PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Capitán JA, Bock Axelsen J, Manrubia S. 2015 New patterns in human biogeography revealed by networks of contacts between linguistic groups. *Proc. R. Soc. B* **282**: 20142947.
http://dx.doi.org/10.1098/rspb.2014.2947

Received: 2 December 2014
Accepted: 23 December 2014

Subject Areas:
ecology, computational biology, theoretical biology

Keywords:
spatial networks, food-web structure, interval graphs, species ranges, niche dimension

Author for correspondence:
José A. Capitán
e-mail: jcapitan@gmail.com

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2014.2947 or via http://rspb.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

New patterns in human biogeography revealed by networks of contacts between linguistic groups

José A. Capitán[1], Jacob Bock Axelsen[2] and Susanna Manrubia[1]

[1]Centro Nacional de Biotecnología (CSIC) c/Darwin 3, 28049 Madrid, Spain
[2]Department of Zoology, Tel Aviv University, Tel Aviv, Israel

JAC, 0000-0002-6245-0088

Human languages differ broadly in abundance and are distributed highly unevenly on the Earth. In many qualitative and quantitative aspects, they strongly resemble biodiversity distributions. An intriguing and previously unexplored issue is the architecture of the neighbouring relationships between human linguistic groups. Here we construct and characterize these networks of contacts and show that they represent a new kind of spatial network with uncommon structural properties. Remarkably, language networks share a meaningful property with food webs: both are quasi-interval graphs. In food webs, intervality is linked to the existence of a niche space of low dimensionality; in language networks, we show that the unique relevant variable is the area occupied by the speakers of a language. By means of a range model analogous to niche models in ecology, we show that a geometric restriction of perimeter covering by neighbouring linguistic domains explains the structural patterns observed. Our findings may be of interest in the development of models for language dynamics or regarding the propagation of cultural innovations. In relation to species distribution, they pose the question of whether the spatial features of species ranges share architecture, and eventually generating mechanism, with the distribution of human linguistic groups.

1. Introduction

Human diversity expresses itself in vastly different ways in terms of cultural traits, personal identity and relationships [1,2]. The human population is genetically quite similar [3], and technological advancements have led to personal mobility and communication on a global scale, but cultural diversity remains pervasive to a degree mostly comparable with biodiversity [4,5]. Most studies comparing cultural and biological diversity rely on the language spoken by individuals to define cultural groups; indeed, human languages are among the most easily quantifiable cultural traits, and display a variety that has intrigued scholars for centuries [6,7]. The analogy between biodiversity and human linguistic groups has led to the application of ecological methods to cultural data, often driven by the intuition that analogous features might arise from common generating processes [5,8,9]. Some remarkable patterns that both systems share are the latitude diversity gradient [10] and the language–area dependence [9,11], which mirrors the species–area relationship in ecology [12]. Also, the allometric dependence between the area occupied by the speakers of a language and the number of speakers of that language [13] finds a counterpart in the allometric dependence between species ranges and their abundances [14,15]. The specific history of particular species or languages has little to no influence in the construction of these collective statistical patterns.

A common representation of the relationships between biological species is in the form of food webs, where links stand for trophic interactions [16]. Food webs display the notable property of being graphs of high intervality, a feature that is deeply related to the existence of a niche space of low dimensionality

[16–18]. A graph is perfectly interval if their nodes can be ordered in such a way that the neighbours of any node occupy positions near that node, with no gaps left in between. A quasi-interval graph has a small number of gaps compared with suitable randomizations of its links. Intervality is deeply related to the existence of an almost one-dimensional configuration space [16,17], which implies that feeding relationships in food webs can be determined using a single species property (a 'niche' variable), and explains the success of models of food-web structure in accounting for many of their topological properties [19–23]. The probability that a food web is interval diminishes with its size [24], though larger food webs maintain high intervality in comparison with appropriate random models [17]. In humans, interactions occur at many levels, from individuals to confederations of countries, involving a hierarchy of connectivity patterns unfolding at different scales of space and time. Often, agents and their contacts can be depicted as networks embedded in space, a geometrical condition that affects their structure and evolution [25].

In this contribution, we construct and analyse networks of contacts between human linguistic groups, or language networks for short. Language networks are undirected, spatial networks that make explicit physical contacts between the areas in which different languages are spoken. We apply several measures commonly used in the analysis of complex networks and show that language networks are characterized by atypical topological properties, among which are a lognormal degree distribution, a one-dimensional local structure and quasi-intervality. The relevance of this latter property is assessed through the introduction of three different constructive hypotheses, which eventually allow us to conclude that the distribution of range sizes, together with a simple perimeter-covering rule among spatial neighbours, explains the patterns described. Nonetheless, we conjecture that the fundamental origin of quasi-intervality in language networks must arise from a non-trivial interaction between environmental variables and settlement of human groups, leaving an interesting question open in the area of linguistics.

2. Material and methods

(a) Language networks

Data on world languages have been obtained from the most comprehensive database currently available, the Ethnologue [26], which contains information on 6900 extant languages. The origin of data in the Ethnologue stems from a collection by SIL International (see http://www.ethnologue.com) and a map developed by Global Mapping International (World Language Mapping System, http://www.gmi.org/wlms/index.htm). In the Ethnologue we find a list of the spatial domains spanned by the speakers of each language and a centroid that is assigned to those domains, a point in latitude–longitude coordinates that best represents their average location. There is only one centroid per language, and centroids are the nodes of language networks. Since a language may have more than one disconnected domain where it is spoken (the sum of domain areas being the range, or total area, covered by the speakers of a language), centroids do not always fall inside speaking domains. Two centroids are connected if the two corresponding languages share boundaries in any of the areas where they are spoken. To avoid insularity effects, only languages found within the 100 largest landmasses of the Earth have been considered. Data in the Ethnologue describe the current distribution of languages, though the observed heterogeneity can be put in correspondence with different evolutionary states [9]. In order to further address the changes in language networks caused by the disappearance of languages and recent mechanisms such as colonization, we have studied how different structural modifications in language network definition affect the topological patterns described below. In addition, we have considered a different dataset regarding the distribution of native languages in North America prior to colonization, to check the robustness of our results (see the electronic supplementary material for further information).

(b) Definitions

The *linkage density* of an undirected network is defined as $z = 2L/N$, where N is the number of nodes and L is the total number of links.

A *planar network* can be drawn on the plane in such a way that its edges intersect only at their endpoints. Planarity has been checked in our networks through application of Kuratowski's theorem to find the minimum number of links that have to be eliminated to obtain a planar graph (see the electronic supplementary material).

The *degree distribution* $p(k)$ of a language network is the probability that a given linguistic group is in contact with k other linguistic groups. Though languages are often spoken in more than one isolated spatial domain, each border contact counts only once for every possible pair of languages.

The *average shortest path* length $\langle d \rangle$ for a network is the average over all possible node pairs of the minimum number of steps required to go from one node to another through existing links.

The *clustering coefficient* C_i of node i is defined as the number of connections between pairs of neighbours of i divided by the maximum value this quantity may take, $k_i(k_i - 1)/2$. The clustering coefficient of a network is the average over all its nodes, $\langle C \rangle = N^{-1} \sum_{i=1}^{N} C_i$. This quantity can be exactly calculated in some simple cases, as for regular networks (i.e. graphs for which linkage density z is uniform for all nodes) embedded in D dimensions, where

$$\langle C \rangle = \frac{\lambda(z - 2D)}{z - D} \tag{2.1}$$

and $\lambda = 3/4$.

A *perfectly interval directed network* admits a permutation of its nodes such that the k_i directed links of any given node i point to a subset of nodes labelled with consecutive indices [17]. This means that the corresponding adjacency matrix (a_{ij})—defined by the conditions $a_{ij} = 1$ if there exists a directed link from j to i and $a_{ij} = 0$ otherwise—has no gaps along its columns. If the network is undirected, it is perfectly interval if and only if there exists a node ordering such that the k_i connections of node i are restricted to k_i circle-neighbours nearby. Therefore, if the node at position $i + j$ is connected to i, so is $i + j - 1$ (and similarly for node $i - n$ and $i - n + 1$, respectively). This implies that the corresponding (symmetric) adjacency matrix has no gaps along its columns and rows.

The intervality of a network can be measured through the overall number of gaps G' along its columns. For a perfectly interval network, $G' = 0$. In particular, a one-dimensional regular graph is an example of a perfectly interval network. Conversely, the larger the number of gaps, the lower the intervality of the network. The overall number of gaps depends on the particular node labelling scheme. Hence there exists a node permutation $\sigma = (\sigma_i)$ such that $G'(\sigma)$ is minimal. This quantity is the empirical number of gaps of the network, $G = \min_\sigma G'(\sigma)$. We have used simulated annealing to estimate the minimum number of gaps G in language networks (see the electronic supplementary material).

3. Results

We constructed the world's network of contacts and extracted from it the subgraphs corresponding to Africa, Asia, Europe and the Americas. For each subgraph, a connected component analysis was performed. World languages can be grouped into a set of connected networks of variable size. To analyse topological properties, we have selected the 13 largest connected components, with sizes ranging from 2126 nodes (Continental Africa) to 33 (a group of languages located in the borders between Argentina, Bolivia and Paraguay—ABP borders; table 1). Figure 1*a* depicts a subset of the network obtained for New Guinea. Analogous results and maps of networks for all other cases studied are provided as electronic supplementary material.

(a) Topological properties

(i) Planarity

Despite the existence of strong spatial restrictions in our networks—a constraint that often facilitates planarity—language networks are non-planar in general. Small language networks are planar or almost planar, but the larger the network, the larger the fraction of non-planar links (table 1). Planarity is broken due to the variable number of isolated domains where a language is spoken and to multilingualism, which causes different domains to overlap (see the electronic supplementary material for details).

(ii) Degree distribution

The distribution $p(k)$ of the number of neighbours of a given linguistic group presents a peak at value 2–4 and a fat tail that extends to high degrees (up to 125 for Mandarin Chinese in Continental Asia). In all cases, the degree distribution of language networks is compatible with a discrete lognormal distribution. This means that most languages have a similar number of neighbours, but there is a small fraction of exceptions with a large number of connections. Figure 1*b* shows the degree distribution for New Guinea's network. In order to assess the likelihood that empirical degrees of nodes arise from independent trials of a lognormal distribution, we have compared this model with two others: an Erdős–Rényi model, characterized by a Poisson degree distribution, and a modified Watts–Strogatz model for which an analytical expression of its $p(k)$—based on the derivation for the original case [27]—has been calculated (see the electronic supplementary material). We have used maximum likelihood for parameter estimation and Akaike's information criterion for model comparison. The lognormal model is rejected only in one case (for New Guinea's degree distribution) at a 5% confidence level. Degree distributions for the remaining networks, together with parameter estimates from lognormal fits as well as the quantitative comparison between the models tested, can be found in the electronic supplementary material.

(iii) Average shortest path length

For each language network, we have calculated the empirical value of the average shortest path length $\langle d \rangle$, which has been compared with lengths rendered by different models for which the functional dependence between $\langle d \rangle$ and the size of the network N is known. Language networks are mostly compatible with two-dimensional, planar networks of similar average degree (square or hexagonal lattices; see the electronic supplementary material), which indicates that nodes are 'separated' on average as if linguistic domains were spatially distributed yielding a perfectly planar network of contacts. Two cases that show significant deviations are Continental Africa and Continental Asia, which actually contain the largest fraction of non-planar links (0.43 and 0.46, respectively) among all networks analysed. In agreement with this fact, they present average shortest paths well below the value expected for regular, planar networks with comparable linkage density z.

(iv) Clustering

The average clustering coefficient $\langle C \rangle$ obtained for language networks has been represented in figure 2 as a function of z. When the functional form (2.1) expected for regular networks is fitted to the data, we obtain a reasonable fit with parameters $D = 0.84$ (95% CI: (0.56, 1.12)) and $\lambda = 0.68$ (95% CI: (0.57, 0.80)). Therefore, language networks seem to behave locally as one-dimensional networks. This is remarkable considering that language networks are naturally embedded in the two-dimensional space, and points to a non-trivial reorganization of neighbouring relationships.

Clustering values are large when compared with random networks with the same linkage density, for which $\langle C^{\mathrm{rnd}} \rangle = z/N$. Hence, though we could not discard that an Erdős–Rényi model matched the degree distribution of New Guinea language network, the random model cannot explain the high clustering measured (table 1). In general, no model without spatial correlations can account for high values of $\langle C \rangle$ when z is low [25].

Contrary to what is observed for the shortest path length, the clustering analysis described above reveals that language networks exhibit local topological features compatible with those of one-dimensional regular networks. This suggests that our networks might be described using a reduced number of variables embedded in a low-dimensional space, as reported in previous work for food webs [16–18,22]. To substantiate this possibility we have quantified to what extent language networks are close to one-dimensional regular graphs by analysing their intervality.

(v) Language network intervality

The values of the empirical number of gaps G obtained for language networks as a measure of their degree of intervality are summarized in table 1. An example of a node ordering that minimizes the number of gaps for New Guinea is shown in figure 3*a* (other networks are provided in the electronic supplementary material).

(b) Assessment of the significance of intervality in language networks

The absolute number of gaps is not informative *per se* of the degree of intervality of a network, since G depends on the network size, on the number of links it has and, in general, on the precise connectivity pattern. Therefore, G has to be compared with appropriate models able to reveal whether the obtained value indeed originates from high intervality or whether it is a generic property of networks sharing some of the topological features described. In order to assess the significance of intervality levels in language networks, we have first devised two random models that

Table 1. Summary of relevant quantities for the largest connected components of each continent and for the whole continent (in bold face). Numbers in parentheses in the connected component column refer to the number of connected components found in each continent; numbers in parentheses in the nodes column indicate the number of isolated nodes. NP stands for 'non-planar', the fourth column thus showing the minimum number of links that have to be removed for the network to become planar; $\langle C \rangle$ is the clustering coefficient, and $\langle d \rangle$ is the average shortest path length. Quantities $\langle C^{\mathrm{RCM}} \rangle$ and $\langle d^{\mathrm{RCM}} \rangle$ correspond to the values yielded by the RCM, averaged over 1000 independent realizations. The optimal number of gaps for language networks (G) and the corresponding averages for the SRM, the PRM and the RCM are listed in the last four columns.

connected component	nodes, N	links, L	NP links	$\langle C \rangle$	$\langle C^{\mathrm{RCM}} \rangle$	$\langle d \rangle$	$\langle d^{\mathrm{RCM}} \rangle$	G	$\langle G^{\mathrm{SRM}} \rangle$	$\langle G^{\mathrm{PRM}} \rangle$	$\langle G^{\mathrm{RCM}} \rangle$
Cont. Africa	2126	6154	2667	0.55	0.42 ± 0.01	12.9	8.1 ± 2.0	74 018	280 449	1 139 565	82 413
Africa (8)	2443 (271)	6218									
Cont. Asia	1375	4093	1883	0.58	0.37 ± 0.01	7.73	6.8 ± 1.6	52 886	293 386	465 972	90 589
New Guinea island	663	1543	307	0.48	0.44 ± 0.02	13.5	9.0 ± 2.1	8737	20 018	65 176	8247
Australia	99	176	3	0.39	0.43 ± 0.04	6.10	5.1 ± 1.2	517	1049	888	466
Sulawesi island	64	121	2	0.58	0.55 ± 0.04	5.09	4.7 ± 1.0	172	401	480	192
Luzon island	56	140	10	0.58	0.56 ± 0.03	2.63	2.3 ± 0.3	476	628	602	468
Asia (31)	3739 (1270)	6444									
Cont. Europe	231	547	67	0.57	0.55 ± 0.02	5.04	6.9 ± 1.8	2336	13 452	7296	2840
Europe (3)	285 (43)	557									
Cont. N. America	171	230	1	0.42	0.32 ± 0.03	2.42	2.9 ± 0.2	4387	4607	4765	5500
Mexico-1	68	120	1	0.49	0.54 ± 0.04	7.24	5.9 ± 1.1	140	232	473	134
Yucatan peninsula	50	111	3	0.54	0.57 ± 0.04	3.70	2.9 ± 0.5	214	455	363	222
Mexico-2	39	71	0	0.53	0.57 ± 0.05	4.81	3.6 ± 0.7	86	155	194	73
N. America (30)	609 (161)	660									
Cont. S. America	234	399	3	0.36	0.40 ± 0.03	11.9	8.7 ± 1.9	1229	1728	2821	1325
ABP borders	33	59	1	0.43	0.57 ± 0.06	3.42	2.8 ± 0.5	100	132	156	86
S. America (34)	549 (148)	583									

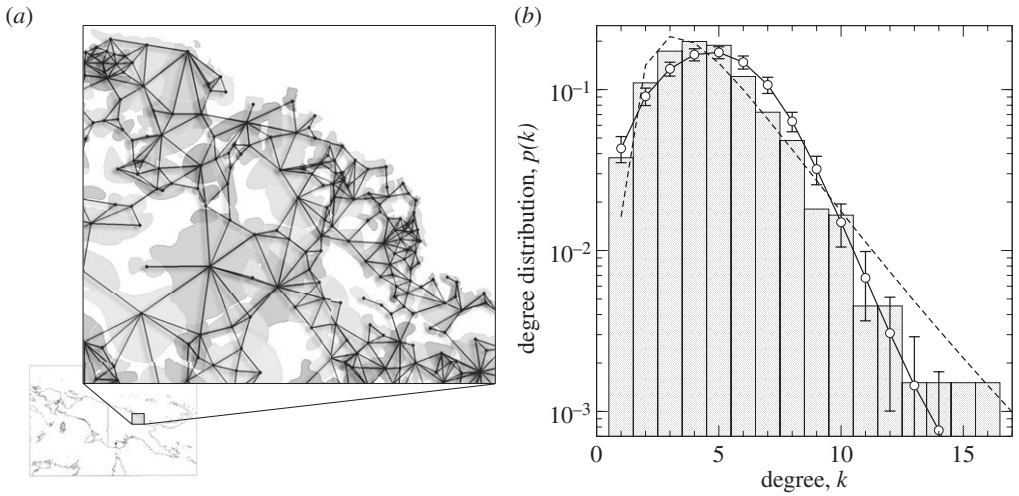

Figure 1. Network of contacts between neighbouring linguistic groups. (*a*) Part of the network corresponding to New Guinea, indicating its localization on the island. (*b*) Degree distribution of the empirical network (bars), maximum-likelihood fit to a lognormal model (dashed line) and average values obtained with the RCM (circles) with its standard deviation (error bars) calculated by averaging over 1000 model realizations. New Guinea has been chosen as a representative example due to its size—large enough to yield good statistical power but manageable so as to produce clear illustrations. There is no other particular feature that singles it out from the set of networks analysed.

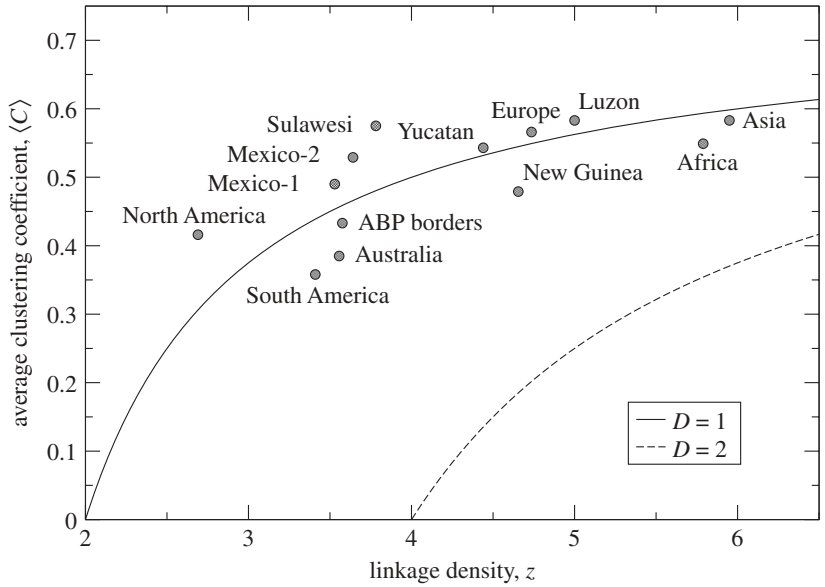

Figure 2. Average clustering coefficient of language networks versus linkage density *z*. The solid line corresponds to the exact coefficient of a regular, one-dimensional network, and the dashed line is the coefficient of a regular, two-dimensional network. Clustering coefficients of language networks fall close to the one-dimensional case.

conserve the degree distribution plus another *a priori* relevant ingredient: the spatial random model (SRM) and the planar random model (PRM). These models fail at recovering, among others, the intervality of language networks. Finally, and inspired by niche models in ecology, we introduce the range contact model (RCM), which is shown to accurately reproduce the structural patterns observed.

(i) Spatial random model

Let us hypothesize that the topological structure of language networks arises from local spatial restrictions in such a way that links can only be established between nodes (centroids) that are at a certain mutual distance on Earth's surface. For this model, we have thus chosen to preserve, in addition to the degree distribution, the empirically obtained distribution of physical distances between pairs of centroids. These empirical distributions are compatible with lognormal distributions in most cases, thus implying the existence of a typical distance for linkage but also a non-negligible probability that distant centroids are linked. The preservation of the distance distribution is a qualitative way to account for the restrictions imposed by a two-dimensional space—it seems unreasonable that links can be drawn arbitrarily between centroids regardless of their mutual separation. We performed $50L$ link rewirings to randomize language networks under the two previous assumptions. Then, we estimated the minimum number of gaps G^{SRM} for the network so obtained, and repeated for 500 independent realizations. The distribution of G^{SRM} values takes a Gaussian shape (figure 3*e*,*f*; averages are reported in table 1) that has been used to estimate the probability p that G^{SRM} is smaller than the empirical number of gaps G. There is only one instance where we cannot reject this hypothesis at a 1–99% confidence interval: ABP borders (see the electronic supplementary material).

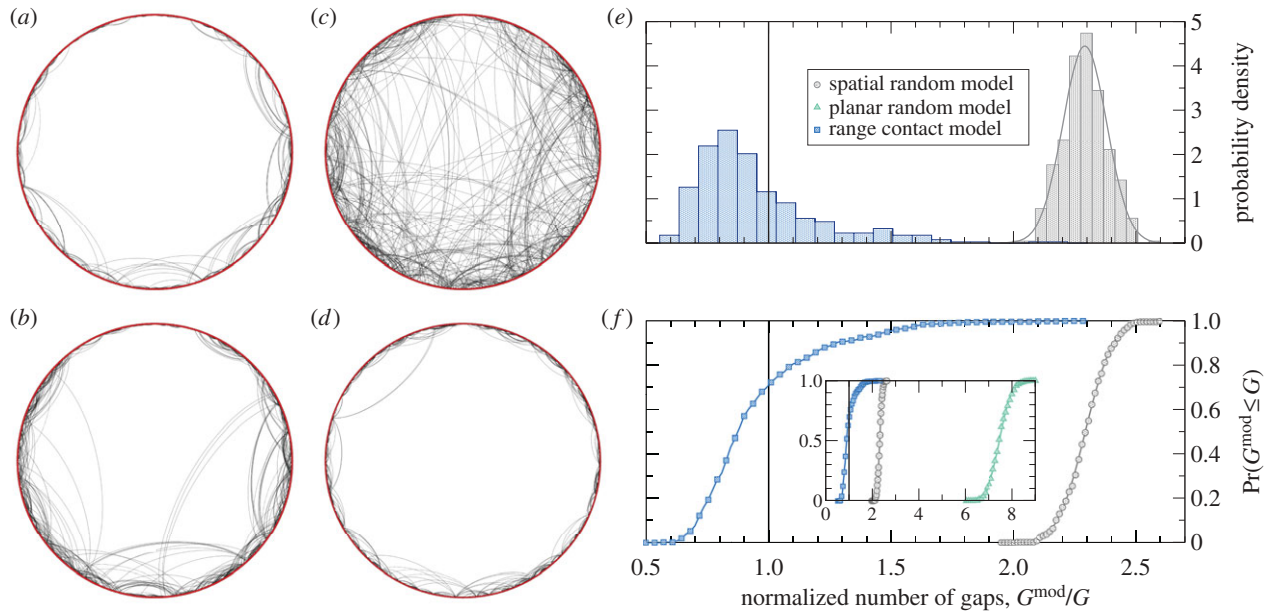

Figure 3. Optimal orderings to minimize the number of gaps for (*a*) the 663-node network of New Guinea ($G = 8737$), and for three realizations corresponding to (*b*) the SRM (average number of gaps $\langle G^{\mathrm{SRM}}\rangle = 20\,018$, deviation $\sigma_G^{\mathrm{SRM}} = 783$), (*c*) the PRM, (average $\langle G^{\mathrm{PRM}}\rangle = 65\,176$, deviation $\sigma_G^{\mathrm{PRM}} = 3741$) and (*d*) the RCM mimicking New Guinea network (average $\langle G^{\mathrm{RCM}}\rangle = 8247$, deviation $\sigma_G^{\mathrm{RCM}} = 2154$). Compare this result to (*a*). (*e*) Probability densities of the normalized number of gaps G^{mod}/G for each model. (*f*) Cumulative distributions of the same variable. Full lines for SRM and PRM (shown in the inset) correspond to Gaussian distributions with the same average and deviation from data. (Online version in colour.)

(ii) Planar random model

Our second model corresponds to networks where the empirical degree of planarity is preserved. To this end, only links in the previously identified planar component are rewired in a way that maintains planarity and the degree of the node. The PRM assumes that the planar component of language networks is strong and should be conserved. The number of rewirings allowed in this case is significantly smaller than under distance-preserving rewiring. Therefore, we have rewired $10L$ links to generate random networks under the PRM conditions, and repeated the procedure 500 times. As above, the minimum number of gaps G^{PRM} has been estimated for each PRM network. The distributions are also well approximated by Gaussian curves, again used to test the likelihood that the PRM explains the observations: in this case, this hypothesis is consistently rejected for all empirical networks (see the electronic supplementary material). The probability distribution of G^{PRM} for New Guinea has been depicted in figure 3*f*. Average values of G^{PRM} are summarized in table 1: they are systematically far from empirical values in language networks.

Figure 3*a–c* depicts optimal orderings obtained for networks generated through SRM and PRM together with the permutation that maximizes intervality for the empirical network corresponding to New Guinea. Both SRM and PRM qualitatively yield many more gaps (i.e. lower levels of intervality) than those in language networks. Interestingly, SRM and PRM implicitly reinforce the two-dimensional structure of contacts between the ranges of linguistic groups, a feature that seems to blur the one-dimensional structure uncovered by clustering and high intervality.

(iii) The range contact model

None of the two putative explanations analysed is able to account for the high intervality observed. At this point, it seems necessary to resort to different kinds of models if we wish to explain not just the high intervality measured in language networks, but also their uncommon degree distribution or the local similarity to networks embedded in a one-dimensional space. Inspired by niche models for food-web structure, which by definition entail a one-dimensional organization, we have devised a model for language networks, the RCM, where the relevant variable is the total area over which linguistic groups are spread. Our working hypothesis is that the lognormal distribution of areas [13] and the lognormal degree distribution of language networks are somehow related through actual spatial contacts between neighbouring linguistic groups ordered along a one-dimensional ring. Group interactions—expressed as conflicts for territory—coupled to demographic growth can quantitatively account for the lognormal shape of the distribution of areas [13]. Our expectation is that other topological properties may also follow from an effective arrangement of areas stemming from an intuitive condition on neighbouring domains: the assumption that the perimeter of any domain is covered by the sum of shared perimeters across all of its neighbours.

The RCM is defined as follows. (i) $N' \gtrsim N$ random numbers are drawn from a lognormal distribution with parameters (μ_a, σ_a). Each of them represents an area a_i. (ii) Areas are arranged along a one-dimensional space in no particular order. (iii) A directed link connects i to its adjacent neighbours $j = i \pm 1,\ i \pm 2, \ldots\ (1 \le j \le N')$ until the condition $\sqrt{a_i} \ge f \sum_{j \in nn(i)} \sqrt{a_j}$ is first fulfilled. This amounts to assuming that the perimeter of domain i is covered by the sum of shared perimeters of all its neighbours. Parameter f weights the average fraction of perimeter shared by domain i with each of its neighbours. (Note that, for regular tilings, $f = 1/z$. In general, f is inversely correlated to the linkage density in empirical networks, but a precise functional relationship cannot be systematically derived.) The set of nodes linked to i, $nn(i)$, is determined as follows: the initial link is established with

Table 2. Summary of topological properties for spatial networks (extracted from [25]) and language networks. Broad degree distributions that have not been proved to be power-law-like in the original reference (no exponent has been calculated) are classified as 'broad'. Exponential or other short-ranged degree distributions have been classified as 'peaked'. If various types have been reported within the same group, we mention both. Language networks are mostly planar and share the $N^{1/2}$ scaling of the average path length with other planar (or almost planar) networks reported. However, as for the clustering values, these networks are similar to non-planar networks (airline, cargo ship or neural networks). The shape of the degree distribution is distinctively new.

network	planar	clustering coefficient	average path length	degree distribution
airline networks	no	large (∼0.6)	∼log N	power law
cargo ship	no	large (∼0.5)	∼log N	power law/broad
neural networks	no	intermediate (∼0.2)	∼log N or ∼$N^{1/3}$	power law/peaked
public transportation	mostly planar	small (∼0.1)	∼$N^{1/2}$	peaked
railway	mostly planar	very small (∼0.01)	∼$N^{1/2}$	peaked
road networks	yes	intermediate (∼0.2)	∼$N^{1/2}$	peaked
power grid/water	yes	very small (∼0.01)	∼$N^{1/2}$	peaked
linguistic groups	mostly planar	large (∼0.5)	∼$N^{1/2}$	lognormal

the left or right neighbour with equal probability, and subsequent links occur with neighbours on alternating sides, not previously considered, and in order of decreasing proximity to i. The procedure is repeated for each area i; note that the order in which areas are selected is so far irrelevant. (iv) By construction, the network generated through steps (i)–(iii) is directed. Since language networks are undirected, we introduce an additional parameter q that sets the probability that a directed link is complemented by its reverse counterpart; with probability $1 - q$ the existing link is removed. In this symmetrization process, some nodes or small groups of nodes in the network might become disconnected from the bulk. We have checked that the final networks used are connected by discarding these small disconnected components, and accepting RCM networks only if their final size has at most a 0.5% size difference to N. The likely elimination of some nodes under application of the algorithm is the reason to begin with $N' \gtrsim N$ nodes.

Variations in parameter μ_a mostly cause a rescaling of the areas, leaving any other topological property of the resulting networks essentially invariant. Therefore, we fix μ_a to its empirical value (see the electronic supplementary material), and the RCM model is left with three relevant parameters: the dispersion σ_a of the lognormal distribution of areas, the fraction of shared perimeter f and the symmetrization probability q.

(iv) Comparison of the range contact model with language networks

The values of parameters that better render the empirical adjacency matrix of each of the 13 studied networks are obtained through a maximum-likelihood procedure (see the electronic supplementary material for further information).

The degree distribution yielded by the RCM is fully compatible with data in all cases. An example of the goodness of fit can be seen in figure 1*b*. The RCM distributions of the remaining networks also show an excellent agreement (see the electronic supplementary material). The RCM reproduces with reasonable accuracy the values of the clustering coefficient $\langle C \rangle$ and the average shortest path length $\langle d \rangle$ (table 1). Probably the most remarkable result concerns the distribution of the minimum number of gaps, G^{RCM}, derived from the model. The distribution of this variable has been obtained through 500 independent RCM realizations for each of the 13 language networks analysed (figure 3*e*,*f* displays the RCM distribution for New Guinea). The hypothesis that the degree of intervality of language networks can be accounted for with RCM networks cannot be rejected in any case at a 1% confidence level. In addition, the RCM accounts for the local structure of language networks measured through the distribution of the number of gaps per node (see the electronic supplementary material for details). An example of an optimal RCM network mimicking New Guinea's language network is represented in figure 3*d*.

The same hypothesis testing has been conducted for networks modified according to three different mechanisms: first, a procedure of domain aggregation that mimics language colonization; second, the removal of hubs (i.e. widespread languages) from language networks; and third, the use of available, high-resolution data of pre-colonial language distributions. High intervality of language networks remains a robust pattern under such modifications, akin to different processes of language network evolution. A detailed account of results can be found in the electronic supplementary material.

4. Discussion

The topological structure of networks of contacts between linguistic groups is consistently similar in all cases analysed despite likely differences in the accuracy of language identification in different world regions. This indicates that the characteristics uncovered are generic and robust under different classifications (such as if more taxonomic levels are considered for languages or different cultural traits are used) and under modifications that mimic the natural processes affecting language networks, as we have shown. Language networks present a particular architecture previously unseen in any other networks described in the literature. A lognormal-like degree distribution has been rarely observed [28,29], and to the best of our knowledge never reported in spatial networks. Table 2 summarizes the main properties of the latter in comparison with

language networks. Language networks constitute the second natural example of quasi-interval graphs, together with food webs [17,18]. This property supports that the architecture of language networks is mainly driven by a single quantitative attribute of nodes, which has been shown to be the area occupied by linguistic groups. Further support that domain area is the quantity that shapes language networks stems from the positive correlation between area and node degree, as the RCM trivially predicts (these results will be published elsewhere). In analogy with niche models, we have introduced the RCM, which successfully reproduces the structure and organization of language networks. Other network models parametrized in two dimensions have been successful in reproducing certain food-web properties [30], a small number of species attributes being needed to explain their global topology [18]. Confronting language network data with those models is a future direction worth pursuing.

The number of neighbours of a given language depends on its area of spread, a quantity strongly correlated to the number of speakers [13]. The number of contacts is also a measure of the likelihood of conflicts between different groups. It has been argued that the frequency and strength of those conflicts affects the area occupied by the group [13]. A particular form of conflict between neighbouring languages is competition for speakers. The dynamics of extinction of languages is influenced by the attractiveness of competing languages [31], by geography [32] and, plausibly, by the number of competitors, which we have shown to vary broadly.

Overall, language networks might be regarded as a first approximation to networks of contacts between cultures. As such, their topology may have implications in the way cultural innovations (e.g. farming, animal domestication or iron tools) spread in the past [33], and in the modelling of the spreading process [34]. A common language is a fast vehicle for the dissemination of knowledge assuming that individuals speaking the same language experience stronger ties than those shared with other linguistic groups. The existence of a complex underlying topology of contacts may entail qualitative changes in the propagation dynamics, as compared with propagation on homogeneous media. This modification echoes how our understanding of epidemic spreading was improved upon the introduction of heterogeneous networks [35] and calls for a deeper study of its effect in the cultural relationships that might be established between human groups.

High intervality, a property reflecting a one-dimensional underlying ordering of nodes (linguistic groups in our case), is indeed a remarkable feature considering that language networks are embedded in two-dimensional space. Although language domains are clustered together, contacts between them are such that the spatial ordering of languages resembles one-dimensional arrays. This suggests that linguistic communities interact along certain directions to a greater extent than would be expected for spatial networks with low intervality. These patterns are robust throughout different regions across the world, and could be used to further improve our understanding of language organization, change and extinction.

The placement of cultural groups is plausibly related to properties of the landscape. Mountain ranges, coastlines, rivers and fertile valleys condition the position and extension of human settlements, as well as preferred directions for movement and group interaction [36–38], which seem to partly eliminate the freedom of a two-dimensional space in favour of linear interactions among neighbouring groups. Indirect evidence of the role played by the environment arises from the significant dependence between linguistic diversity and, especially, landscape roughness and river density [9]. Whether an explicit consideration of topography might explain the quasi-intervality of language networks is an open question that deserves additional attention.

Widespread languages play a relevant role in several of the issues tackled. Usually, they have many neighbours, responsible for most 'shortcuts' in our networks and, consequently, for decreasing intervality. The elimination of those languages in the Ethnologue networks, or their progressive appearance through models that effectively consider modern evolutionary processes of language extinction and growth, shows however that they are not essential in determining the topological properties uncovered. Widespread languages are the hubs of language networks, though at the same time they percolate across continents, and cause the isolation and fragmentation of groups of minority languages. That is the case for North America, with 609 remaining languages forming 30 disconnected components located on the continental landmass. Asia also holds an astonishingly large number of solitary languages (34% of its total diversity) and many disconnected components. However, the latter are mostly due to the abundance of large islands, not to fragmentation on the mainland. The structure of language networks is in continuous transformation due to the sustained growth of widespread languages and to the disappearance of many others: 3500 languages are predicted to become extinct within the next century [39]. Extinction dynamics are likely to be affected by variations in contacts with potentially competing languages, but also by increasing isolation and area shrinkage. These factors find their counterpart in ecology. Habitat fragmentation leads to the isolation of species, to a reduction of their home ranges, and eventually to an accelerated extinction [40]. It would be interesting to extend our analysis to networks of contacts between species ranges. An intriguing question is whether the architecture of those networks belongs to the class here described and, in that case, whether cultural and biological diversity patterns are the final products of generic constructive processes. As our knowledge increases, so does evidence supporting the qualitative and quantitative parallelisms between both evolutionary systems.

Acknowledgements. The authors are indebted to Jacobo Aguirre, Sara Cuenda, José A. Cuesta, Anxo Sánchez, Daniel B. Stouffer, Damián H. Zanette and two anonymous reviewers for constructive criticism of the manuscript.

Author contributions. J.A.C. and S.M. conceived and designed the research; J.A.C. performed the research; J.A.C., J.B.A. and S.M. analysed the data; J.A.C. and J.B.A. contributed materials/analysis tools; J.A.C. and S.M. wrote the paper.

Funding statement. The authors acknowledge financial support from Spanish MICINN through projects FIS2011–27569 and FIS2011–22449 (J.A.C.), from Comunidad de Madrid through project MODELICO, S2009/ESP-1691 (J.A.C., S.M.) and from the Carlsberg Foundation (J.B.A.).

References

1. Pagel M, Mace R. 2004 The cultural wealth of nations. *Nature* **428**, 275–278. (doi:10.1038/428275a)
2. Diller JV. 2011 *Cultural diversity: a primer for the human services*. Belmont, CA: Brooks/Cole, Cengage Learning.
3. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002 Genetic structure of human populations. *Science* **298**, 2381–2385. (doi:10.1126/science.1078311)
4. Maffi L. 2005 Linguistic, cultural, and biological diversity. *Annu. Rev. Anthropol.* **29**, 599–617. (doi:10.1146/annurev.anthro.34.081804.120437)
5. Burnside WR, Brown JH, Burger O, Hamilton MJ, Moses M, Bettencourt L. 2011 Human macroecology: linking pattern and process in big-picture human ecology. *Biol. Rev.* **87**, 194–208. (doi:10.1111/j.1469-185X.2011.00192.x)
6. Darwin C. 1871 *The descent of man, and selection in relation to sex*. London, UK: John Murray.
7. Nettle D. 1999 *Linguistic diversity*. Oxford, UK: Oxford University Press.
8. Moore JL, Manne L, Brooks T, Burgess ND, Davies R, Rahbek C, Williams P, Balmford A. 2002 The distribution of cultural and biological diversity in Africa. *Proc. R. Soc. Lond. B* **269**, 1645–1653. (doi:10.1098/rspb.2002.2075)
9. Axelsen JB, Manrubia S. 2014 River density and landscape roughness are universal determinants of linguistic diversity. *Proc. R. Soc. B* **281**, 20133029. (doi:10.1098/rspb.2013.3029)
10. Collard IF, Foley RA. 2002 Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? *Evol. Ecol. Res.* **4**, 371–383.
11. Gomes MAF, Vasconcelos GL, Tsang IJ, Tsang IR. 1999 Scaling relations for diversity of languages. *Physica A* **271**, 489–495. (doi:10.1016/S0378-4371(99)00249-6)
12. Preston FW. 1962 The canonical distribution of commonness and rarity: part I. *Ecology* **43**, 185–215. (doi:10.2307/1931976)
13. Manrubia SC, Axelsen JB, Zanette DH. 2012 Role of demographic growth and conflict in the population–area relationship for human languages. *PLoS ONE* **7**, e40137. (doi:10.1371/journal.pone.0040137)
14. Harte J, Blackburn T, Ostling A. 2001 Self-similarity and the relationship between abundance and range size. *Am. Nat.* **157**, 374–386. (doi:10.1086/319323)
15. Gaston KJ, He F. 2002 The distribution of species range size: a stochastic process. *Proc. R. Soc. Lond. B* **269**, 1079–1086. (doi:10.1098/rspb.2002.1969)
16. Cohen JE. 1977 Food webs and the dimensionality of trophic niche space. *Proc. Natl Acad. Sci. USA* **74**, 4533–4536. (doi:10.1073/pnas.74.10.4533)
17. Stouffer DB, Camacho J, Amaral LAN. 2006 A robust measure of food web intervality. *Proc. Natl Acad. Sci. USA* **103**, 19 015–19 020. (doi:10.1073/pnas.0603844103)
18. Eklöf A *et al*. 2013 The dimensionality of ecological networks. *Ecol. Lett.* **16**, 577–583. (doi:10.1111/ele.12081)
19. Williams RJ, Martinez ND. 2000 Simple rules yield complex food webs. *Nature* **404**, 180–183. (doi:10.1038/35004572)
20. Cattin M-F, Bersier L-F, Banasek-Richter C, Baltensperger R, Gabriel J-P. 2004 Phylogenetic constraints and adaptation explain food-web structure. *Nature* **427**, 835–839. (doi:10.1038/nature02327)
21. Stouffer DB, Camacho J, Ng CA, Amaral LAN. 2005 Quantitative patterns in the structure of model and empirical food webs. *Ecology* **86**, 1301–1311. (doi:10.1890/04-0957)
22. Allesina S, Alonso D, Pascual M. 2008 A general model for food web structure. *Science* **320**, 658–661. (doi:10.1126/science.1156269)
23. Capitán JA, Arenas A, Guimerà R. 2013 Degree of intervality of food webs: from body-size data to models. *J. Theor. Biol.* **334**, 35–44. (doi:10.1016/j.jtbi.2013.06.004)
24. Cohen JE, Briand F, Newman CM. 1990 *Community food webs: data and theory*. Berlin, Germany: Springer.
25. Barthélemy M. 2011 Spatial networks. *Phys. Rep.* **499**, 1–101. (doi:10.1016/j.physrep.2010.11.002)
26. Gordon RG, Grimes BF, Summer Institute of Linguistics. 2005 *Ethnologue: languages of the world*. Dallas, TX: SIL International.
27. Barrat A, Weigt M. 2000 On the properties of small-world network models. *Eur. Phys. J. B* **13**, 547–560. (doi:10.1007/s100510050067)
28. Perc M. 2010 Growth and structure of Slovenia's scientific collaboration network. *J. Informetrics* **4**, 475–482. (doi:10.1016/j.joi.2010.04.003)
29. Clauset A, Shalizi CR, Newman MEJ. 2009 Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703. (doi:10.1137/070710111)
30. Staniczenko PPA, Smith MJ, Allesina S. 2014 Selecting food web models using normalized maximum likelihood. *Methods Ecol. Evol.* **5**, 551–562. (doi:10.1111/2041-210X.12192)
31. Abrams DM, Strogatz SH. 2003 Modelling the dynamics of language death. *Nature* **424**, 900. (doi:10.1038/424900a)
32. Patriarca M, Heinsalu E. 2009 Influence of geography on language competition. *Physica A* **388**, 174–186. (doi:10.1016/j.physa.2008.09.034)
33. Ackland GJ, Signitzer M, Stratford K, Cohen MH. 2007 Cultural hitchhiking on the wave of advance of beneficial technologies. *Proc. Natl Acad. Sci. USA* **104**, 8714–8719. (doi:10.1073/pnas.0702469104)
34. Kandler A, Perreault C, Steele J. 2012 Cultural evolution in spatially structured populations: a review of alternative modeling frameworks. *Adv. Complex Syst.* **15**, 1203001. (doi:10.1142/S0219525912030014)
35. Barthélemy M, Barrat A, Pastor-Satorras R, Vespignani A. 2005 Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *J. Theor. Biol.* **235**, 275–288. (doi:10.1016/j.jtbi.2005.01.011)
36. Davison K, Dolukhanov PR, Sarson G, Shukurov A. 2006 The role of waterways in the spread of the neolithic. *J. Archeol. Sci.* **33**, 641–652. (doi:10.1016/j.jas.2005.09.017)
37. O'Shea JM. 2011 A river runs through it: landscape and the evolution of Bronze Age networks in the Carpathian basin. *J. World Prehist.* **24**, 161–174. (doi:10.1007/s10963-011-9046-6)
38. Greenhill SJ. 2014 Demographic correlates of language diversity. In *The Routledge handbook of historical linguistics* (eds C Bowern, B Evans), pp. 557–578. Abingdon, UK: Routledge.
39. Krauss M. 1992 The world's languages in crisis. *Language* **68**, 4–10. (doi:10.1353/lan.1992.0075)
40. Fahrig L. 2003 Effects of habitat fragmentation on biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **34**, 487–515. (doi:10.1146/annurev.ecolsys.34.011802.132419)