5.8 (2.5)** $p < 0.001$	14.4 (3.6) $p = 0.125$	**6.2 (2.6)** $p < 0.001$
Molise	1	0.6 (0.8) $p = 0.47$	1.3 (1.1) $p = 0.743$	0.7 (0.9) $p = 0.498$
Piemonte	110	**49 (8.3)** $p < 0.001$	**78.8 (9)** $p = 0.001$	**52.9 (8.4)** $p < 0.001$
Puglia	222	**34.7 (6.7)** $p < 0.001$	**100 (9.8)** $p < 0.001$	**36.9 (6.8)** $p < 0.001$
Sardegna	159	**13.4 (4)** $p < 0.001$	124.5 (11.4) $p = 0.003$	**14.5 (4.1)** $p < 0.001$
Sicilia	495	**65 (9.5)** $p < 0.001$	**365.1 (20.2)** $p < 0.001$	**70.8 (9.6)** $p < 0.001$
Toscana	166	**62.6 (9)** $p < 0.001$	**120.5 (10.8)** $p < 0.001$	**66.9 (9.1)** $p < 0.001$
Trentino-Alto Adige	2	1.9 (1.4) $p = 0.557$	2 (1.3) $p = 0.602$	2 (1.5) $p = 0.596$
Umbria	30	**9.2 (3.3)** $p < 0.001$	17.1 (4) $p = 0.003$	**9.8 (3.4)** $p < 0.001$
Valle D'Aosta	0	0 (0.2) $p = 1$	0 (0) $-$	0 (0.2) $p = 1$
Veneto	88	**43.6 (7.6)** $p < 0.001$	66.7 (8) $p = 0.007$	**46.1 (7.7)** $p < 0.001$

Table S19. Observed and expected number of IPs for the dataset Italy, 2005. For each region, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

field	observed	by country	by city	by field
Agr	103	**30.9 (6.3)** $p < 0.001$	75.4 (10.2) $p = 0.008$	**29.4 (5.4)** $p < 0.001$
Bio	157	**50.4 (8.1)** $p < 0.001$	122.3 (13.2) $p = 0.008$	**55.6 (7.7)** $p < 0.001$
Chem	88	**20.5 (5)** $p < 0.001$	**52.6 (8.2)** $p < 0.001$	**20.6 (4.5)** $p < 0.001$
CivEng	118	**45.2 (8.1)** $p < 0.001$	94.2 (12.3) $p = 0.037$	**51.6 (7.8)** $p < 0.001$
Econ	127	**37.7 (6.8)** $p < 0.001$	**77.9 (10)** $p < 0.001$	**38 (6.3)** $p < 0.001$
Geo	5	2.7 (1.7) $p = 0.148$	7 (2.8) $p = 0.813$	3.1 (1.7) $p = 0.191$
Hum	90	**54.8 (8.4)** $p < 0.001$	120.4 (12.6) $p = 0.996$	**47.1 (6.9)** $p < 0.001$
IndEng	227	**79.9 (11.1)** $p < 0.001$	**157.1 (16.5)** $p < 0.001$	**86.3 (10)** $p < 0.001$
Law	154	**37.9 (6.8)** $p < 0.001$	**94.5 (11)** $p < 0.001$	**41.7 (6.5)** $p < 0.001$
Math	64	**21.5 (5.1)** $p < 0.001$	46.7 (7.6) $p = 0.019$	**23.4 (4.9)** $p < 0.001$
Med	861	**312.7 (24.9)** $p < 0.001$	**669.6 (36)** $p < 0.001$	**336 (20.9)** $p < 0.001$
Ped	109	**47.5 (7.9)** $p < 0.001$	100.1 (11.5) $p = 0.227$	**47.7 (7.1)** $p < 0.001$
Phys	36	**11.5 (3.6)** $p < 0.001$	27.3 (5.8) $p = 0.082$	**12.9 (3.6)** $p < 0.001$
Soc	19	**6.1 (2.6)** $p < 0.001$	13 (3.8) $p = 0.084$	**7.1 (2.7)** $p < 0.001$

Table S20. Observed and expected number of IPs for the dataset Italy, 2010. For each field, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

region	observed	by country	by city	by field
Abruzzo	25	**9 (3.3)** $p < 0.001$	17.7 (4) $p = 0.05$	**9.5 (3.3)** $p < 0.001$
Basilicata	5	0.8 (0.9) $p = 0.004$	1.8 (1.3) $p = 0.026$	0.8 (0.9) $p = 0.003$
Calabria	36	**6.8 (2.8)** $p < 0.001$	32.8 (5.4) $p = 0.296$	**7.2 (2.9)** $p < 0.001$
Campania	317	**70 (10)** $p < 0.001$	**238.8 (16.9)** $p < 0.001$	**74.2 (10)** $p < 0.001$
Emilia-Romagna	171	**64.4 (9.3)** $p < 0.001$	137.5 (11.6) $p = 0.003$	**66.6 (9.3)** $p < 0.001$
Friuli-Venezia Giulia	12	7 (2.8) $p = 0.063$	6.4 (2.4) $p = 0.025$	7.1 (2.8) $p = 0.066$
Lazio	309	**200 (21.1)** $p < 0.001$	254.9 (21.2) $p = 0.012$	**212.2 (20.2)** $p < 0.001$
Liguria	41	**13.2 (4.1)** $p < 0.001$	33.3 (5.8) $p = 0.108$	**14.1 (4.2)** $p < 0.001$
Lombardia	276	**140.9 (15)** $p < 0.001$	230.3 (17.8) $p = 0.009$	**150.5 (15)** $p < 0.001$
Marche	21	**6.4 (2.7)** $p < 0.001$	15.4 (3.7) $p = 0.086$	**6.6 (2.7)** $p < 0.001$
Molise	1	0.7 (0.9) $p = 0.509$	1.1 (1) $p = 0.696$	0.7 (0.9) $p = 0.522$
Piemonte	85	**47.2 (8.2)** $p < 0.001$	65.3 (8.1) $p = 0.012$	**49.6 (8.3)** $p < 0.001$
Puglia	136	**28.1 (6)** $p < 0.001$	**73.2 (8.3)** $p < 0.001$	**29.1 (6)** $p < 0.001$
Sardegna	141	**10.9 (3.6)** $p < 0.001$	**108 (10.5)** $p = 0.002$	**11.3 (3.6)** $p < 0.001$
Sicilia	366	**52.2 (8.4)** $p < 0.001$	**268.5 (16.7)** $p < 0.001$	**55.2 (8.4)** $p < 0.001$
Toscana	119	**49 (7.9)** $p < 0.001$	94 (9.6) $p = 0.008$	**51.2 (8)** $p < 0.001$
Trentino-Alto Adige	3	2.5 (1.6) $p = 0.441$	2.4 (1.5) $p = 0.43$	2.6 (1.7) $p = 0.475$
Umbria	27	**8.9 (3.3)** $p < 0.001$	17.6 (4.1) $p = 0.021$	**9.2 (3.3)** $p < 0.001$
Valle D'Aosta	0	0.1 (0.2) $p = 1$	0 (0) $-$	0.1 (0.2) $p = 1$
Veneto	67	**41.4 (7.4)** $p = 0.002$	59.1 (7.5) $p = 0.162$	42.8 (7.4) $p = 0.003$

Table S21. Observed and expected number of IPs for the dataset Italy, 2010. For each region, we report the observed number of pairs, as well as the expectation when randomizing last names either within academic system (by country), within regions (by region), etc. The number in parenthesis is the standard deviation. In bold values that have a $p-$value $\leq 0.05/$number of tests.

Region	2000	2005	2010	2015	% change (00-05)	% change (05-15)
Lombardia	6960	8514	8748	8318	22.33	-2.30
Lazio	6598	7789	7725	6895	18.05	-11.48
Emilia-Romagna	5179	5712	5423	5097	10.29	-10.77
Toscana	5019	5439	4850	4087	8.37	-24.86
Campania	4593	5559	5524	5074	21.03	-8.72
Sicilia	4489	4947	4678	4154	10.20	-16.03
Veneto	3386	3694	3622	3446	9.10	-6.71
Piemonte	3021	3276	3250	3149	8.44	-3.88
Puglia	2388	3317	3095	2826	38.90	-14.80
Liguria	1719	1709	1395	1294	-0.58	-24.28
Friuli-Venezia Giulia	1624	1724	1535	1398	6.16	-18.91
Sardegna	1580	1885	1714	1617	19.30	-14.22
Marche	1281	1501	1568	1420	17.17	-5.40
Abruzzo	1244	1584	1575	1430	27.33	-9.72
Umbria	1157	1251	1231	1174	8.12	-6.16
Calabria	863	1177	1372	1330	36.38	13.00
Trentino-Alto Adige	417	571	711	775	36.93	35.73
Basilicata	322	308	311	305	-4.35	-0.97
Molise	164	289	309	263	76.22	-9.00

Table S22. Number of professors by region for the periods considered in the study. The last two columns report the percent change between 2000 and 2005 and between 2005 and 2015.

Field	2000	2005	2010	2015	% change (00-05)	% change (05-15)
Med	9559	11242	10484	9270	17.61	-17.5
Hum	5215	5875	5461	4762	12.66	-18.9
Bio	4530	5207	4969	4619	14.94	-11.3
Hist-Ped-Psi	4164	4954	4859	4248	18.97	-14.2
Eng-Ind	4130	4923	5193	5208	19.20	5.8
Law	3721	4616	4785	4511	24.05	-2.3
Econ	3493	4412	4792	4669	26.31	5.8
Eng-Civ	3368	3827	3645	3317	13.63	-13.3
Chem	3140	3257	2994	2797	3.73	-14.1
Math	2926	3285	3270	2996	12.27	-8.8
Agr	2787	3215	3078	2927	15.36	-9.0
Phys	2408	2584	2326	2133	7.31	-17.4
Soc	1287	1611	1722	1639	25.17	1.7
Geo	1276	1280	1114	1006	0.31	-21.4

Table S23. Number of professors by field for the periods considered in the study. The last two columns report the percent change between 2000 and 2005 and between 2005 and 2015.

Institution	2000	2005	2010	2015	% change (00-05)	% change (05-15)
Roma La Sapienza	4246	4656	4230	3574	9.66	-23.2
Bologna	2828	3095	2929	2781	9.44	-10.2
Napoli Federico II	2673	2983	2683	2359	11.60	-20.9
Firenze	2184	2360	2057	1668	8.06	-29.3
Padova	2127	2248	2207	2058	5.69	-8.4
Milano	1968	2418	2198	1980	22.87	-18.1
Torino	1958	2106	2027	1947	7.56	-7.5
Pisa	1806	1833	1584	1428	1.50	-22.1
Palermo	1795	2018	1797	1554	12.42	-23.0
Genova	1719	1709	1395	1294	-0.58	-24.3
Bari	1478	1917	1676	1443	29.70	-24.7
Catania	1431	1592	1512	1300	11.25	-18.3
Cattolica Sacro Cuore	1315	1386	1412	1367	5.40	-1.4
Messina	1263	1337	1275	1141	5.86	-14.7
Pavia	1119	1132	1029	922	1.16	-18.6
Perugia	1117	1199	1172	1116	7.34	-6.9
Roma Tor Vergata	1061	1378	1542	1339	29.88	-2.8
Politecnico Milano	1014	1264	1360	1316	24.65	4.1
Parma	1012	1093	991	915	8.00	-16.3
Cagliari	1000	1194	1052	978	19.40	-18.1
Trieste	991	945	756	681	-4.64	-27.9
Siena	849	1036	943	723	22.03	-30.2
Politecnico Torino	785	838	813	804	6.75	-4.1
Campania	774	946	1028	959	22.22	1.4
Modena e Reggio Emilia	685	846	858	783	23.50	-7.5

Table S24. Number of professors by institutions for the periods considered in the study. The last two columns report the percent change between 2000 and 2005 and between 2005 and 2015.

References

1. Allesina S (2011) Measuring nepotism through shared last names: the case of Italian academia. *PLoS one* 6(8):e21160.
2. Durante R, Labartino G, Perotti R (2011) Academic dynasties: Decentralization, civic capital and familism in Italian universities. *National Bureau of Economic Research, Working paper* (No. 17572).
3. Zei G, Guglielmino CR, Siri E, Moroni A, Cavalli-Sforza LL (1983) Surnames as neutral alleles: observations in Sardinia. *Human Biology* pp. 357–365.
4. Kessler DA, Maruvka YE, Ouren J, Shnerb NM (2012) You name it–how memory and delay govern first name dynamics. *PloS one* 7(6):e38790.