The Italian primary school-size distribution and the city-size: a complex nexus

Alessandro Belmonte1, Riccardo Di Clemente1,2 & Sergey V. Buldyrev3

1IMT Institute for Advanced Studies Lucca, Piazza S. Ponziano 6, 55100, Lucca, Italy, 2Istituto dei Sistemi Complessi - CNR, Via dei Taurini 19, 00185 Rome, Italy, 3Department of Physics, Yeshiva University, New York, NY 10033, USA.

We characterize the statistical law according to which Italian primary school-size distributes. We find that the school-size can be approximated by a log-normal distribution, with a fat lower tail that collects a large number of very small schools. The upper tail of the school-size distribution decreases exponentially and the growth rates are distributed with a Laplace PDF. These distributions are similar to those observed for firms and are consistent with a Bose-Einstein preferential attachment process. The body of the distribution features a bimodal shape suggesting some source of heterogeneity in the school organization that we uncover by an in-depth analysis of the relation between schools-size and city-size. We propose a novel cluster methodology and a new spatial interaction approach among schools which outline the variety of policies implemented in Italy. Different regional policies are also discussed shedding lights on the relation between policy and geographical features.

There is a growing literature that nowadays sheds light on complexity features of social systems. Notable examples are firms and cities1–4, but many others have been proposed5,6. These systems are perpetually out of balance, where anything can happen within well-defined statistical laws7,8. Italian schools system seems to not escape from the same characterization and destiny. Despite several attempts of the Italian Ministry of education to reduce the class-size to comply with requirements stated by law9–11, no improvements have been made and still heterogeneity naturally keeps featuring the size distribution of the Italian primary schools.

In this paper we characterize the statistical law according to which the size of the Italian primary schools distributes. Using a database provided by the Italian Ministry of education in 2010 we show that the Italian primary school-size approximately distributes (in terms of students) as a log-normal distribution, with a fat lower tail that collects a large number of very small schools. Similarly to the firm-size12,13, we also find the upper tail to decrease exponentially. Moreover, the distribution of the school growth rates are distributed with a Lapussian probability density function (PDF). These distributions are consistent with the Bose-Einstein preferential attachment process. These results are found both at a provincial level and aggregate up to a national level, i.e. they are universal and do not depend on the geographic area.

The body of the distribution features a bimodal shape suggesting some source of heterogeneity in the school organization. The evidence of the bimodality underlies the interplay between different processes that define thresholds and boundaries that are very peculiar for the Italian primary school-size distribution. The question that we attempt to address in this paper is whether such regularity might depend on the complex geographic features of the country that in turn determines the way population (and in particular young people) distributes. We address these questions by analyzing in depth the spatial distribution of the schools, with particular regard to the areas where commuting is more effortful. We then proceed by investigating the complex link between schools and comuni, the smallest administrative centers in Italy, addressed by the introduction of a new binning methodology and a new spatial interaction analysis. Our conclusions indicate that the bimodality of the Italian primary school-size distribution is very likely to be due to a mixture of two laws governing small schools in the countryside and bigger ones in the cities, respectively.

Several examples of different regional schooling organizations are analyzed and discussed. We use GPS code positions for schools in two very different Italian Regions: Abruzzo and Tuscany. We introduce a measure of the average spatial interaction intensity between a school and the surrounding ones. We show that in regions like Abruzzo, that are mainly countryside, a policy favoring small schools uniformly distributed across small comuni has been implemented. Abruzzo small schools are generally located in low density populated zones, belonging to
small comuni. They are also very likely to have another small school as closest and the median distance between them is 8 km that is also the distance between small comuni. In Tuscany, a flatter region with a very densely populated area along the metropolitan area composed by Florence, Pisa and Livorno, we conversely find 1) a higher school density; 2) a stronger interaction between small and big schools; 3) a greater average proximity among schools. We address these stylized facts by arguing that the Italian primary school organization is basically the result of a random process in the school choice made by the parents. Primary education is not felt so much determinant to drive housing choice, like in US, because of the absence of any territorial constraint in school choice. Even if there is a certain mobility within a comune toward the most appealing schools, primary students generally do not move across comuni to attend a school. As a result, school density and school-size are prevalently driven by the population density and then by the geographical features of the territory. This generates a mixture in the schooling organization that turns into a bimodal shape distribution.

Results

Empirical evidence. We analyze a database on the primary school-size distribution in Italy that provides information on public and private schools, locations, and the number of classes and students enrolled. Data are collected, at the beginning of every academic year, by the Italian Ministry of education to be used for official notices. Our dataset covers $N = 17187$ primary schools in 2010 of which 91.31% were public. Almost seven thousands are located in mountain territories, (which represent the 40%) and 4101 are spread among administrative centers (provincial head-towns).

In Italy primary education is compulsory for children aged from six to ten. However, the parents are allowed to choose any school which they prefer, not necessarily the school closest to their home. We define $x_j$ the size of the school $i$ as the number of students enrolled in each school. Fig. 1(a) shows the histogram of the logarithm of the size of all primary schools in Italy. The red solid curve is the log-normal fit to the data

$$P(\ln x) = \exp \left( -\frac{(\ln x - \mu)^2}{2\sigma^2} \right) \frac{1}{\sqrt{2\pi\sigma}}$$

(1)

using the estimated parameters $\mu = 4.77(\mu/\ln(10) = 2.07)$, the mean of the $\ln x$ of the number of students per school, and its standard deviation, $\sigma = 0.85(\sigma/\ln(10) = 0.37)$. On a non-logarithmic scale, $\exp(\mu) = 118$ and $\exp(\sigma) = 2.34$ are called the location parameter and the scale parameter, respectively. The histogram in Fig. 1(a) suggests that log-normal fits data quite well. However, even a quick glance reveals that there are too many schools with a small dimension and much less mass in the upper tail with respect to the fit, suggesting that the number of students of the largest schools is smaller than would be the case for a true log-normal. In other words, similarly with firms-size distribution, tails seem to distribute differently from the log-normal distribution. Also Fig. 1(a) reveals a bimodal shape of the school-size distribution that we will extensively investigate below.

These findings can be detected in a more powerful way by plotting the histogram in a double logarithmic scale, comparing the tails of the log-normal distribution with those of the empirical one. We do this in Fig. 1(b) where y-axes represents the logarithm of the number of schools in the bins whereas in the x-axes the logarithm of the number of students stands. The empirical distribution differs significantly from the theoretical distribution which is a perfect parabola (the red curve), both in the tails and in the central bimodal part. A functional form of the right tail of the empirical distribution is revealed in the inset of Fig. 1(b) where we plot the cumulative distribution $P(X > x)$ of school sizes in semi-logarithmic scale. The straight line fit suggests that the right tail decreases exponentially $P(X > x) = \exp(-x \lambda)$ with a characteristic size $x = 1/\lambda$. This in turn means that there are approximately 120 students per school and also that the distribution of large schools declines exponentially. The exponential decay of the right tail of size distribution is consistent with Bose-Einstein preferential attachment process and is observed in the distribution of sizes of universities and firms.

Next we investigate the growth rates of elementary schools. Since temporal data are not currently available, we look at the single academic year, the 2010, and define the growth rate $g_i$ as follows:

$$g_i = \frac{x_i^5 - x_l^5}{\sum_{j=1}^5 x_j} = \lambda_i - \mu_i,$$

(2)

where $x_j^5$ stands for the number of students attending the $j$-th grade in school $i$, with $j \in [1, 5]$; $\lambda_i \equiv x_i^5/\sum_{j=1}^5 x_j$ is the fraction of students that have been enrolled in the first grade at six years old in school $i$, whereas $\mu_i \equiv x_i^5/\sum_{j=1}^5 x_j$ is the fraction of students that exit the school after the 5-th grade. Fig. 2(a) shows the relation between growth rate $g_i$ and school-size $x_j$. The numbers of grades $j$ provided by each school $i$, named $J_i$, is defined by the color gradient bar on the right side of the Fig. 2(a). Blue circles identify schools with $J_i = 1$. Such a group collects schools just established only providing the 1-st grade, i.e. with $\lambda_i = 1$ and $\mu_i = 0$, or that are going to close providing only the 5-th grade, i.e. with $\mu_i = 1$ and $\lambda_i = 0$. As soon as more grades are provided (colors switching to the warm side of the bar) schools tend to cluster around a null growth rate.

In Fig. 2(b) we investigate the growth/size relationship in depth. We demonstrate the applicability of the Gibrat law that states that the average growth rate is independent on the size. We define the average of the school size in each bin as $(x_i)$. The number of schools in each bin $n_i$ is represented by the size of the circle and the average number of grades $(J_i)$, is depicted according to the color gradient on the right side (the same of Fig. 2(a)). Independently from the size and the number of grades provided, schools do not grow on average. Nevertheless, we find more variability in smaller schools, apart from schools with $x_i < 10$, namely hospital-based schools mostly similar to one another, and the standard deviation of the growth rate $\sigma_\rho((x_i))$ is found to be decreasing as $(x_i)^{-\beta}$ with school-size by a rate of $\beta \approx 0.60$ (sub Fig. 2(b) inset). This is consistent with what has been found for other complex systems like firms or cities.

In Fig. 2(c) we study the growth rate distribution, where the probability density function $P(\rho = g_i)$ of growth rate has been plotted. The blue line represents the full sample (all the schools) distribution. Black and red colors identify the full capacity schools ($J_i = 5$) and the schools with $J_i < 5$, respectively. Regardless of the number of grades provided, the growth distribution underlines a Laplace PDF in the central part of the sample. The not-fully covered schools show a three peak behavior, where the left peak represents schools which are going to close, the central peak gathers schools that provide several grades but still in equilibrium phase, and the right peak is made up by the growing schools. Fig. 2(d) reports empirical tests for the tails of the PDF of the growth rate of the full sample (the upper one in blue, and the lower one in black). The asymptotic behavior of $g$ can be well approximated by power laws with exponents $\zeta = 4$ (the magenta dashed line), bringing support to the hypothesis of a stable dynamics of the process. All these findings are consistent with the Bose-Einstein process according to which the size distribution has an exponential right tail, a tent-shaped distributed growth rate and is observed also that the distribution of large schools declines exponentially. Therefore, the right tail of size distribution is consistent with Bose-Einstein preferential attachment process and is observed in the distribution of sizes of universities and firms.

City size and school size. Fig. 1(a) features the coexistence of two peaks, the first peak corresponding to log$_{10} x = m_1 = 1.7$ and the second one to log$_{10} x = m_2 = 2.3$, divided by a splitting point in
correspondence of $\log_{10} x_i \equiv m \approx 2.1$. The school sizes corresponding to these features are $\mu_1 = 10^{m_1} = 50$, $\mu_2 = 10^{m_2} = 200$, and $\mu = 10^{m} = 128$, with $\mu$ approximately equal to the average school size. 39% of the Italian primary schools distribute on the right of $\mu$, and more than 60% distribute on the left side. We test the alternative hypothesis of unimodality by looking at the probability that the number of schools in the two central bins $n_1, n_2$ are not smaller and the numbers of schools in the next three bins $n_3, n_4, n_5$ are not larger than a certain number $n^*$ provided that the standard deviation of the number of schools in these bins due to small statistics is $\sqrt{n}^*$. This probability is equal to $p(n^*) = \Pi, \text{erfc}\left(\left|\frac{n_n - n^*}{\sqrt{2n^*}}\right|/2\right)$ and it reaches maximum $p_{\text{max}} \approx 4 \times 10^{-15}$ at $n^* = 980$. Accordingly, we establish the bimodality with a very high confidence. This is also consistent with the bimodality index that we find to be equal to $\delta = (\mu_1 - \mu_2)/\sigma = 0.45^{23}$.

In this section we investigate the source of this heterogeneity that we find to be related to geographical and political features of the country and remarkably to the size of the comuni, the smallest administrative centers in Italy (information on comuni are provided by the Italian statistical institute, ISTAT), also here referred interchangeably as cities regardless of the size, $p_k$. A particular treatment is devoted to the nexus between the school-type (private versus public) and the geographical features of the comuni in the supplementary information, where we show that private schools are much less variable in size than public schools and have a narrow unimodal distribution peaked at approximately 100 students which contributes to the left peak of the entire school size distribution (Figure S1).

We denote a comune with letter $k = [1, \ldots, K]$. In 2010, $K = 8,092$ comuni have been counted in Italy, the 40% of which located in the mountains. We define $\mathcal{M}$ the set of mountain comuni and, accordingly, we call school $i$ a mountain school iff it resides in a comune $k \in \mathcal{M}$ (in the Supplementary we explain the laws according to which Italian comuni are classified as mountains). Each city $k$ has $n_k \geq 0$ schools (more than 15% of the cities have no schools) and population $p_k$, which distributes approximately as a log-normal PDF (see Fig. 3(a)), except for the right tail that is distributed according to a Zipf law, i.e. $p_k \sim r(p_k)^{-\xi}$ with slope $\xi \approx 1^{23,26-28}$. In Fig. 3(b) we find $\xi \approx 0.8$, in Italy, that is exactly the slope of the power law $p_k \sim r(n_k)^{-\xi}$ which links the population $p_k$ with the rank of this city in terms of number of schools $n_k$ (blue circles in Fig. 3b), i.e. $\xi \approx 0.8$. This means that the first city, Rome, has almost the double number of schools than Milan, and triple of Naples, while Rome has almost the double of inhabitants of Milan, and the triple of Naples. This amounts to say that $n_1$ is a good proxy for the city size.

We use the number of schools to assign comuni to different clusters $h \in [1, \ldots, H]$, according to

$$h = \{ k \in [1, \ldots, K] : 2^k - 1 \leq n_k < 2^k \}.$$  \hfill (3)

Accordingly, the first bin $h = 1$ gathers all the comuni with only one school; the second one collects all the comuni with $n_k = [2, 3]$, and so on. Though we find the average population $(p_h)_h$ to increase across different city-clusters $h$, less comuni $K_h$ lie in more populated clusters (the magenta and black lines in Fig. 3c). Interestingly, we find the interaction term $K_h(p_h)_h$ the green line in Fig. 3c), to distribute uniformly across different comuni-clusters, meaning that in small comuni with $n_k = 1$ live the same population than in bigger ones with much more schools.

Nevertheless, population is differently composed across city-clusters and a smaller fraction of young people is found in smaller comuni. To see that we also introduce a clustering of comuni according to population. Each comune is assigned to a cluster $c \in [1, \ldots, C]$ composed by all the comuni $k$ with population $p_k$ ranging from $\psi^{-1}$ to $\psi$, i.e.

$$c = \{ k \in [1, \ldots, K] : \psi^{-1} < p_k \leq \psi \}.$$  \hfill (4)

Setting the parameter $\psi = 2$ yields $C = 23$ clusters. Although the first seven sets are empty because no comuni in Italy has less than 128 inhabitants, the first (non-empty) cluster, $c = 8$, collects very small comuni with $p_k \in [128, 256]$. The last one, $c = 23$, conversely, is composed by the biggest cities with $p_k \in (2^{20}, 2^{21}]$. In Fig. 3(d) we plot
the average number of schools \( n_c \) (magenta line) and the average school-size \( x_i \) (the blue line) against the comuni size \( p_c \) for each non-empty cluster \( c \). We find that the average number of schools increases as a power law with exponent \( \beta = 0.88 \). This is consistent with the literature\(^2\,^3\,^26\,^28\) that has stressed the emergence of scale-invariant laws that characterize the city-size distribution. The average school-size increases with the population of the city reaching an asymptotic value at \( x_h \approx 230 \) students per school in the large cities. As expected, the interaction term, representing the average number of school-aged population in comuni belonging to cluster \( c \), \( \tilde{x}_c = \langle x_i \rangle_c \), behaves linearly with the comuni size except for small comuni with \( p_c < 10^3 \), for which the school-aged population constitutes a smaller fraction of the total population than in large cities.

In Fig. 4 we investigate the school-size distribution according to the comuni features. To this end, Fig. 4(a) draws the distributions of \( \log_{10} x_i \) conditionally on the number of schools, \( n_c \), in the comune \( k \). It yields 8 curves, one for each cluster \( h \) defined in Eq. 3. The first cluster is drawn in red (+) distributing all the schools located in comuni where only one school is provided. The orange line (○) distributes all the schools provided in comuni with two or three schools (i.e. \( h = 2 \)); and so on. The interesting point of Fig. 4(a) is that only the school-size distribution of the smallest comuni (with \( n_c = 1 \)) features a unimodal shape. The reason for that relies on the fact that comuni with only one school are geographically similar: they are the 57% of the total, with little more than 2000 inhabitants, the 81% of which are located in mountain territories.

The relationship between school-size and altitude is investigated in Fig. 4(b). Instead of conditioning on \( M \), here we propose a more consistent procedure according to which comuni are assigned to different bins on the basis of the altitude. In such a way, we can analyze comuni with 1,000 meters above the sea differently to those
with 600 meters of altitude that would be gathered in the same cluster \( M \) throwing away informative heterogeneity. It yields 5 bins: the first bin (drawn as a blue + line) gathers all the comuni whose altitude is lower than 125 meters above the see level (labeled 125 in Fig. 4(b)). Comuni with an altitude between 125 and 250 meters above the sea level composed the second bin (the green O line). These two distributions cluster around the second mode \( m_2 \) and in the Supplementary Information we additionally demonstrate that the hypothesis of bimodality can be rejected for the latter distribution.

However, the greater the altitude of the comuni the greater is the shift of the corresponding school-size distribution toward the small schools and the greater is the contribution of these comuni to the first mode \( m_1 \). Such a shift becomes evident for comuni with an altitude between 250 m and 500 m (red \( \Delta \) line). Comuni located between 500 m and 1000 m (cyan \( \times \) line) and above 1000 m (purple \( \square \) line) clusterize around \( m_1 \).

Even the largest cities are very different from each other in terms of their school size distribution. This heterogeneity is very likely to be driven by geographical features. We argue this point in Fig. 4(c), where we restrict our interest on the largest Italian cities belonging to cluster \( h = 8 \) (and to the first two bins in terms of altitude in Fig. 4(b)). These cities provide a number of schools \( n_h \) between 127 and 255, whose overall size distribution shows a three-peak shape with a third peak around 300 students absent in smaller cities (the bottom violet \( \Delta \) line in Fig. 4(a)). The presence of the three peaks around 100, 200 and 300 students might suggest the presence of architectural standards of school buildings supporting these particular sizes. However, by plotting the distribution by city we show that...
biggest Italian cities. With the exception of Rome, the hypothesis of unimodality may not be rejected in none of the biggest cities. In particular, flatter cities, such as Milano and Torino, mostly contribute to second mode $m_2$, whereas in Genova, Italian city built upon mountains that steeply ended on the sea, all the school-size distribution stands on the left side.

all the traces of trimodality disappear. In particular flatter cities, such as Milano and Torino, mostly contribute to second mode $m_2$, whereas in Genova, an Italian city built upon mountains that steeply slope towards the sea, the school-size distribution is unimodal contributing mostly to the first mode $m_1$.

Another way to look at the effect of geography on the comunal school-size is to compute the fraction of large schools within each comune $k$:

$$P_k(x_i > \bar{\mu} \mid \forall i \in k) \equiv \frac{n_k(x_i > \bar{\mu})}{n_k} \quad \forall i \in k,$$

where $n_k(x_i > \bar{\mu})$ stands for the number of schools that, in each comune $k$, are larger than the minimum $\bar{\mu}$ of the school-size distribution shown in Fig. 1(a). It can also be interpreted as the contribution rate of a comune $k$ to the second mode $m_2$. The upper panel of Fig. 5(a) diagrammatically explains how $P_k(\cdot)$ is computed.

We firstly study the relationship between $P_k(\cdot)$ and population, then looking at the spatial distribution across the Italy. In Fig. 5(a), we clusterize comuni according to Eq. 3, and for each bin $h$ we compute the average $\langle P_k(x_i > \bar{\mu} \mid \forall i \in k) \rangle_h$ and population $\langle P_k \rangle_h$. Interestingly, the plot shows that $P_k(\cdot)$ does not increase monotonically with population, demonstrating the existence of two city-patterns. More precisely, cities with less than $10^4$ inhabitants follow a pattern according to which the fraction of big schools, with $x_i > \bar{\mu}$, increases, on average, with population at a rate of $\beta_1 \approx 22$; in cities with more than $10^6$ we find the effect of population to be smaller, corresponding to $\beta_2 \approx 15$. For the cities with population between $10^4$ and $10^6$, the fraction of large schools does not increase with size suggesting that exogenous shocks such as altitude, rugged terrain and age might shift a city in this transition zone to either mode $m_1$ or mode $m_2$.

Overall, the distribution of $P_k(x_i > \bar{\mu} \mid \forall i \in k)$ is strongly correlated with the geographical features of the comuni territory. The map in Fig. 5(b) clarifies this point; all the mountain territories, Apennines that represent the spine of the peninsula and the Alps on the northern side, turns to be comuni with small schools, since the share of small schools in mountain comuni is equal to $P(x_i \leq \bar{\mu} \mid k \in M) = 0.72$. As soon as the probability to contribute to $m_2$ increases the colors get warmer; but this is very unlikely to be in the mountain territories, because less than $30\%$ of mountain comuni contribute to the antimode. Some regional patterns are also shown in the insets. The first upper panel depicts the area around Milan, which is surrounded by warm colors that mostly dye the Pianura Padana around. On the south side, Apennines approach and colors get blue with a lot of comuni with no schools (depicted in white). This pattern is more evident in the lower panel, which maps the region of Apulia, flat and mostly red, and the Basilicata on the left side, mountainous and mostly blue colored.

**Countryside versus dense regions.** In this last section, we bring more evidence on the effect of geography and comuni organization on the school-size by restricting our attention at two Italian regions: Abruzzo and Tuscany. But same results stand by looking at regions with the same geographical features. The two regions have very peculiar and representative geographical and administrative characteristics. Abruzzo is a mostly mountain region with a little flat seaside; it has four main head towns divided from each other by mountains. Conversely, Tuscany has many flat zones in the center...
is fatter in the former region. The cutoff for splitting the mixed
population distributions in Tuscany and Abruzzo. We clusterize
comune using the algorithm in Eq. 4. As Fig. 6(a) makes clear, comune
distribute approximately as a log-normal PDF in both regions, i.e. as
a parabola in a log-log scale (the green-
line stands for Abruzzo
PDF, the magenta-Δ for Tuscany). Nevertheless, Tuscany has bigger
cities. Figure 6(a) also shows the average number of schools as func-
tion of the population. The fact that
 cities. Figure 6(a) also shows the average number of schools as func-
tion of the population. The fact that
cities. Figure 6(a) also shows the average number of schools as func-
tion of the population. The fact that

distributions amounts to 128 in Abruzzo and 151 in Tuscany, and
31% of the schools are clustered in the second peak in the former
region; \( P(x_i > \mu_{TOS}) \forall i \in TOS = 0.38 \) in the latter.
In Fig. 6(c) we show, following the same clustering technique used
in Fig. 5(a), that in both regions the fraction of big schools within
comune \( k, P_k(x_i > \mu) \forall i \in k \), increases monotonically with respect to
the number of inhabitants for \( (p_k)_h < 2000 \). In this interval, a
comparison with figures for entire Italy, plotted in Fig. 5(a), reveals
that both regions follow the same national pattern. Yet, mountain
regions, such as Abruzzo, have a significantly smaller concentration
of big schools. In particular in Abruzzo, only about 1/10 of comuni
with just one school, with an average population of roughly 2000,
have a school with more than 125 students. In Tuscany, they are the
25%, about the same as national ratio. In larger comuni, with an
average population of 5000 and two schools provided (the second
bin), the probability of having big schools raises to 0.2 in Abruzzo,
still smaller than Tuscany where \( P(x_i > \mu_{TOS}) \forall i \in k \) for

Small schools are mainly located in the countryside, and for that
reason they cluster together, i.e. it is more likely to find a small school
near a small one. In Abruzzo this clustering effect is stronger than in
Tuscany. We investigate this point in Fig. 6(d), where we compute,
and plot on the x-axis, the cumulative probability \( P(x_i \leq x^*) \), as
function of \( x^* \), and the correspondent conditional probability
\( P(x_i \leq x^* | x_i \leq x^* \) on the y-axis, which is the fraction of schools with
the size smaller than \( x^* \) among the schools closest to a school of size
\( x^* \). This quantity is equal to 74% and 65% for \( x^* = \mu_{reg} \) in Abruzzo
and Tuscany respectively, meaning that there is a greater probability
that a small school matches with another of the same kind in the
former region. If the conditional probability were equal to the
cumulative, as indicated by the black line in Fig. 6(d), the sizes of

and the mountain areas shape the region boundaries. Remarkably, it
has a very high densely populated zone along the metropolitan area
composed by Florence, Pisa and Livorno.

These two regions also differ in terms of administrative organiza-
tions. Abruzzo favoring the establishment of comuni with a smaller
size due to the presence of mountains. Figure 6(a) shows the comuni
population distributions in Tuscany and Abruzzo. We clusterize
comuni using the algorithm in Eq. 4. As Fig. 6(a) makes clear, comuni
distribute approximately as a log-normal PDF in both regions, i.e. as
a parabola in a log-log scale (the green-
line stands for Abruzzo
PDF, the magenta-Δ for Tuscany). Nevertheless, Tuscany has bigger
cities. Figure 6(a) also shows the average number of schools as func-
tion of the population. The fact that

...
neighboring schools would be independent. This is not the case in either of the two regions. The probability that a small school has a smaller nearest neighbor is larger than the probability that any school is smaller than a given one. Indeed, the two curves (green for Abruzzo and magenta for Tuscany) are significantly above the 45 degree line for \( P(x_i < x^*) \), 0.6 in Tuscany and for \( P(x_i < x^*) \), 0.7 in Abruzzo.

These probability values roughly correspond to the probabilities \( P(x_i < x^*) \), \( \langle x_i \rangle \), in respectively Tuscany and Abruzzo, indicating that in both regions small schools are likely to belong to the small mountainous comuni, whose nearest neighbors are of the same class.

We further study the attraction intensity among small schools by disentangling the effect between the countryside and dense zones. To this end, we analyze the GPS location of the schools in the two regions and, for each school \( i \), we compute the number of schools \( n_i^\prime_m \) belonging within a circle of radius \( r_m \) centered at each school \( j \). We exclude from \( n_i^\prime_m \) all the schools which do not belong to Tuscany or Abruzzo, respectively. To eliminate the effect of region’s boundaries, we also compute areas \( D'_m \) as the areas of the intersections of these circles with a given region (Abruzzo or Tuscany). Thus \( D'_m \leq \pi (r'_m)^2 \), because these areas do not include the seaside and administrative territories of other regions. The difference between two subsequent circles yields the area of the annulus \( A'_m = D'_m - D'_{m-1} \). The density of schools in the area \( A'_m \) is then defined as:

\[
\rho'_m = \frac{n'_m - n'_{m-1}}{A'_m},
\]

and the average density of schools as function of a distance to a randomly selected school is

\[
\langle \rho'_m \rangle = \frac{\sum_n n'_m - \sum_n n'_{m-1}}{\sum_n A'_m}.
\]

In Fig. 7(a) red lines represent the average school-density around all the schools in Tuscany and Abruzzo, which are 472 in the former and

Figure 6 | Regional analysis. (a). The figure distributes the city-size in Abruzzo (○-green) and Tuscany (△-magenta) by plotting the number of comuni, \( K_c \), against the number of inhabitants, \( p_c \). Also shown is the average number of schools in a comune in Abruzzo and Tuscany, belonging to a bin \( c \) defined by Eq. 4, by the circled- and triangled-connected lines respectively. (b). School-size distribution in Abruzzo (○-green) and Tuscany (△-magenta). Both PDFs are approximately lognormal and bimodal with splitting point equal to 128 and 151 students per school respectively. (c). Average fraction of big schools in each comuni bin, defined by Eq. 3, in Abruzzo (○-green) and Tuscany (△-magenta). The plot shows that more populated comuni are, on average, more likely to have schools sized around \( m_c \), in both regions. Yet, in mountain regions, such as Abruzzo, smaller comuni have also smaller schools on average. (d). The conditional probability is plotted in the y-axis, for an arbitrary school size \( x^* \), as function of \( x^* \) against the cumulative probability \( P(x_i \leq x^*) \). The conditional probability is equal to the cumulative in correspondence of the black line. Along these points, there is no attraction between schools of the same size. This is not the case in both the two regions.
The analysis of schools growth rates highlights that the schools dynamics follows the Gibrat law, and both the growth rate distribution and the size distribution are consistent with a Bose-Einstein
phenomena and can be useful for a more effective policy making. Moreover, the availability of new longitudinal school data will be relevant to a more in-depth analysis and further discussions. In Fig. 8(b), we plot \((\rho_{\text{w}2})_k\) averaged over all the \(K = 8092\) Italian comuni as a function of the radius \(r_{m}\) that goes up to \(10^3\) Km across the entire Italy. The red line represents the average school-density among all the cities in Italy. On average, Italian comuni stand within very dense zones providing almost 1 school per \(10^3\) km². The dense zones generally last for \(10^3\) km and, after that, a smoothed depletion zone is experienced. However, the average distance between a comune \(k\) and a very large city with many schools is about \(100\) km, accordingly we see a second peak in the average school density at distance \(100\) km.

The full sample analysis basically averages heterogeneous characteristics that feature different types of comuni. The interaction among schools can be better understood by splitting the sample according to \(P_k(x_i > \tilde{\mu} \forall \text{comune}) \leq 1/2\). We see a second peak in Fig. 8(b), comuni with \(P_k(x_i > \tilde{\mu} \forall \text{comune}) = 1/2\), i.e. with predominantly small schools, are named \(M1\). The others, with predominantly big schools, are called \(M2\).

- \(M2\)-comuni, the blue line, are (on average) more likely to be surrounded by school-dense comuni. These comuni are located in densely populated areas (depicted in red in Fig. 5(b)) where the school density is large (1.3 schools stand on average within \(10^3\) km²). As far as the distance increases mountainous areas (and hence \(M1\)-comuni) are encountered and, as a result, the density of schools is found to drastically decrease.
- The green line describes instead comuni labeled \(M1\) where a smaller school density is found. Within \(10^3\) km, in fact, almost one school per \(20^3\) km² is encountered on average, about the half of what we find for the \(M2\)-comuni. This is

\[ Z_{km} = \{ x \in \{1, \ldots, K\} : d(x_i, x_j) \leq r_{km} \}. \]
because M1-comuni mainly stand along the countryside (those depicted in blue in Fig. 5(b)) where school density slowly increases with distance and reach a maximum at approximately 40 km, which can be interpreted as a typical distance to a densely populated area in a neighboring mountain valley. After this distance the density of schools around M1 and M2 comuni behave approximately in the same way.

1. Gabaix, X. Power Laws in Economics and Finance. *Annu. Rev. Econ.* 1, 255–93 (2009).
2. Gabaix, X. Zipf’s Law for Cities: An Explanation. *Q J Econ.* 114, 739–67 (1999).
3. Allen, P. M. Cities and regions as self-organizing systems: models of complexity. (Routledge, 1997).
4. Bak, P. How nature works (Oxford University Press, 1996).
5. Kauffman, S. *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, 1993).
6. Caves, R. E. Industrial Organization and New Findings on the Turnover and Mobility of Firms. *J Econ Lit.* 36, 1947–82 (1998).
7. Bak, P. *How nature works* (Oxford University Press, 1997).
8. Kauffman, S. At Home in the Universe: The Search for the Laws of Self-Organization and Complexity: The Search for the Laws of Self-Organization and Complexity. (Oxford University Press, 1996).
9. Fu, D. Decreto del Presidente della Repubblica del 20 marzo 2009. *Gazzetta Ufficiale* 81, 2 ( luglio 2009).
10. Fu, D. Decreto del Presidente della Repubblica (Paris: Recueil Sirey 1931).
11. Bak, P. *How nature works* (Oxford University Press, 1997).
12. De Wit, G. Firm Size Distributions: An Overview of Steady-State Distributions Resulting from Firm Dynamics Models. *Int J Ind Organ.* 23, 423–50 (2005).
13. Fu, D. et al. The growth of business firms: Theoretical framework and empirical evidence. *Proc Natl Acad Sci USA.* 102, 18801–18806 (2005).
14. Fu, D. et al. A Generalized Preferential Attachment Model for Business Firms Growth Rates-I. Empirical Evidence. *Eur Phys J B.* 57, 127–130 (2007).
15. Pammolli, F. et al. A generalized preferential attachment model for business firms growth rates: II. Mathematical treatment. *Eur Phys J B.* 57, 131–138 (2007).
16. Stanley, M. H. R. et al. Zipf distribution of U.S. Firm Sizes. *Science* 293, 1818–20 (2001).
17. Gibrat, R. *Gibrat’s Law for (All) Cities.* *Am Econ Rev.* 94, 1429–51 (2004).
18. Reed, W. J. The Pareto, Zipf and Other Power Laws. *Econ Lett.* 74, 15–9 (2001).
19. Black, S. E. Do better schools matter? Parental valuation of elementary education. *Q J Econ.* 114, 577–599 (1999).

**Acknowledgments**

We acknowledge Stefano Sebastio for many interesting and important discussions. We are also grateful to all the participants at the LIME seminars in IMT Lucca. S.V.B. thanks the Dr. Bernard W. Gamson Computational Science Center at Yeshiva College for support.

**Author contributions**

A.B. & R.D.C. analyzed the data. R.D.C. created the maps. A.B., R.C.D. & S.B. devised the research, wrote and revised the main manuscript text.

**Additional information**

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Belmonte, A., Di Clemente, R. & Buldyrev, S.V. The Italian primary school-size distribution and the city-size: a complex nexus. *Sci. Rep.* 4, 5301; DOI:10.1038/srep05301 (2014).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/