Supplementary Material
Modeling the polycentric transition of cities

Rémi Louf* and Marc Barthelemy†
Institut de Physique Théorique, CEA, CNRS-URA 2306, F-91191, Gif-sur-Yvette, France

CONTENTS

DETAILS ON THE SIMULATIONS

The simulation results presented in these supplementary materials are obtained in the following way. We first distribute randomly a number N_c of potential activity centers uniformly in the unit circle. Then, at each time step, we add a worker i at a random position in the circle, and compute the cost functions Z_{ij} for all potential activity centers j. We then connect the worker with the center maximizing the cost function, and add one to the value of the traffic corresponding to this center. We repeat this procedure until all workers are connected to a center.

Strictly speaking, traffic congestion do not arise only at the exact location of the activity center, but in an area of typical size R around that center. People who are not commuting to this specific center, but who have to travel close to it when commuting to their working place, might add to the traffic in the surroundings of this center. Therefore, we should consider the traffic $T_R(j)$ in the area around the center j, that is to say the number of people whose commute path crosses the area of size R around this center. For the sake of simplicity and clarity, we choose to ignore this in our model and leave it for further investigation.

ATTRACTIVITY-DRIVEN AND DISTANCE-DRIVEN POLYCENTRIC REGIMES

Definition and consequences

Fig. 1 in the main text suggests the existence of two different types of polycentric regimes that we called attractivity-driven and distance-driven polycentric regimes. In the attractivity-driven regime, individuals decide to work at the most attractive center, provided that the traffic is not too large. Therefore, as the traffic increases, new centers are going to appear in the decreasing order of attractivity. In the distance-driven regime, however, individuals tend to connect to the closest center, and thus the centers should appear in a random order.

More specifically, the attractivity-driven regime appears when the term $\frac{d_{ij}}{\ell}\left[1+\left(\frac{T(j)}{c}\right)^{\mu}\right]$ is negligible compared to the attractivity η_j at small values of traffic. Then, maximising the value of Z_{ij} amounts to maximising the attractivity of the center j. This will typically be the case when:

$$\ell \gg \ell^* = L \tag{1}$$

On the other hand, when $\ell \ll \ell^*$, the attractivity of the centers becomes irrelevant and maximising Z_{ij} amounts to connecting the closest center j to i. An important consequence is that the assumptions used to derive Eq. 6 and Eq. 10 in the main text are not valid in the distance-driven regime. In fact, there cannot be any stable monocentric state in the distance-driven regime and we start from the beginning (after a few iterations) with a polycentric state. Therefore, the difference between attractivity-driven and distance-driven polycentric regime draws the distinction between a system which grows from a single center, and a system where several centers appear from the beginning. This may be interpreted as the emergence of a single city when people can afford to travel a longer distance than the extension of the system (attractivity-driven), and of a system of cities when they cannot (distance-driven regime).

In between these two regime lies an intermediate regime, which is neither completely driven by space, not completely driven by the attractivity. This interplay between space and attractivity makes this intermediate regime difficult to explore analytically, we thus limit ourselves in this Letter to the two extreme regimes.

The transition is towards the attractivity-driven regime

We now show that if the parameters are such that there exists a monocentric state, then the system necessary evolves to an attractivity-driven polycentric structure. Let us assume that the system is in a monocentric state: P individuals commute to the center $j = 1$ characterised by the attractivity η_1. Let us assume that the monocentric state is unstable and that the $P+1$th individual i has two different possibilities. First, she could choose to go to the next most attractive subcenter characterised by the attractivity η_2, the largest attractivity among all the remaining potential centers. This translates into

$$Z_{i2} = \eta_2 - \frac{d_{i2}}{\ell} > Z_{i1} = \eta_1 - \frac{d_{i1}}{\ell}\left[1+\left(\frac{P}{c}\right)^{\mu}\right] \tag{2}$$

In general, the most attractive center is at a random fraction of the system size L, that is to say $d_{i1} \sim d_{i2} \sim L$. We are then led to Eq. 6 of the main text (see further in the Supplementary Material for a derivation)

$$P^* = c\left(\frac{\ell}{LN_c}\right)^{1/\mu} \tag{3}$$

The second possibility is for the individual to choose the closest center, characterized by η_j. We then have

$$Z_{ij} = \eta_j - \frac{d_{ij}}{\ell} > Z_{i1} = \eta_1 - \frac{d_{i1}}{\ell}\left[1+\left(\frac{P}{c}\right)^{\mu}\right] \tag{4}$$

Since j is the closest subcenter to i we can write $d_{ij} \sim L/\sqrt{N_c}$ and after a simple calculation we obtain a new value for P^* (in the limit of large N_c)

$$P'^* = c\left(\frac{\ell}{2L}\right)^{1/\mu} \tag{5}$$

We immediately observe here that for all values of the parameters, we always have $P^* < P'^*$. This result indicates that if a monocentric state exists, the transition is always towards an attractivity-driven polycentric structure. Or conversely, that there cannot be any stable monocentric state in the distance-driven regime.

Numerical verifications

Ordering in the apparition of centers

In order to verify the ordering in the apparition of secondary centers –and therefore the existence of two distinct regimes–, we increase the population in a certain configuration until all the the centers are populated. Each time a new center is populated, we note its attractivity. At the end, we compare the list obtained with the list of all the attractivities of the centers in decreasing order and compute Kendall's *tau* function defined by:

$$\tau = \frac{N_c - N_d}{\frac{n(n-1)}{2}} \tag{6}$$

where n is the number of centers, N_c the number of concordant pairs and N_d the number of discordant pairs. A pair i, j is said to be concordant if $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$. It is said to be discordant otherwise. One can see that if the orderings are identical we have $\tau = 1$, if they are completely opposite we have $\tau = -1$ and $\tau = 0$ if there is no correlation between the two lists.

Fig. 1 shows the evolution of τ with the value of ℓ. We see that we have $\tau = 0$ for small values of ℓ, indicating that the centers appear in a random order, and that we are in a distance-driven regime. On the other hand, for large values of ℓ we have $\tau = 1$, indicating that the centers appear in increasing order of their attractivity, thus that we are in an attractivity-driven regime. The study of the nature of the transition between the two regimes goes beyond the scope of this paper, and we leave it for further investigation.

Existence of a monocentric state

We confirm numerically the absence of a stable monocentric state in the distance-driven regime, and its existence in the attractivity-driven regime. We plot on Fig. 2 the value of P^*, the population at which a second center appears for different values of ℓ/L. We see that when $\ell < L$, $P^* \approx 1$, which means that there is no stable monocentric state. On the other hand, when $\ell > L$, P^* is different from 1, there exists a stable monocentric regime. We show later in these Supplementary Materials that we can compute an analytical expression for P^* when $\ell \gg L$.

SPATIAL COHERENCE

In the derivation of Eq. 13 in the main text, we assume that cities are systems such that $\ell > L$ (and Eq. 11 holds) while still exhibiting some spatial coherence. By spatial coherence, we mean that even if we are in a regime both controlled by the attractivity and space. In particular, it implies that the overlap between the basins of attraction of each subcenter is relatively small. In order to justify this assumption, we study numerically the behaviour of the average commute distance $\overline{d}$.

If we are in a regime where the individuals connect to the closest center, we have $\overline{d} \sim L/\sqrt{k}$ where k is the number of active centers. On the other hand, if we are in a regime where individuals tend to connect to the most attractive centers, then the average commute distance scales as $\overline{d} \sim L$. We thus plot $\sqrt{k}\ \overline{d}$ as a function of ℓ/L on Fig. 3. We can see that $\sqrt{k}\ \overline{d}$ increases as ℓ/L increases: people travel smaller distances in the distance-driven regime that in the attractivity-driven regime. Yet, for values of ℓ/L which are not too large (typically 10) the value of the average commute distance lies between the two extremes, closer to the value corresponding to the distance-driven regime, which supports the assumption that has been made in the main text to derive Eq. 13: up to large values of ℓ/L for which we know from figure 1 that we are in the attractivity driven regime, we can still use the assumption of the existence of well-defined basins of attraction for each subcenter with small overlap with each other.

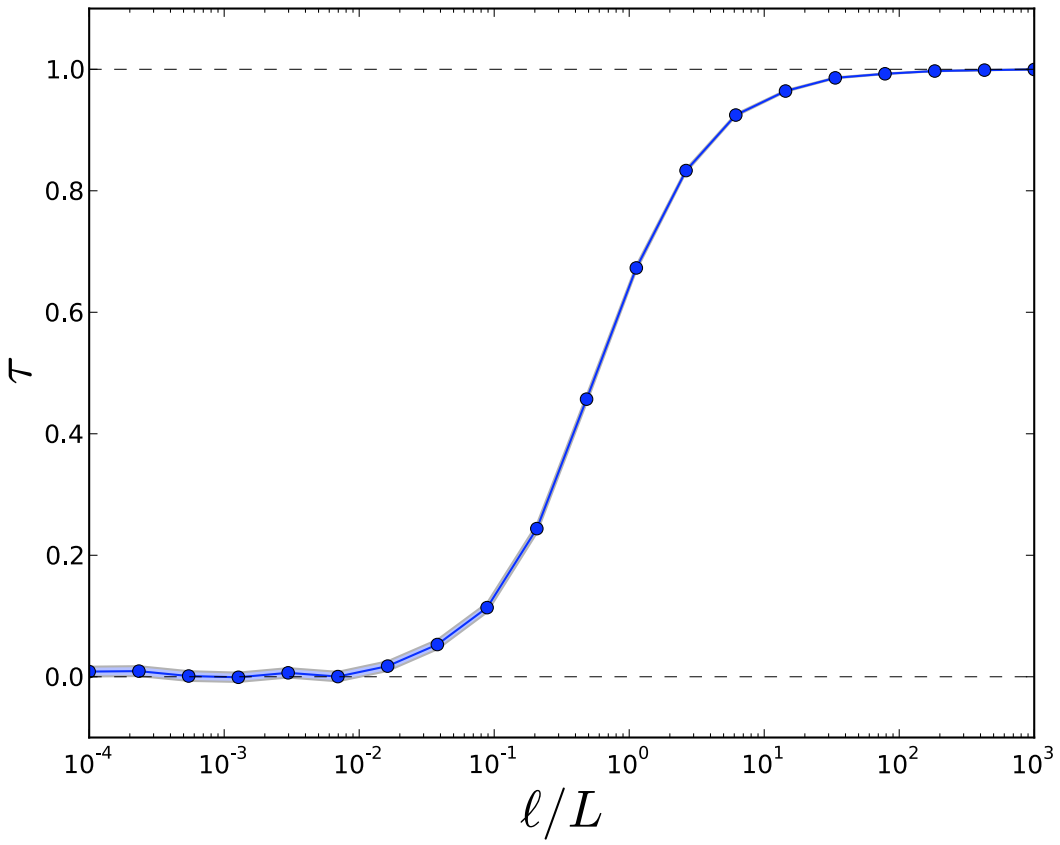

FIG. 1. Evolution of Kendall's τ with ℓ/L for $N_c = 10$, $\mu = 4$, and $c = 1$. The curve shows the average for 1000 configurations, and the standard deviation is shown by the shaded area. The dotted black lines highlight the extreme values $\tau = 0$ and $\tau = 1$.

COMPUTATION OF P^* IN THE ATTRACTIVITY-DRIVEN REGIME

Analytical derivation

In this section, we explicit the details of the calculation of the population P^* for which a first secondary center appears in a city within our model. We are here interested in the regime $\ell \gg L$ in which the monocentric state is stable for small values of the population (cf Fig.2). We assume that the system is in a situation where all the existing workers are connected to a single center 1, the most attractive of all the possible centers. We would like to know what happens when we add the next workers, that is to say whether a secondary center is going to appear, and if so, for what value of the population.
A secondary center k will appear when the $(P+1)^{th}$ worker i is added to the system if:

$$Z_{ik} \geq Z_{i1} \tag{7}$$

Following our assumptions, the traffic to 1 is given by $T(1) = P$ and the traffic to k by $T(k) = 0$. Therefore, the previous equation, in combination with the definition of Z_{ij} reads:

$$\eta_k - \frac{d_{ik}}{\ell} \geq \eta_1 - \frac{d_{i1}}{\ell}\left[1 + \left(\frac{P}{c}\right)^{\mu}\right] \tag{8}$$

Which can be written as:

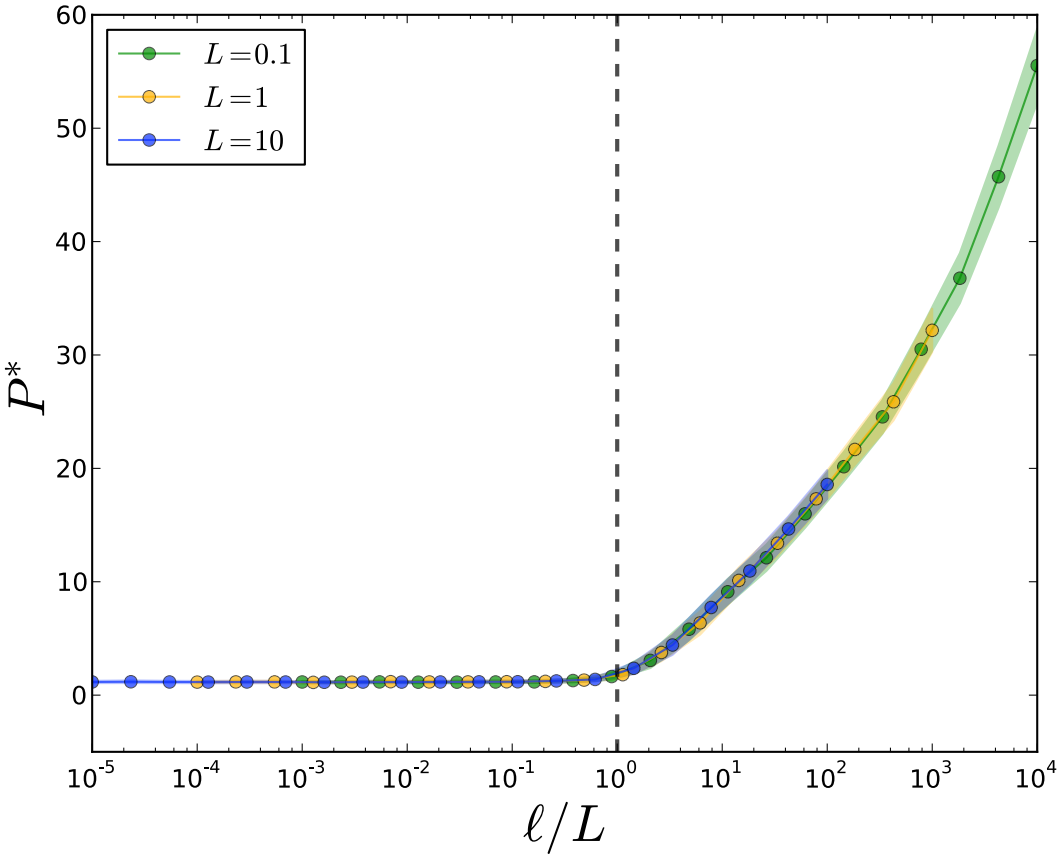

FIG. 2. Evolution of P^*, the value of population at which a second center appears, with ℓ/L for several values of L ($\mu = 4$,$N_c = 10$,$c = 1$, average of 10000 configurations). All the plots collapse on the same curve, and indicate that there is no stable monocentric state for $\ell/L < 1$.

$$\left(\frac{P}{c}\right)^{\mu} \geq \frac{\ell}{d_{i1}}\left[\eta_1 - \eta_k + \frac{d_{ik}}{\ell}\right] - 1 \tag{9}$$

Which is so far an exact result. In order to derive the average value of P^*, we make some further assumption. First, because households and activity centers are distributed at random, the distance to the different centers is taken to be approximately $d_{i1} \sim d_{ik} \sim L$, the typical size of the system. Also, we know from order statistics [S1] that in the case of a random variable η uniformly distributed in $[0, 1]$, such that $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_k$ we have

$$\overline{\eta_1 - \eta_k} = \frac{k-1}{N_c + 1}$$

on average. Therefore, the first secondary center to appear is necessarily the center k which maximises $\overline{\eta_1 - \eta_k}$, that is to say the second most attractive center with attractivity η_2. When $N_c \gg 1$ we find that a secondary center appears when the population P of the city is such that:

$$P \geq P^* = c\left(\frac{\ell}{LN_c}\right)^{1/\mu} \tag{10}$$

Thus, if the parameters are such that a monocentric configuration can exist in the first place, there will always be a value of the population over which a secondary center is going to appear.

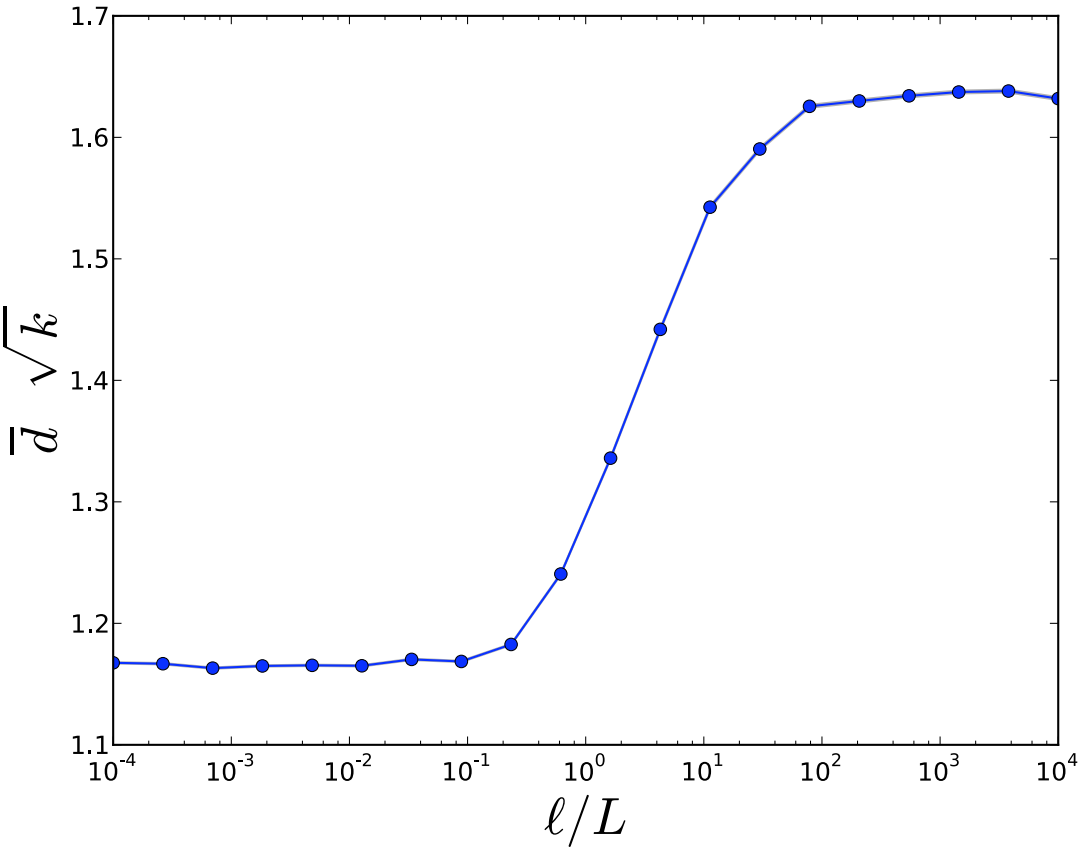

FIG. 3. Evolution of $\overline{d}\,\sqrt{k}$ with the ratio ℓ/L computed over 10000 generated systems ($\mu = 4$, $c = 100$, $N_c = 10$). We generate the systems such that all the centers are occupied, $k = N_c$. For small values of ℓ/L we are in the distance-driven regime, people commute to the closest center and the ratio is close to one. On the other hand, when ℓ/L is large, people commute to the most attractive center disregarding of the distance and the ratio is close to $\sqrt{N_c}L/2$. In the intermediate regime, the transition is relatively slow, meaning the system keeps a spatial coherence for a reasonable range of ℓ/L.

Numerical verification

In order to check the validity of our assumptions and thus of the previous formula, we numerically generate a lot of systems for different values of ℓ , L, N_c and μ and determine the population P^* at which a secondary center appears. We then plot the average of P^* as a function of $c\left(\frac{\ell}{LN_c}\right)^{1/\mu}$ (Fig. 4) and see that all the plots collapse on the same curve, confirming the above expression of P^*.

Existence of the monocentric regime

We saw previously that there always exists a value of the population size over which a secondary center appears. The same expression for P^* also implies that even in the regime $\ell \gg L$ there is a range of parameters for which the monocentric regime simply does not exist, that is when $P^* < 1$. Assuming N_c and μ are fixed, for every value of the capacity c there exists a value $\tilde{\ell}$ of ℓ given by :

$$\tilde{\ell} = L\ N_c\ \left(\frac{1}{c}\right)^{\mu} \tag{11}$$

under which there is no monocentric regime and the system exhibits a polycentric structure from the beginning of

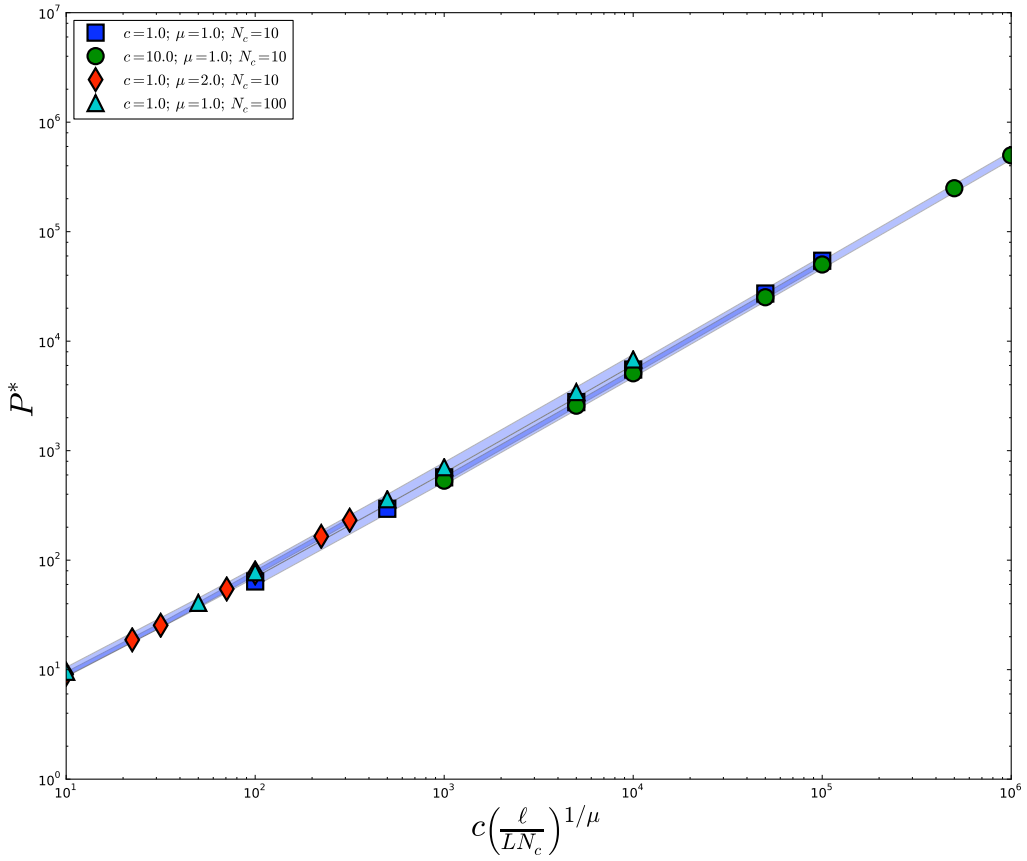

FIG. 4. Verification of the expression for P^*. Shows the average of the measured P^* on many simulated configurations of the system as a function of $c\left(\frac{\ell}{LN_c}\right)^{1/\mu}$, for different values of the parameters. The points all collapse on the same curve, confirming our expression for P^*.

its evolution. Therefore, we witness a transition from a monocentric to a polycentric regime only if the parameters are such that $\ell > \ell*$ and $\ell > \tilde{\ell}$

SOME DETAILS ABOUT P_k

Distribution of $T(j)$

In the derivation of P_k in the main text, we assume that the traffic $T(j)$ of the $(k-1)$ existing centers is rougly the same: $T(j) \sim P/(k-1)$. A simple argument supports this assertion: in order for the k^{th} center to appear, Eq. 7 in the main text must be satisfied, i.e. Z_{ik} must be larger than all the Z_{ij} $j \in [1, k-1]$. Seen differently, it means that as long as a center j is such that Z_{ij} is larger than Z_{ik}, people will connect to that center. Therefore, all the centers j are going to 'fill up' until Z_{ij} is too large. We can thus expect that, on average, all the centers will have the same traffic. Of course, the influence of space complicates this reasoning, and to convince the reader, we plot the distribution of $T(j)$ in the regime $\ell \gg L$ on Fig. 5.

We measure $\langle T(j) \rangle \approx 100 = P/N_c$ and the relative dispersion of the values is equal to 8%. Thus, our approximation of an equal repartition of people in the different centers in justified.

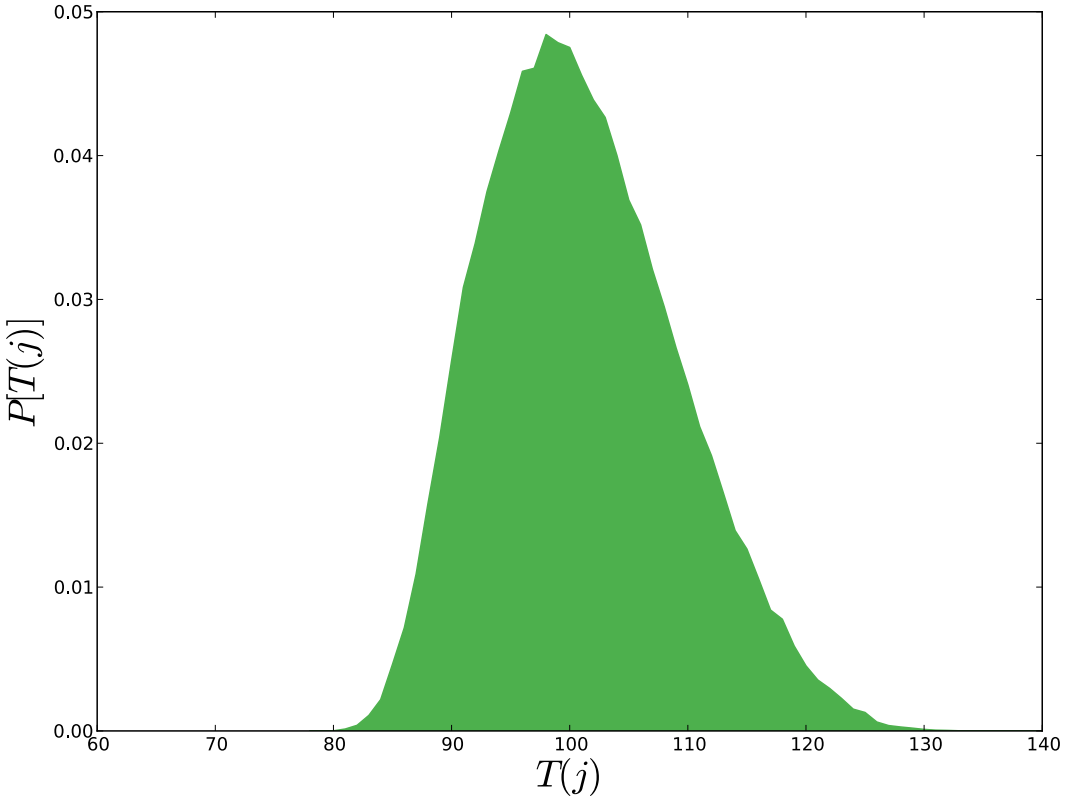

FIG. 5. Distribution of $T(j)$ obtained from the simulation of 10000 systems with the parameter $\mu = 4$, $c = 1$, $L = 1$, $\ell = 10^4$, $N_c = 10$ and $P = 10^3$

Sublinear relation between P_k and k

General argument

The expression given in the main text for P_k obviously relies on the fact that the η_j are uniformly distributed, and it is legitimate to wonder how the scaling would be modified it their distribution was different. It is usefull to note that the previously derived exponent is the result of two contributions. First, $(k-1)$ from the assumption that the commuters are approximately equally distributed among the existing centers. We do not expect this to change with η_j's distribution and thus we still have $T(j) \sim N/(k-1)$. Second, $(k-1)^{1/\mu}$ from the maximum of the difference $\eta_1 - \eta_k$, which obviously depends on the distribution of η. Nevertheless, we can reasonably expect $\eta_1 - \eta_k = f(k, N_c)$ where f is an increasing function of k and decreasing function of N_c. If we only consider the evolution of f with k, the function can be approximated around the current value of k as:

$f \sim (k)^\theta$

with $\theta > 0$. It follows that we locally have:

$$k \sim P^{\frac{\mu}{\mu+\theta}} \tag{12}$$

Which is still sublinear. The value of θ might change with k depending on the distribution, but it will always be positive. Thus, the dependence of k with P might not be a power-law, but it will definitely be sublinear.

Illustration in the case of a Pareto distribution.

In order to illustrate this point, we compute the average of $\max_{j\in[1,k-1]}(\eta_j) - \eta_k = \eta_1 - \eta_k$, that is to say $\mathbb{E}[\eta_1 - \eta_k]$ for the Pareto distribution.

We assume that the random variables possess a probability density function $f(x) = x^{-1-\mu}$ defined on $[1, +\infty[$, and a cumulative distribution function $F(x) = 1 - x^{-\mu}$ defined on the same interval. The probability density function of the r^{th} *largest* sampled value η_r is given by:

$$f_{\eta_r}(x) = \frac{N_c!}{(r-1)!(N_c-r)!}\left[1 - F(x)\right]^{r-1} F(x)^{N-r} x \tag{13}$$

And $\mathbb{E}(\eta_r)$ is given by:

$$\mathbb{E}[\eta_r] = \int_1^{+\infty} x\, f_{\eta_r}(x)\, dx \tag{14}$$

Which after calculations we find to be:

$$\mathbb{E}[\eta_r] = \frac{\Gamma(N+1)\,\Gamma\left(r - \frac{1}{\mu}\right)}{\mu\,\Gamma\left(N+1-\frac{1}{\mu}\right)\,\Gamma(r)} \tag{15}$$

Where we denote by Γ the usual Gamma function. Using the Stirling approximation for the terms which have a dependance in r, we show that in the leading order:

$$\mathbb{E}[\eta_r] = \frac{\Gamma(N+1)}{\mu\,\Gamma\left(N+1-\frac{1}{\mu}\right)}\, r^{-1/\mu} \tag{16}$$

$\mathbb{E}[\eta_r]$ is thus a decreasing function of r. Therefore, $\mathbb{E}[\eta_1 - \eta_r]$ is an increasing function of r. Thus, assuming that the values of η are drawn from a Pareto distribution, we still have a sublinear dependance of the number of subcenters in the population.

Numerical verification of the expression of P_k

We show in the main text that the value of the population P_k at which the k^{th} subcenter appears is given by:

$$P_k = P^*\,(k-1)^{\frac{\mu+1}{\mu}} \tag{17}$$

In order to verify this formula, we numerically generate many configurations of the system and we measure the mean P_k for several values of k and several values of μ. We then plot the measured average P_k/P^* as a function of $(k-1)^{\frac{\mu+1}{\mu}}$ (Fig. 6) and see that all the plots collapse on the same curve, confirming the accuracy of our expression for P_k.

DATA ANALYSIS

Definition of the number of subcenter

Extracting the number of activity centers from the Zip Code Business Patterns (ZBP) data is a priori a non-trivial task. ZBP data provide, among other things, the total number of employees per Zip Code Area in the United States every year between 1994 and 2010. Because Zip Code areas can have different areas, we first decide to normalise the

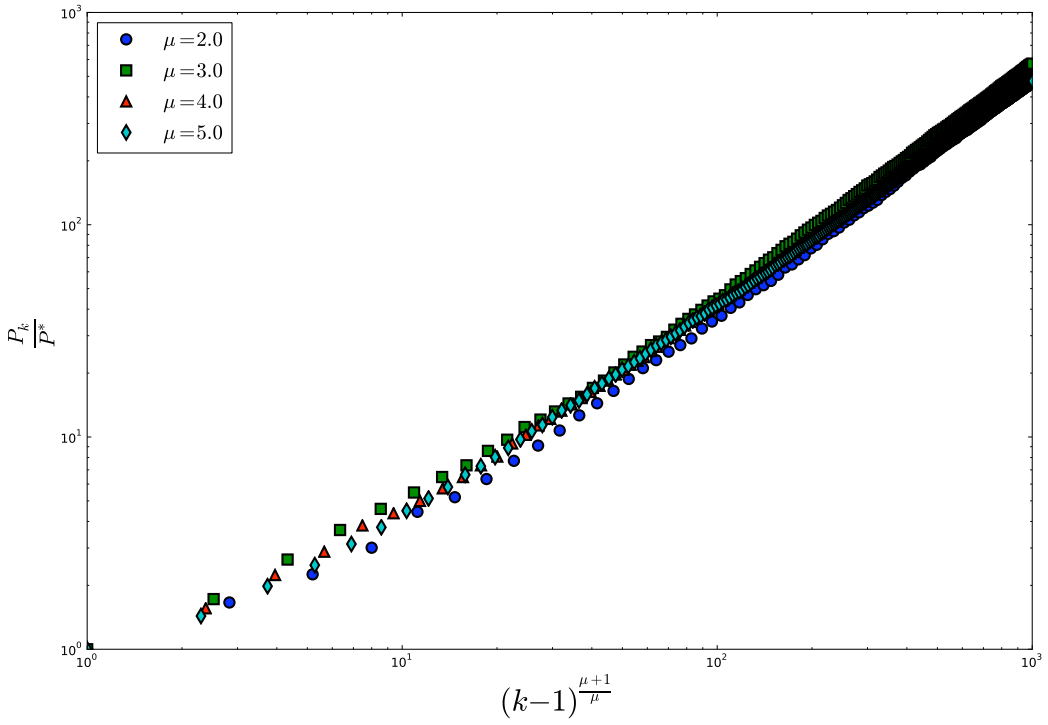

FIG. 6. Verification of the expression for P_k. Shows the average of the measured P_k on many simulated configurations of the system as a function of $(k-1)^{\frac{\mu}{1+\mu}}$, for different values of the parameters. The points all collapse on the same curve, confirming our expression for P_k. The plots have been obtained for $N_c = 10^3$, $L = 1$, $c = 1$ and $\ell = 10^5$, in the attractivity-driven polycentric regime.

number of employees to the area and obtain the employment density per Zip Code area.
If we further sort the employment densities in decreasing order and plot the employment densities of the different Zip Code areas as a function of their rank in the ordering (Supplementary Figure 3), we obtain a curves which decreases faster than exponentially for the first ranks, and then exponentially. We interpret this result as there existing a natural scale in the number of subcenters. Indeed, if the decrease is exponential, one can write for the employment density:

$$\rho_e \propto e^{-r/r_0} \tag{18}$$

where r_0 is the typical value of the rank in this situation, and that we would interpret here as the number of subcenters. However, the rank-plot is not strictly exponential and we therefore define the number of subcenters using a threshold value α. If we denote by ρ_0 the highest employee density in the city, we define ρ_m as:

$$\rho_m = \frac{\rho_0}{\alpha} \tag{19}$$

And the number of subcenters k is taken to be equal to the number of values of ρ_e such that $\rho_e \in [\rho_m, \rho_0]$. (Fig. 7)

In the case where the rank-plot is exactly exponential, we thus have:

$$\alpha = \frac{\rho_0}{\rho_m} = e^{m/r_0} \tag{20}$$

where m is the number of activity centers. It follows:

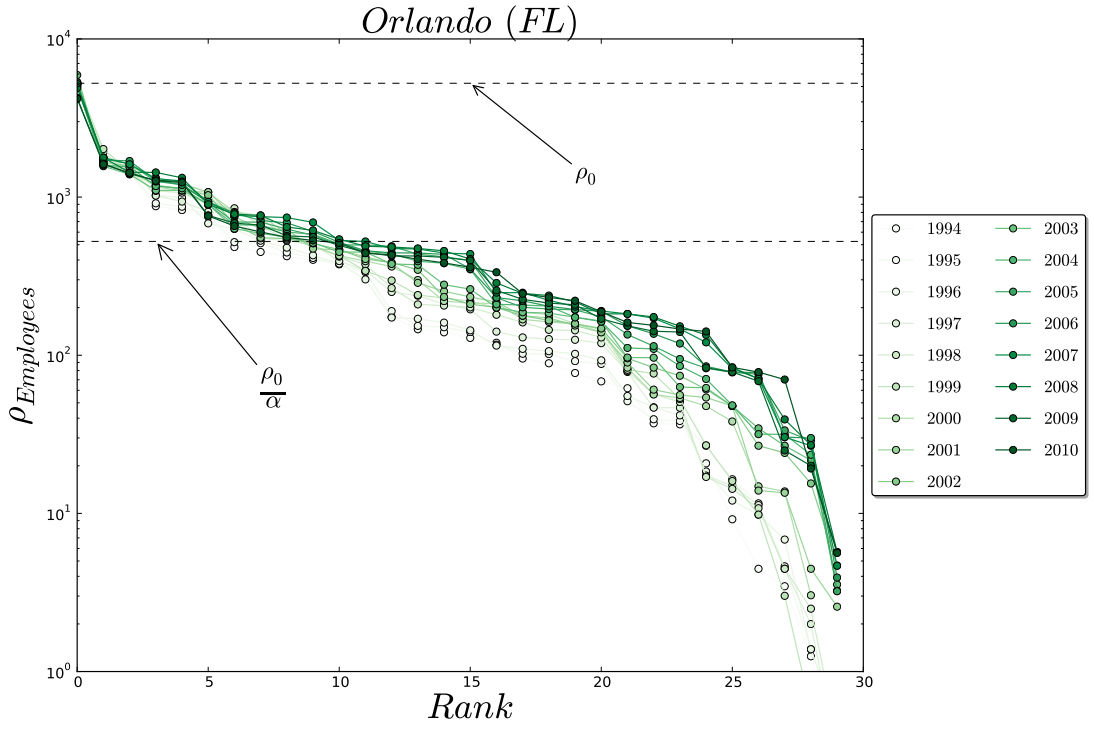

FIG. 7. Illustration of the definition of the number of subcenters. Shows the rank-plot of the employment density (in employees per square km) for Orlando(FL) between 1994 and 2010. We represent the largest employment density ρ_0 and the threshold value of employment ρ_0/α for 1994 with two dashed lines. The number of subcenters is taken to be the number of values of the employment density between those two values.

$$m = r_0 \ln(\alpha) \tag{21}$$

The order of magnitude of the number of centers is given by r_0 and relatively small variations of α therefore do not affect greatly the number of subcenters.

Some details

In order to plot the population as a function of the number of subcenters, we first extract the number of subcenters for every city in the dataset and for every year between 1994 and 2010, as indicated in the previous section.
We then apply the following treatment to the data:

- If a given city corresponds to a single Zip Code, we only keep one point for that city: its population and $k = 1$ center;

- We perform a Kolmogorov-Smirnov [S2] test between the rank-plots of the employment density in 1994 and 2010 for the remaining cities. If the probability that the two curves are different is greater than a certain threshold p_{KS}, we keep all the points for the city. Otherwise, we just keep a single point.

At the end of this process, we obtain points that can be understood as coming from different realisations of a city. The following section gives the robustness of the results with regards to the value of the threshold p_{KS} we choose.

Robustness of the empirical results

Choice of α

The empirical results presented in the main text will a priori depend on the value of the α we choose. We summarize the results obtained when fitting $k = f(\overline{P})$ assuming a power-law dependance and a multiplicative noise in Table 8. We see that for a reasonable range of values for α we always have a sublinear behaviour, and that given the estimated variance of the measured exponents, they are all compatible with each other.

α	5	10	15	20	30
β	1.53	1.56	1.35	1.37	1.37
σ_β	0.27	0.15	0.14	0.12	0.11
R^2	0.74	0.87	0.81	0.84	0.85

FIG. 8. Robustness of the empirical results. Shows the result for the fit for different values of α and $p_{KS} = 0.50$.

Choice of p_{KS}

It is also legitimate to wonder whether the empirical value we found for the exponent is affected by the value of the threshold p_{KS}. We summarize the results obtained with $\alpha = 10$ in Table 9. Again, we obtain a sublinear behaviour whatever the choice for the threshold p_{KS} and the different values are compatible with each other.

p_{KS}	0.5	0.6	0.7	0.8	0.9	1
β	1.56	1.50	1.52	1.56	1.64	1.66
σ_β	0.19	0.18	0.18	0.17	0.17	0.15
R^2	0.87	0.81	0.83	0.85	0.86	0.89

FIG. 9. Robustness of the empirical results. Shows the results for the fit for different values of p_{KS} and $\alpha = 10$.

Comments

The exponent derived from the plot on Fig. 3 in the main text is very sensitive to the first (k=1) and last point (k=26). Indeed, a fit excluding these points gives $\delta \sim 1.03 \pm 0.13$ ($R^2 = 0.83$). This exponent still agrees with the prediction of our model of $\beta < 1$, and would correspond to a very large value for $\mu \sim 32$ implying that the traffic dependence in Eq. 3 is a sharp threshold function. This indicates, if anything, that our empirical results should be confirmed through other studies based on different measures and datasets.

* remi.louf@cea.fr
† marc.barthelemy@cea.fr

[S1] H.A. David, H.N. Nagaraja, *Order statistics* (John Wiley & Sons, Inc., 1970).
[S2] F.J. Massey, *Journal of the American statistical Association* **46**, 68-78 (1951).