Supporting Information

Expert et al. 10.1073/pnas.1018962108

SI Text

This supplement to the paper "Uncovering space-independent communities in spatial networks" contains detailed information on relations between the spatial (Spa) null model and the standard Newman–Girvan (NG) null model, multiscale modularity, datasets, effect of bin size on results, optimal bipartitions of the mobile phone networks, and randomizations used to assess the significance of the results.

Spatial vs. NG Null Model. As discussed in the main text, the null model P_{ij} is a crucial ingredient of modularity defined as

$$Q = \frac{1}{2m} \sum_{C \in \mathcal{P}} \sum_{i,j \in C} [A_{ij} - P_{ij}]. \quad [S1]$$

The most standard choice is

$$P_{ij}^{\mathrm{NG}} = k_i k_j / 2m \quad [S2]$$

where the probability for two nodes to be connected is proportional to their degree. Our spatial null model incorporates nonstructural ingredients, namely, a dependence on the physical distance d_{ij} between two nodes

$$P_{ij}^{\mathrm{Spa}} = N_i N_j f(d_{ij}), \quad [S3]$$

where N_i is a measure of the importance of node i and where the deterrence function

$$f(d) = \frac{\sum_{i,j|d_{ij}=d} A_{ij}}{\sum_{i,j|d_{ij}=d} N_i N_j} \quad [S4]$$

is measured from the empirical data. This expression directly comes from the constraint

$$\sum_{i,j|d_{ij}=d} P_{ij}^{\mathrm{Spa}} = \sum_{i,j|d_{ij}=d} A_{ij} \quad [S5]$$

that the total weights between nodes at a certain distance is preserved. When analyzing the mobile phone network, we have taken N_i as the number of clients in commune i, in analogy with simple versions of gravity models. In that case, the above expression for $f(d)$ is a weighted average of the probability $A_{ij}/(N_i N_j)$ to have a call between clients in i and in j.

An interesting choice for N_i is to choose the degree itself (i.e., $N_i = k_i$), such that the set of null models

$$P_{ij}^{\mathrm{Spa}} = k_i k_j f(d_{ij}) \quad [S6]$$

includes P_{ij}^{NG}. Indeed, if $f(d)$ does not depend on d, one finds $f(d) = 1/2m$ and P_{ij}^{Spa} reduces to P_{ij}^{NG}.

Finally, we would like to briefly introduce a possible further generalization of the modularity function. Even when dealing with systems with strong spatial constraints, one might want to favor the importance of topological effects over spatial ones. A way of weighing both aspects in the modularity function is to introduce a mixing parameter, ξ, in order to interpolate between the two previous null models, $P_{ij}(\xi) = [\xi P_{ij}^{\mathrm{Spa}} + (1-\xi) P_{ij}^{\mathrm{NG}}]$.

Multiscale Modularity. Modularity optimization suffers from the limitation of producing one single partition, which is not satisfactory when dealing with multiscale or hierarchical systems, that is, systems made of (typically nested) modules at different scales. Different methods have been proposed to overcome this limitation (1). A first naive approach consists in reapplying modularity optimization on the communities found in the whole system. This approach provides a first guess but has the drawbacks of neglecting the global organization of the system when uncovering finer modules and of being unable to uncover coarser partitions than those obtained by the original modularity optimization. A second set of methods looks for local maxima of the modularity landscape, which has been shown to produce modules at different scales (2). Finally, a third class of methods is based on multiscale quality functions where a resolution parameter is incorporated such as to tune the characteristic size of the modules. A popular quantity is the parametric modularity introduced by Reichardt and Bornholdt (3),

$$Q_\gamma = \frac{1}{2m} \sum_{C \in \mathcal{P}} \sum_{i,j \in C} [A_{ij} - \gamma P_{ij}], \quad [S7]$$

where γ plays the role of a resolution parameter. Increasing γ tends to decrease the characteristic size of the modules in the optimal partition.

Because Q_{Spa} only differs from standard modularity by the choice of its null model, all three approaches can directly be applied to search for multiscale modules in spatial networks. Although an exhaustive analysis of such hierarchical organization is beyond the scope of the current work, we have performed a preliminary redecomposition of the two largest communities found in the mobile phone network (see Fig. S1). One finds $Q = 0.019$ and z score $= 91$ in the case of the northern community, and $Q = 0.064$ and z score $= 425$ in the case of the southern community (the z score is calculated for the ensemble of random networks where weights are randomized). Values of the z score are high but smaller than in the whole system (z score $= 803$). Moreover, the higher value observed in the southern community than in the northern community is expected due to the presence of bilingual Brussels in the former community. These results confirm that the linguistic division is the dominant factor, but also suggest that more regional factors (e.g., importance of local dialects, for instance in the Flemish community, cultural differences between cosmopolitan Brussels and the more rural Walloon, etc.) still play a role and lead to observable, finer subdivisions of the country.

Weights in the Belgian Mobile Phone Data. In our analysis of the mobile phone data, we have considered the fully connected matrix $\{A_{ij}\}_{i,j=1}^{571}$ where A_{ij} is the total number of calls between users in commune i and in commune j. Different types of weights could have been chosen for this aggregated network where nodes correspond to communes instead of individual users. In ref. 4, the authors focus on another network where weights $A_{ij}/(N_i N_j)$ correspond to the probability that users in i and j have called each other. This sensible choice gives, on average, the same importance to each commune and thus removes the effect of heterogeneity coming from different sizes of communes. In this article, we have instead preferred the first option, mainly for two reasons:

i. One of the aims of modularity is to properly account for the importance of nodes in the null model, thereby producing balanced modules in terms of this measure of importance (5). Because the definition of a proper null model is at the heart of this paper, we have preferred to preserve a strong heterogeneity (Fig. S2) in the system and to let the definition of modularity "deal with it."
ii. By focusing on a metanetwork where the weights of the links between communes is the sum over the links between their users, the same importance is given to each user. More importantly, modularity at the commune level is related to modularity at the user level: By optimizing the modularity of A_{ij}, one finds the best partition of the user network with the constraint that users in the same commune must be in the same community (6).

Size of the Communities. In the main text, we point out that the two largest communities found using the spatial null model account for about 75% of all communes. The remaining communes are assigned to 29 small communities, most of them close to Brussels. This fact can be attributed to the blindness of the algorithm we used to overlapping communities and the strong interaction of Flemish speaking communes with Brussels. To clearly illustrate this point, we plot in Fig. S3 the size of the communities found by the two null models. This plot clearly shows that the communities found by Spa other than the two largest are of negligible size. The sizes of the communities found by NG, on the other hand, are rather homogeneous. Fig. S3 also shows the size of the communities in terms of number of customers. Again the two largest communities found by Spa account for more than 70% of the customers, whereas NG again divides the Belgian population into communities fairly similar in size.

Binning Distance. The evaluation of $f(d)$ depends on the size of the bins used to measure distances. Two extreme cases are 1 and 200 km (the largest distances in Belgium are of the order of 300 km and we need at least two bins). To choose the most appropriate size for the bins in that range, we computed the deterrence function $f(d)$ and the partitions obtained for eight different bin sizes s: 1, 2, 5, 10, 20, 50, 100, and 200 km. The different deterrence functions are shown in Fig. S4. There is no clear discrimination for distances smaller than 5 km and the noise in the tail of the distribution is negligible from 20 km. Considering the size of Belgium, the number of communes (571), and the typical distance between them, a bin size of 5 km is a reasonable choice (7).

In order to support this choice, we have also computed the average normalized variation of information $\langle V(s)\rangle \equiv \frac{1}{N_s}\sum_{s'\neq s} V(s,s')$ between the partition at bin size s and those at other bin sizes (see Fig. S5). The size that is closest to all the others, thus the most representative of the system, is 5 km.

Gravity Model Benchmark: Averages. In the main text, we tested how close uncovered partitions were from the known underlying community structure as a function of the interaction strength and the density of links. The results for one single realization of the benchmarks were overwhelmingly in favor of Spa. Here we produce the same graphs, but averaged over 100 different realizations of the random networks, thus leading to a smoother surface, Fig. S6. We also present a "phase diagram" ($\lambda_{\text{different}}$, ρ) in which we highlight values where the partitioning ceases to be perfect (i.e., the normalized variation of information becomes larger than 0), Fig. S7. One observes that Spa offers a perfect reconstruction over a significantly broader range of parameters than NG. Visualizations of the benchmarks are shown in Fig. S8.

Bipartition of the Mobile Phone Data. From modularity, it is always possible to partition a network into two communities by assigning each node to a community according to its sign in the leading eigenvector of the modularity matrix. In this procedure, a negative second largest eigenvalue indicates that this bipartition is a good approximation to the full optimization of modularity (8). When applied to the Belgian mobile phone network, the second eigenvalue for NG modularity matrix is positive, contrary to Spa, thereby suggesting that a bipartition is a reasonable approximation to the full optimization for Spa. This observation is confirmed visually in Fig. S9, where NG picks Brussels and its neighborhood, and the rest of Belgium as the best bipartition, whereas Spa gives a North-South bipartition consistent with the linguistic bipartition of the country.

1. Lambiotte R (2010) Modeling and optimization in mobile, ad hoc and wireless networks (WiOpt). *Proceedings of the Eighth International Symposium on Multi-Scale Modularity in Complex Networks* 546–553; arXiv:1004.4268.
2. Sales-Pardo M, Guimera R, Moreira AA, Amaral LAN (2007) Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci USA* 104:15224–15229.
3. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Phys Rev E Stat Nonlin Soft Matter Phys* 74:016110.
4. Blondel VD, Krings G, Thomas I (Oct 4, 2010) Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies* (42), http://perso.uclouvain.be/gautier.krings/docs/EN_129_BruS42EN.pdf
5. Delvenne J-C, Yaliraki SN, Barahona M (2010) Stability of graph communities across time scales. *Proc Natl Acad Sci USA* 107:12755–12760.
6. Arenas A, Duch J, Fernández A, Gómez S (2007) Size reduction of complex networks preserving modularity. *New J Phys* 9:176.
7. Lambiotte R, et al. (2008) Geographical dispersal of mobile communication networks. *Physica A* 387:5317–5325.
8. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E Stat Nonlin Soft Matter Phys* 74:036104.

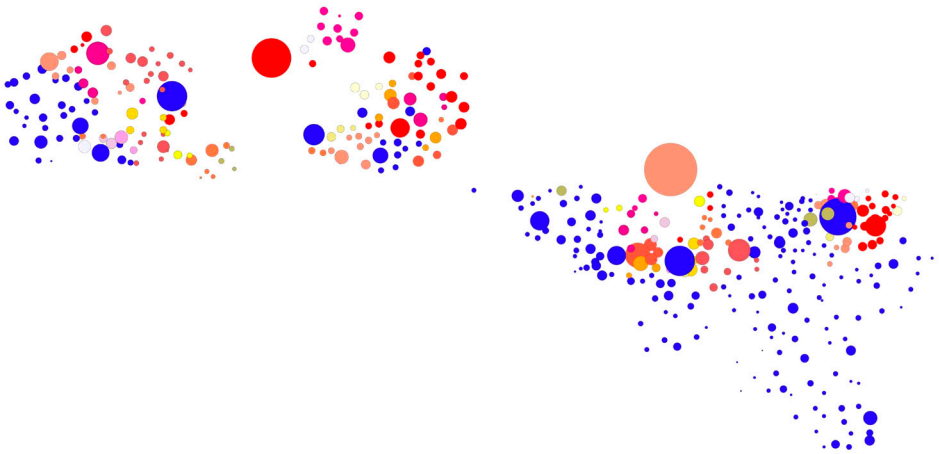

Fig. S1. Decomposition of the two main communities found in the Belgian mobile phone network by Spa into subcommunities (see main text). Each node represents a commune and its size is proportional to its number of clients N_i: (*Left*) northern community, (*Right*) southern community.

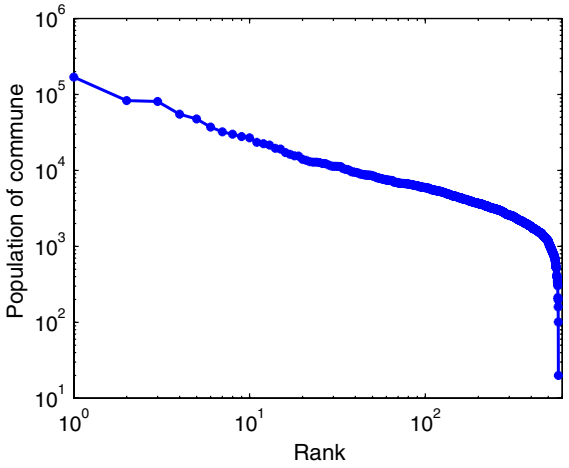

Fig. S2. Zipf plot of the commune sizes. The system is highly heterogeneous with several orders of magnitude between largest and smallest communes.

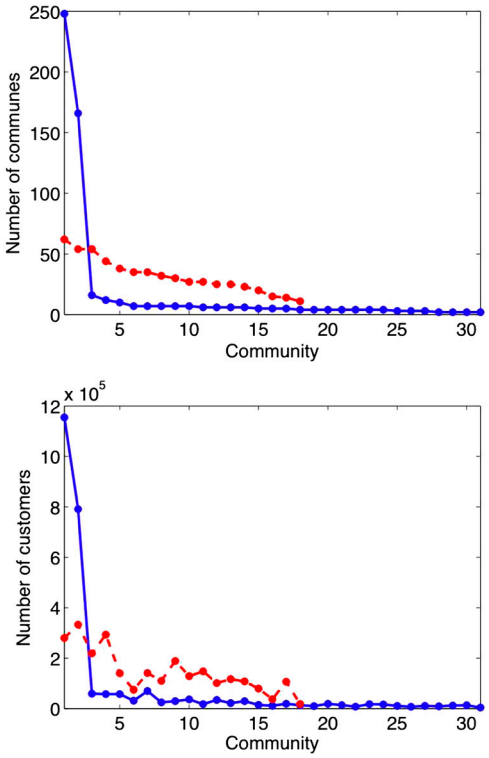

Fig. S3. Sizes of the communities found by Spa (full blue line) and NG (dashed red line). The size of each community is measured by the number of communes it contains (*Upper*) and by the number of customers living in it (*Lower*). In the partition found by Spa, two communities are large while the others are of negligible size. In the partition found by NG, all communities are of similar size. The labeling of the communes is the same in both figures.

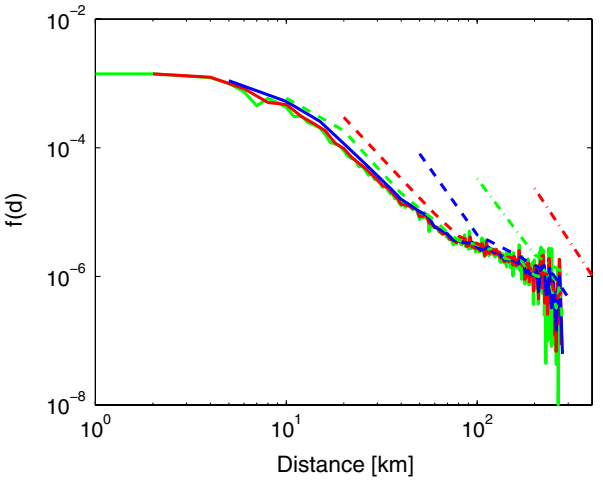

Fig. S4. Deterrence function $f(d)$ for different size of bins. Solid lines: green, 1 km; red, 2 km; blue, 5 km. Dashed lines: green, 10 km; red, 20 km; blue, 50 km. Dashed and dotted line: green, 100 km; red, 200 km.

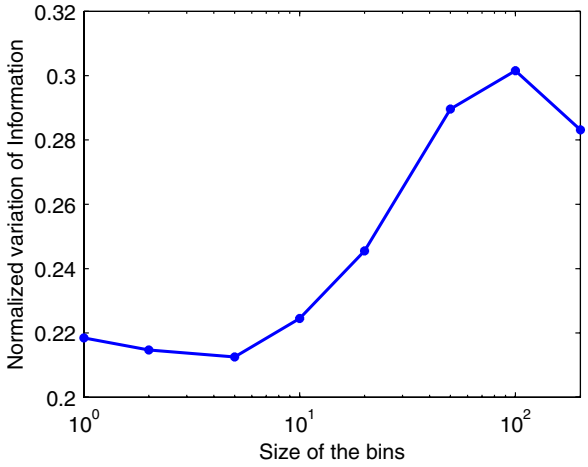

Fig. S5. Averaged normalized variation of information as a function of bin size. The minimum is reached at 5 km.

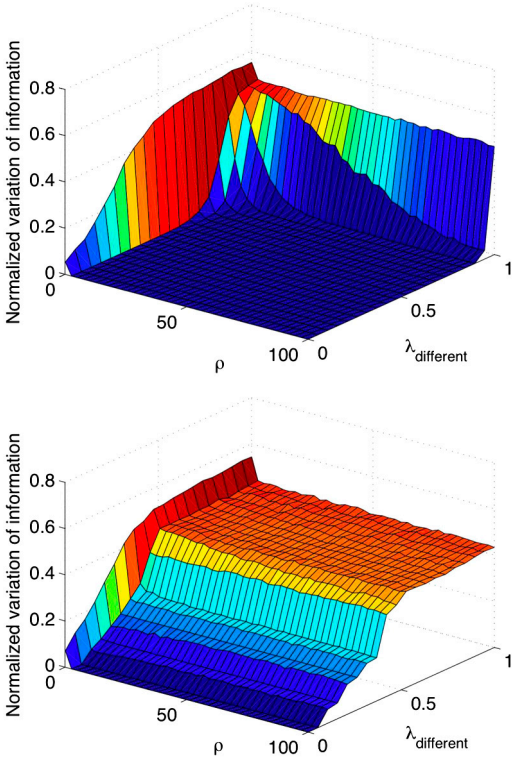

Fig. S6. Averaged normalized variation of information over the ($\lambda_{different}$, ρ) parameter space for Spa (*Upper*) and NG (*Lower*) when tested on the spatial benchmark. Spa is able to recover the correct communities over a wide range of parameters' values, whereas NG fails to find the correct communities almost as soon as the interaction $\lambda_{different}$ is turned on.

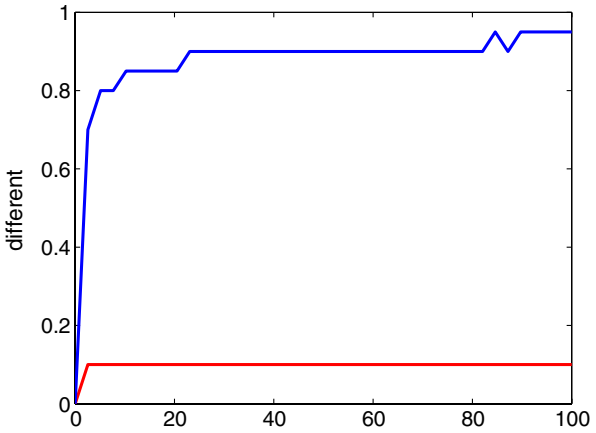

Fig. S7. Parameter space $\lambda_{\text{different}} - \rho$ separation. Solid red line, NG; solid blue line, Spa. Below the line, the partition found matches the real one; above, there are discrepancies. Spa is able to recover the original communities on a much larger range of parameters.

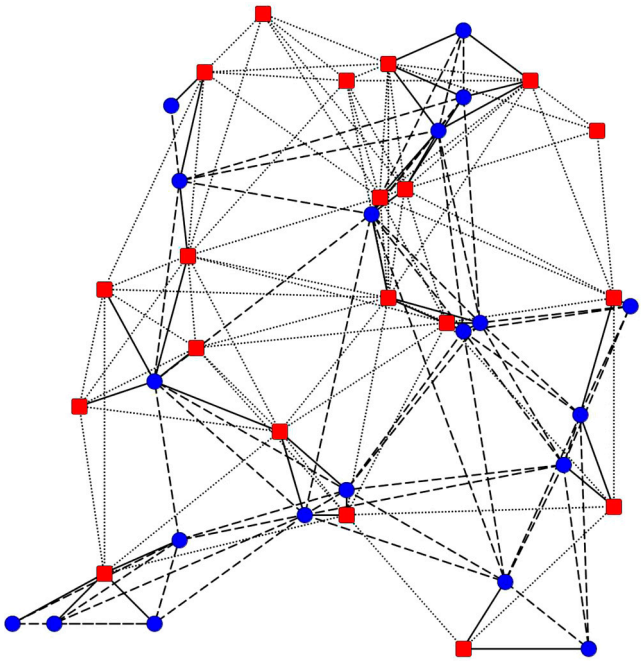

Fig. S8. Illustration of the probabilities used for a Gravity Model Benchmark made of 40 nodes with $\lambda_{\text{different}} = 0.5$. Only 20% of links with highest probability are represented. Hidden communities are represented by squares and circles. Different types of interactions are highlighted by their line style: red-red, dotted; blue-blue, dashed; red-blue, solid.

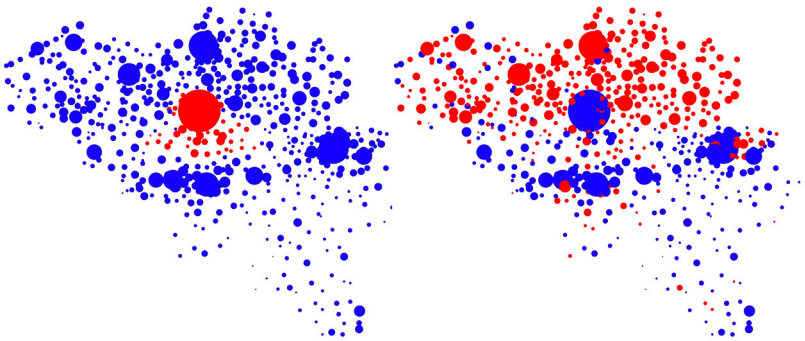

Fig. S9. Decomposition of a Belgian mobile phone network into two communities (see main text). Each node represents a commune and its size is proportional to its number of clients N_i. (*Left*) Bipartition found by the leading eigenvector of the NG modularity matrix, (*Right*) bipartition found by the leading eigenvector of the Spa modularity matrix.