Network communities within and across borders

Federica Cerina\textsuperscript{1,3}, Alessandro Chessa\textsuperscript{2,3}, Fabio Pammolli\textsuperscript{2} & Massimo Riccaboni\textsuperscript{2}

\textsuperscript{1}Universit\`a degli Studi di Cagliari, Cagliari, Italy, \textsuperscript{2}IMT Institute for Advanced Studies Lucca, Lucca 55100, Italy, \textsuperscript{3}Linkalab, Complex Systems Computational Laboratory, Cagliari 09129, Italy.

We investigate the impact of borders on the topology of spatially embedded networks. Indeed territorial subdivisions and geographical borders significantly hamper the geographical span of networks thus playing a key role in the formation of network communities. This is especially important in scientific and technological policy-making, highlighting the interplay between pressure for the internationalization to lead towards a global innovation system and the administrative borders imposed by the national and regional institutions. In this study we introduce an outreach index to quantify the impact of borders on the community structure and apply it to the case of the European and US patent co-inventors networks. We find that (a) the US connectivity decays as a power of distance, whereas we observe a faster exponential decay for Europe; (b) European network communities essentially correspond to nations and contiguous regions while US communities span multiple states across the whole country without any characteristic geographic scale. We confirm our findings by means of a set of simulations aimed at exploring the relationship between different patterns of cross-border community structures and the outreach index.

Over recent decades, political change and new transportation and information technologies have enhanced international openness and cross-border integration. Globalization has made social networks more international and human communities more integrated across cultural and political borders. This is witnessed by the increasing number of long-range connections in multiple networks, such as trade, human mobility, communications, financial investments and scientific collaborations\textsuperscript{1–4}.

Enabled by modern technology, people from all over the world are offered a myriad of opportunities for social interactions and group assembly with increasingly larger geographic ranges\textsuperscript{5}. Nonetheless, this does not mean that networks can stretch across a borderless world indefinitely: as for climate networks one can detect geographical regions with the same climate variability\textsuperscript{6–8}. As individual nodes in socio-economic networks occupy a given region in space, it is reasonable to assume that geographical proximity also plays a crucial role in social link formation\textsuperscript{9}. Indeed, a power-law decay in link probability with distance acting as a spatial constraint has been observed\textsuperscript{5,7,10,11}. In a recent meta-analysis estimating the role of distance in international trade, it has been shown that $t \propto d^{-\gamma}$, where $t$ is trade, $d$ is the distance and $\gamma \approx 1$ over more than a century of data\textsuperscript{12}. Arguably, distance is not the only spatial constraint on link formation, since natural and artificial borders also have the power to hamper connectivity. Communication and transportation routes, on the contrary, facilitate long-distance interactions\textsuperscript{13}. Geographical and institutional borders are relevant in all networks where distance matters, such as power grid networks, transportation and communication networks as well as collaboration networks. Natural, artificial and administrative borders can substantially reduce the probability of link formation by introducing a major constraint in terms of cost, service, capacity and reliability of global networks. Two of the most widely accepted results in international economics are that trade is impeded by distance and that the crossing of national borders also sharply reduces trade. It has been shown, for instance, that national borders are responsible for a fivefold decrease in world trade when compared to a borderless world\textsuperscript{14}. Well-known global networks are transportation and communication networks, such as the airline network and the World Wide Web, for which the role of borders has been recently documented\textsuperscript{15,16}.

Although physical and social propinquity still dominates the organization of human activities, increasing long-range interactions are radically transforming the architecture of global networks across cultural, political and geographical borders. Different measures of globalization have been used in the literature. Traditionally, international openness has been proxied by the share of cross-border links over the total number of connections. More recently, various network-based measures of cross-border integration have been introduced\textsuperscript{17–19}. Similarly, a multinational corporation consists in a group of geographically dispersed organizations that include its headquarters and the various national subsidiaries. Such an entity can be conceptualized as an international spatially
Table 1 | Home base of communities according to three different criteria

| Community | Europe | United States |
|-----------|--------|---------------|
|           | Ñ(dQ) | Ñ(dQ + S) | Ñ(Wint) | Community | Ñ(dQ) | Ñ(dQ + S) | Ñ(Wint) |
| I         | DE     | DE          | DE       | I         | CA     | CA         | CA       |
| II        | DE     | DE          | DE       | II        | NJ     | NJ         | NJ       |
| III       | FR     | FR          | FR       | III       | MA     | MA         | MA       |
| IV        | DE     | DE          | DE       | IV        | OH     | OH         | OH       |
| V         | DE     | DE          | DE       | V         | PE     | PE         | PE       |
| VI        | NL     | NL          | NL       | VI        | MN     | MN         | MN       |
| VII       | DE     | DE          | DE       | VII       | IL     | IL          | IL       |
| VIII      | UK     | UK          | UK       | VIII      | CA     | CA         | CA       |
| IX        | DK     | Fi          | Fi       | IX        | TX     | TX          | TX       |
| X         | DE     | DE          | DE       | X         | OH     | OH          | OH       |
| XI        | IT     | IT          | IT       | XI        | NC     | NC          | NC       |
| XII       | AT     | AT          | AT       | XII       | CT     | CT          | CT       |
| XIII      | ES     | ES          | ES       | XIII      | NY     | NY          | NY       |
| XIV       | DE     | DE          | DE       | XIV       | GA     | GA          | GA       |

The table compares the home base of countries found for each community using three different criteria: simple | dQ |, | dQ | + S where S is the node strength, and internal link density Wint. Results do not vary significantly with the only exception of the Nordic cluster that, when we use | dQ |, has its center in Denmark instead of Finland | dQ | + S and Wint.

We proceed as follows. In “Methods” we describe the methodology we used to define an outreach index, while in “Data” we analyze the patent dataset by building a geo-coded network of inventors. Section “Results” illustrates our main findings on (a) the distance distribution of links (b) the geographical distribution of network communities and the selection of core regions and (c) the outreach and geographical span of communities. As a robustness check we run a set of simulations to show how the outreach index can be used to identify different patterns of cross-border communities. The concluding section outlines our contribution to the analysis of cross-border spatially embedded network communities. We also highlight the relevance of our methodology to the proper assessment of the future emergence of an integrated European Research Area and similar regional integration efforts.

Table 2 | European versus US communities: geographical span, \( D_s \), and outreach index, \( O_s(N_s) \)

| Community | Core | Country | \( D_s(N_s) \) | Community | Core | Country | \( D_s(N_s) \) |
|-----------|------|---------|----------------|-----------|------|---------|----------------|
| I         | Mannheim | DE     | 1.9090 0.0509 | I         | San Jose | CA     | 13.8236 0.8775 |
| II        | Düsseldorf | DE   | 1.1016 0.0038 | II        | New Brunswick | NJ     | 7.8733 0.9379 |
| III       | Paris     | FR     | 6.0087 0.0181 | III       | Cambridge | MA     | 11.2038 0.8911 |
| IV        | Berlin    | DE     | 2.2098 0.0477 | IV        | Cincinnati | OH     | 3.9768 0.7525 |
| V         | Stuttgart | DE     | 1.4729 0.0087 | V         | Philadelphia | PE    | 9.6278 0.9792 |
| VI        | Eindhoven | NL     | 1.5392 0.5440 | VI        | Minneapolis | MN     | 7.9104 0.9803 |
| VII       | Munich    | DE     | 1.1793 0.0000 | VII       | Chicago   | IL     | 7.9630 0.9909 |
| VIII      | Cambridge | UK     | 2.5560 0.1335 | VIII      | San Diego | CA     | 17.8420 0.9288 |
| IX        | Helsinki | FI     | 6.3203 0.6942 | IX        | Houston   | TX     | 6.9423 0.9803 |
| X         | Nuremberg | DE     | 1.5577 0.0223 | X         | Cleveland | OH     | 6.7129 0.9734 |
| XI        | Milan     | IT     | 3.4392 0.0309 | XI        | Raleigh   | NC     | 6.4801 0.9876 |
| XII       | Wien      | AT     | 1.9614 0.0537 | XII       | New Haven | CT     | 11.3220 0.9641 |
| XIII      | Barcelona | ES     | 5.2386 0.5065 | XIII      | Schenectady | NY    | 6.7066 0.9718 |
| XIV       | Lörrach   | DE     | 0.2957 0.6356 | XIV       | Atlanta   | GA     | 5.7034 0.9364 |
| Average   | Europe    | 2.6278 0.1964 | Average | US       | 8.8634 0.9394 |

The table compares cross-border communities in the US and Europe. Communities are identified based on the main city of the NUTS3 region with the highest \( dQ | \) or \( dQ | + S \). We report the values of geographical span \( D_s \) and the outreach index \( O_s(N_s) \) with respect to home country for European regions and states in the US. Home country has been selected based on the maximum \( dQ | \) or \( dQ | + S \) and \( W_{int} | \).
Starting from these data we define $w_{ij}$ as the number of links between regions $i$ and $j$. In our network $w_{ij}$ will be equal to the number of patents jointly invented by the two regions. We use a full-counting approach so that a patent with $N$ number of patents jointly invented by the two regions. We use a self-loops.

Results

Table 2 reports the value of the geographical span on our data. Larger values of $D$ means that the members of the community are geographically spread out; at first glance one could conclude that US community members are more spread out than the European ones on average.

Distance distribution. It is well known that social interactions negatively depend on distance. More precisely, it has been shown that the probability of a tie between any pair of nodes decays with distance as a power-law $-d^{-\alpha}$ where $1 \leq \alpha \leq 2$. Figure 2 compares the distance distribution of links in the European and US networks from 1986 to 2009. The two distributions clearly depict different behaviors: the US distribution is well approximated by a power-law as reported in the literature with an exponent $\alpha \approx 1$, whereas the European one shows an exponential behavior.

Figure 3 shows the results of the analysis performed using the modularity method. For the sake of clarity, every node that corresponds to a geographical region, which is geo-referenced and displayed on a map, is given the same color as the community it belongs to. This results in nodes in different communities having different colors as well. Figure 3 also shows the results of the core analysis performed on the partition obtained using the Newman-Girvan Modularity. In this representation each community has been given a different color. In the European case, the community structure almost perfectly matches the national boundaries of the Member States of the European Union. The only significant difference seems to be Germany, which is sectioned off into multiple communities with an average size in the order of a Land Region (NUTS2 level).

---

**Table 3 | Simulated Outreach index**

| $N_i$ | Outreach index, simulated | Parameters |
|------|--------------------------|-----------|
|      | $d = 0$ | $d = 10^1$ | $d = 10^2$ | $d = 10^3$ | $d = 10^4$ | $\gamma_{in}$ | $\gamma_{out}$ | $\gamma_{across}$ |
| FR   | 0.019 ± 0.002 | 0.022 ± 0.003 | 0.034 ± 0.005 | 0.121 ± 0.007 | 0.285 ± 0.019 | 1 | 0 | 0.08 |
| NL   | 0.533 ± 0.021 | 0.542 ± 0.021 | 0.538 ± 0.025 | 0.544 ± 0.022 | 0.534 ± 0.017 | 1 | 0.1 | 0.10 |
| FI   | 0.711 ± 0.018 | 0.721 ± 0.015 | 0.722 ± 0.018 | 0.721 ± 0.025 | 0.725 ± 0.011 | 1 | 0.5 | 0.20 |
| CA   | 0.887 ± 0.011 | 0.889 ± 0.009 | 0.884 ± 0.009 | 0.889 ± 0.008 | 0.886 ± 0.008 | 1 | 0.1 | 0.20 |

The table reports the values of the simulated outreach index $O_i(N_i)$ obtained for the 4 cases FR, NL, FI, CA and for different values of the separating distance $d$. We use here different values of $\gamma$ to differentiate the four cases of the simulation: $\gamma_{in}$ is the network density within borders, $\gamma_{out}$ is the network density outside borders and $\gamma_{across}$ regulates cross-border links between the inner and the outer part of the community.
US community structure reveals, on the other hand, a practically opposite behavior: communities are stretched out over more than one state, at great distances from the alleged geographical core. For each community, colors are graduated according to the $|dQ|$ of the node: darker colored nodes have a higher $|dQ|$ and are therefore “more central” while lighter colored ones are less central since they have a lower $|dQ|$. Combined with our previous results regarding the different decay of connectivity at a distance (power-law in the US, exponential in Europe), we clearly show that the presence of national borders in Europe has a strong role in shaping the topology of the network, both reducing connectivity at a distance and constraining networks community in space.

As previously said, Europe is a genuine transnational network whereas the US system is not. Accordingly, their different behaviors do not come as a surprise. On the one hand, since the US innovation system is rather homogeneous, the probability for coast-to-coast interaction (up to a cut off distance of $10^8$ kilometers) is high. On the other hand, Europe is a collection of almost independent national systems of innovation (see Figure 2). Namely, in the European case there is a cut off in the distribution due to strong country border effects, that eventually results in the exponential decay behavior with a characteristic length that is roughly of the size of the average country diameter (about 363 Kilometers). The European network thus differs sharply from the US case, where the state border effect is almost negligible. In the US, scientific and technological communities span throughout the country without any characteristic scale. Moreover we find a power law decay of connectivity at a distance; even when we focus on a single European nation such as Germany, we still note some interesting differences with borders that play a stronger role in reducing connectivity between German Länders when compared to their US counterparts.

Next we proceed to consider the outreach index of the communities. Before doing that, we must determine which one of the three criteria reported in the Section “Methods” is the most appropriate. In Table 1 we show the home base country of the communities we identify. As one can see, the outcome is the same in all the cases except for the ninth European community, for which we obtained Denmark as the home base according to the first criterion, and Finland for the other two. All in all, it turns out that the final result does not crucially depend on the method we use to identify the home country. Therefore, the outreach index has an high degree of universality. In the following analysis we will opt for the sensible solution of using a balanced method which takes both the topological centrality inside the community ($|dQ|$) and the total weight ($S$) attached to that node (second criterion) into account. In the cases in which the community detection does not come from a modularity optimization and the $|dQ|$ value is not available, the third criterion can also be considered as a viable alternative.

Table 2 reports the value of the outreach index by choosing the home base according to the second criterion. Given that the outreach index always lies between 0 and 1, we are allowed to compare the outreach of European communities with US counterparts (see Table 2). As expected, the outreach value is about 0 for almost every community in Europe, while this value is always close to 1 in the US. This means that the communities in the United States undertake more outreach than in Europe. However, we should remember that the United States are not a truly transnational network, and accordingly it makes sense to compare the US with Germany as we did before. Indeed, community detection showed that Germany behaves

---

**Figure 2 | Distance distribution.** This figure shows the distance distribution of the links both for Europe (blue dots) and USA (red dots) in log-log scale and their best fits: a power law $y = \beta d^{-\alpha}$ for the US (black), and exponential distribution $m \exp(-nx)$ for Europe (green), with $\beta = 426.583 \pm 29.474$, $\alpha = 0.960 \pm 0.017$, $m = 65030 \pm 570.5$, $n = 33.7 \pm .35$. 
The community structure of the patent co-inventorship network: Europe versus the United States. This figure shows the results obtained performing a community detection analysis both on the (a) European and the (b) US networks using the Newman-Girvan method. In this representation, each community has been given a different color. It can be seen that, in the European case, the community structure almost perfectly matches the national boundaries of the 15 member states of the European Union. The only significant difference seems to be Germany, which is sectioned off into more than one community. The US community structure reveals almost an opposite behavior: communities are stretched out over more than one state and at great distances. Each community color is graduated according to the \(|dQ|\) of the node: darker colored nodes have a higher \(|dQ|\), so they are “more central”, while lighter colored ones are less central since they have a lower \(|dQ|\). The nodes with the highest \(|dQ|\) are considered community core regions. We use the same community color coding for networks (top) and maps (bottom). Maps and networks were generated using the open source software Gephi and QGIS, respectively.

Simulations. The inspection of the outreach index for the US and Europe reveals the existence of four main community types: the French case, which exemplifies European national communities, the Benelux and Nordic clusters, which are two cases of regional integration, and the California case, which is representative of the general behavior of long-range out-reaching communities in the United States. In order to uncover the internal mechanics involved in the creation of these cases we reproduced the relevant patterns that emerged from real data by simulating the internal structure of each community.

The artificial model is defined as follows. Out of a total number \(N\) of nodes we decide what fraction of them, say \(N_c\), to place into the central/home region and what fraction \(N_e\) to place into the external one(s). For the sake of simplicity we choose to shape the regions as circles whose radii are proportional to the number of nodes, so that the more the nodes the bigger the region.

As the regions belonging to the same community can either be adjacent to each other or not, we introduce the parameter \(d\) to regulate the spatial separation between them.

Once the nodes have been placed into communities, we randomly create links between them until the total number of links is reached. In general, if \(M\) is the maximum number of possible links differently and splits into several sub-clusters. Then, if we take the NUTS2 level Länd (the German equivalent of a US state) as the reference nation \(N_e\) instead of Germany as a whole (which is NUTS1), the outreach values are sensibly different. They become comparable to the US values ranging from .49 for the region centered around Munich (Oberbayern) to .95 for the region of Mannheim (Karlsruhe).

We then calculated the outreach index for different values of \(d\) (see Table 3). Even when \(d = 0\), the case for which there is no separation among the regions (see Figure 4), in addition to a set of parameters extracted from the data, the model closely reproduces the spatial organization of the four real cases. As one can also notice in Table 3, the simulated outreach indexes are similar to the empirically observed values.

Discussion

The role of distance in spatially embedded complex networks has been recently investigated. Empirically speaking, it has been found that connectivity tends to decay with distance according to a power-law relationship. This is in line with previous results in the economic literature, where an inverse power relationship has been repeatedly observed in gravity-like models of international trade, human migration and foreign investment. Despite this growing body of evidence regarding complex networks in space, still little is known about the role of national borders in the formation of cross-national networks. In this paper we aim at understanding more about this role by analyzing the structure of network communities within and across borders. We show that, while the connectivity of US scientific communities decays as a power of distance, European scientific communities tend to be confined within national borders. We introduce a new measure for the outreach of network communities across borders and confirm our results via simulations. All in all, our findings reveal that Europe is still a collection of national systems of innovation and the European Research Area is still far from becoming reality. Our methodological approach can be used to keep track of scientific collaboration though the international landscape.
of the progress toward the integration of the European Research Area. More in general, the outreach index we discuss in this paper is worth using to detect the impact of borders on the formation and dynamic evolution of spatially embedded networks.

Methods

Community detection and core regions. Beyond the local topological features, many networks have groups of nodes marked by the high density of their internal links with respect to the outgoing links that connect the groups with each other. This is especially true if the nodes are embedded in space and subject to geographical constraints. Indeed, if geography has such a strong role in link formation, after performing a community detection analysis we would expect to find well-defined communities of spatially connected nodes. However, it has been already shown that geographical clusters and network communities do not perfectly overlap. There are now many community detection algorithms and in the following passage we will use modularity optimization introduced by Newman and Girvan through the “Louvain” algorithm.

By definition, if the modularity associated with a network has been optimized, every perturbation in the partition leads to a negative variation in the modularity (dQ). If we move a node from a partition we have M − 1 possible choices (with M as the number of nodes). The modularity can be defined as possible targets for the node’s new host community. It is possible to define the dQ associated with each node as the smallest variation in absolute value (or the closest to 0 since dQ is always a negative number) for all the possible choices. This is a measure of how central that node is to its community.

In the following we introduce two measures of geographical span and community outreach to properly assess the role of borders in network formation.

Geographical span and community outreach. The geographical dispersion of a community s, or geographical span Ds, can be measured as:

\[ D_s = \frac{1}{n_s} \sum_{c \in C_s} \sqrt{(X_i - x_c)^2 + (Y_i - y_c)^2} \]

where \( n_s \) is the number of nodes in the community \( C_s \) and \((X_i, Y_i)\) are the coordinates of the geographical center of the community, with \( X_s = (1/n_s) \sum_{c \in C_s} X_i \) and \( Y_s = (1/n_s) \sum_{c \in C_s} Y_i \).

The geographical dispersion is a pure spatial index and does not contain any information about the network structure of the possible links connecting the nodes embedded in space. Since this index is neither normalized nor weighted, it is inadequate for the comparison of different structures, like Europe and the US, where the distances are considerably different. Moreover, the geographic span does not measure how communities reach out. To do this, we introduce a new index, the outreach index, defined as follows:

\[ O_s(N_c) = 1 - \sum_{i,j \in N_c} \sum_{\delta_{ij} = w_{ij}} \frac{d_{ij}}{w_{ij}} O_s(N_c) \in [0,1] \]

where \( N_c \) is the home base of the community \( s \), \( d_{ij} \) and \( w_{ij} \) are the distance and the weight of the link between nodes \( i \) and \( j \) belonging to \( C_s \), with respect to the same quantity calculated for all pairs of nodes \( i \) and \( j \) belonging to \( C_s \). Figure 1 provides a schematic representation of the way in which the outreach index is obtained.

Multiple criteria can be used to select the home base \( N_c \):

1. the home base is located in the region with the highest \( dQ \). The regions with the highest \( dQ \) can be defined as the core of the community, based on the intensity of intra-community ties;
2. the center of the community can be chosen as the one with the highest \( dQ \) + \( S \), where \( S \) is the sum of the weights of all outgoing and incoming links of a node. This index accounts for both the role the node plays in the intra-community connectivity (\( dQ \)) and the overall centrality of the region, as measured by the node strength (\( S \));
3. the area that scores the highest internal link density is selected as the home base. Intuitively, this criterion identifies the region with the highest share of inner linkages in the community. In such a case the selected regions will be the biggest ones, regardless of where the core region is located.

In our analysis we find that the above listed criteria tend to provide similar results. Therefore the choice between them depends on the selected community detection method and data availability.

Acknowledgments

Authors acknowledge funding from the PNR project CRISIS Lab. MR acknowledges funding from the MIUR (PRIN project 20093Z32BF and FWG (G07S31N1). Federica Cerina gratefully acknowledges Sardinia Regional Government for the financial support of her PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007–2013 – Axis IV Human Resources, Objective I.3.1, Line of Activity I.3.1.).

Author contributions

F.C. carried out the simulation and prepared Figures 2–4, M.R. prepared Figure 1, M.R., A.C., F.C. and F.P. wrote the main manuscript text. All authors contributed to data analysis and manuscript revisions.
Additional information

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

How to cite this article: Cerina, F., Chessa, A., Pammolli, F. & Riccaboni, M. Network communities within and across borders. Sci. Rep. 4, 4546; DOI:10.1038/srep04546 (2014).