Infrastructures connecting people. A mechanistic model for terrestrial transportation networks.

Luce Prignano[1,2], Lluc Font-Pomarol[3], Ignacio Morer[1,2], and Sergi Lozano[1,4]

[1] *Universitat de Barcelona Institute of Complex Systems (UBICS) Universitat de Barcelona, Barcelona, Spain*
[2] *Departament de Fisica de la Matèria Condensada, Universitat de Barcelona, Barcelona, 08028, Spain*
[3] *Departament d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona, Spain*
[4] *Departament d'Història Econòmica, Institucions, Política i Economia Mundial, Universitat de Barcelona, Barcelona, 08028,Spain*

March 25, 2022

Abstract

The structure and evolution of Terrestrial Transportation Infrastructures (TTIs) are shaped by both socio-political and geographical factors, hence encoding crucial information about how resources and power are distributed through a territory. Therefore, analysing pathway, railway or road networks allows us to gain a better understanding of the political and social organization of the communities that created and maintained them. Network science can provide extremely useful tools to address quantitatively this issue. Here, focusing on passengers transport, we propose a methodology for mapping a TTI into a formal network object able to capture both the spatial distribution of the population and the connections provided by the considered mean of transport. Secondly, we present a simple mechanistic model that implements a wide spectrum of decision-making mechanisms which could have driven the creation of links (connections). Thus, by adjusting few parameters, for any empirical system, it is possible to generate a synthetic counterpart such that their differences are minimized. By means of such inverse engineering approach, we are able to shed some light on the processes and forces that moulded transportation infrastructures into their current configuration, without having to rely on any additional information besides the topology of the network and the distribution of the population. An illustrative example is also provided to showcase the applications of the proposed methodology and discuss how our conclusions fit with previously acquired knowledge (and literature) on the topic.

1 Introduction

Historically, transportation infrastructures have had huge impact over the development of the territories. They allow the movement of people and goods, affecting the long-term capability to cope with changing socio-economic scenarios. Among them, terrestrial transportation infrastructures (hereafter referred to as TTIs) differ from other transportation systems mainly in that their connections are physical.

This implies that, unlike in the case of air, maritime, or river transportation, whose growth concentrates the costs on localized structures –such as airports or harbours– to enable the hosting of more connections, most of the costs of TTIs are associated to the construction of connections themselves –roads, railways, etc.– while a much smaller fraction of the total budget is usually for toll booths or stations.

This particular feature opens the door to the exploitation of TTIs as a fundamental source of information about the societies that created and maintained them.

Indeed, only in the case of terrestrial infrastructures, the decision of connecting two previously disconnected places is a crucial, not easily reversible one. The balance between the cost of building a new connection – which typically increases with the distance – and the provided benefit to the system affects the development of TTIs, which can hence be regarded as the result of multiple interests and competing constraints.

Their structure and evolution are influenced, on the one hand, by the changing needs that they are supposed to satisfy, and on the other hand, by how resources and power are distributed through a territory [1, 2].

In this sense, each TTI as a whole is the emerging output of a complex process involving many interacting actors and different spatial and temporal scales.

In the context of archaeological research, the importance of TTIs for the understanding of the political and social organization of the societies that created and maintained them was initially assessed in relation to the Roman Empire, and more recently in a number of studies in the New World and in Pre-Roman Europe (see, for instance, [3, 4, 5, 6, 7, 8]).

In previous works, building on this literature, we took a further step and tried to infer aspects of the political organization of a region from the quantitative analysis of TTI.

We proposed a methodology to assess the relationship among political entities from the structure of transportation networks in some proto-historical case-studies [9, 10, 11].

Here, focusing on passengers transport (PTTIs), we present a generalised version of such a methodology, extending its applicability to current-time case-studies. In particular, this generalised approach allows to address questions like to what extent an infrastructure was conceived as an independent system rather than as an auxiliary network that complements another; or whether a particular PTTI was shaped according to the demographic spatial distribution.

The article is organised as follows. The Second Section summarizes our baseline methodology.

Afterwards, we discuss how to adapt both ingredients to the characteristics of current-time scenarios. In Section 3, we focus on how to process empirical data about PTTIs that are very different in nature from that of 'archaeological PTTIs'. For instance, present case studies are much richer in details whose relevance needs to be assessed and that could be treated in multiple ways.

In Section 4, we discuss what kind of hypotheses are suitable to be taken into consideration in modern and contemporary scenarios, connecting them with relevant research questions and integrating the knowledge available about the administrative and institutional organization of the systems under study. Such a reflection is translated to specific modifications to the original approach.

In Section 5, we discuss the results of applying our methodology to an illustrative case-study: the Catalan railway system, a regional PTTI with several decades of history that coexists with a road system. Special attention is paid to the assessment of the performance of the models and to how to draw conclusions properly from the comparison of empirical and synthetic data. Finally, the article closes with some concluding remarks. Our preliminary results show how this approach has good potential, calling for further research.

2 Baseline methodology

In previous works, we tackled the issue of quantitatively inferring aspects of the political organization of a region from the structure of transportation networks [9, 10, 11].

Our goal was to identify the nature of the interplay between different human communities, going beyond tasks usually addressed by archaeological quantitative analysis, such as establishing the existence of a certain degree of regional organization.

To this aim, we addressed the analysis of the decision-making processes prioritizing some paths over others by envisioning a methodology consisting of two fundamental ingredients:

(i) a procedure for extracting relevant quantitative data from road maps; (ii) formal models implementing alternative mechanisms for generating synthetic TTIs to be compared against the empirical ones.

The underlying idea is that some models reproduce relevant structural features of the empirical TTI with higher accuracy than others. In other words, they provide a higher quality explanation of the empirical evidence and, therefore, we assume them to be more likely to resemble the actual mechanisms of regional organization [9].

We adopted network science as a natural framework to address the interplay between connectivity and functionality of TTIs. Indeed, network science provides us both with tools to identify and measure structural characteristics of empirical TTIs and with a conceptual framework for formal model building [12, 13, 14].

The task of translating road maps into networks is not straightforward and can be performed in many alternative, not equivalent ways. Since we were studying inter-settlement interactions, we

needed our nodes to represent the human communities connected through the regional TTI. Then, as the simplest possible option, we established a bidirectional link between any two sites that were directly connected by a terrestrial route, with no other settlement in between. To include the geographical factor in a simple way, we represented sites as geo-localized nodes and assigned weights to the links according to the geodesic distance between the nodes they connected.

We designed a minimalist set up in which each node, at each step, expressed a preference concerning the new link to be established, according to a variably well informed assessment of costs and benefits. Then they could either compete against each other or reach an agreement about which connection was going to be built next.

It is worth stressing that our goal was not to understand why in a certain region there were more or less settlements, or more or less roads. On the contrary, we were addressing the question of why and how the settlements that existed in the region built those roads instead of others. Consequently, the models took the set of settlements with their corresponding geographic locations and the amount of available resources – here quantified as the total link length L_{tot} – as inputs, not as parameters to be fitted.

The process ended when the total length of the connections added was equal to the total link length of the corresponding empirical network. Consequently, any synthetic graph generated by a network model replicated the following characteristics of the corresponding empirical network: (1) the total number of nodes N, (2) its geographic density $\delta = L_{tot}/\sum_{i=1}^{N}\sum_{j=i+1}^{N} d_{ij}$, and (3) the average node strength $\langle s\rangle = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in V} l_{ij}$, where V is the set of neighbors of i and l_{ij} is the length of the link between i and its neighbor j.

Any other metric is, in principle, suitable to be used for comparing different synthetically generated graphs against their empirical counterpart.

2.1 The Equitable Efficiency Model (EE Model).

Out of all the models that we devised, one presented a higher explanatory power when applied to our proto-historical case studies (*i.e.*, it generated the synthetic networks more similar to the corresponding empirical ones). It equipped nodes with global information about their connectivity, but also with the ability to make coordinated decisions. More concretely, it assumed that each settlement knew the (weighted) length of each one of the existing path joining it with any other settlement. Then links were prioritized globally according to their normalized distance R that was calculated as follows:

$$R_i(j) = R_j(i) = R_{ij} = \frac{d_{ij}}{L_{ij}}, \tag{1}$$

where d_{ij} is the geodesic distance between node i and node j, and L_{ij} is the length of the shortest existing path between them. The normalized distance R can also be seen as the inverse of what is called the route factor or detour index [15]. If i and j are disconnected, *i.e.*, belong to different connected component, this means that there exist no finite length path between them. Hence

$$R_{ij}^{[d]} = \lim_{L_{ij}\to\infty} \frac{d_{ij}}{L_{ij}} = 0 \tag{2}$$

and the comparison between disconnected node pairs is performed by determining

$$\text{sign}\left(R_{ij}^{[d]} - R_{lm}^{[d]}\right) = \text{sign}\left(\lim_{L^{[d]}\to\infty} \frac{d_{ij} - d_{lm}}{L^{[d]}}\right) = \text{sign}(d_{ij} - d_{lm}) \tag{3}$$

which is obtained assuming that the path length $L^{[d]}$ is the same for all the disconnected nodes.

The function $R_i(j)$ balances costs and benefits, prioritizing those links that shorten long paths (large L_{ij}) while wasting little resources (short d_{ij}).

More concretely, the EE model follows a three-step procedure:

1. For each node i, all the $R_i(j)$ values are calculated.

2. Each node i proposes the creation of a link between itself and a node j^* such that the $R_i(j^*)$ was the minimum value among all the $R_i(j)$ (local interest expressed by node i).

3. All the proposals are ranked according to their R value and a link is created between the pair corresponding to the global minimum (coordinated decision-making).

Step 1, 2, and 3 are repeated until the summed lengths of all created links reaches that of the empirical system[1].

2.2 EE model with preferential attachment

The EE model considered all settlements to be on the same ground and the links to be built were selected among the individual preferences according to a fair criterion. In order to explore a slightly different scenario, where preferences of nodes with more and longer links were entitled to a higher priority level, we devised a variant of the EE model. Mathematically, such a bias is obtained by weighting the ratio R with a (negative) power of the strength (or weighted degree, *i.e.*, the total length of its adjacent links) of the proposing node, thus introducing preferential attachment among the nodes with the greatest strength. The trade-off between the two ingredients in determining the priority of each link is tuned by the exponent a of such power. Hence, the new value of the biased ratio R' for a connection between node i and j proposed by node i is:

$$R'_{ij} = R_{ij} s_i^{-a} = \frac{d_{ij}}{L_{ij}} s_i^{-a} \tag{4}$$

where s_i is the strength of node i. Therefore, when a is equal to zero, we recover the EE Model.

It is worth noting that in all the scenarios considered, including the last one, we assumed that all node-settlements were intrinsically equally important. When prioritising new links to be established, the node-settlements based their choice on geographical (distances) and topological (already existing links) information, but on node-settlement attributes (such as power, richness or attractiveness).

3 From maps to accessibility networks

In the case-studies considered, all of them from the Iron Age, the region under study was provided of a single PTTI that embedded the footprint of the relationships between the human communities living in the area. When human societies started building roads [16, 17], they created, in each territory, a network made of a combination of artificially adapted natural paths and manufactured ways that served for displacements at all the scales –from local to supra-regional– and for all means of transport (pedestrian, by wheeled vehicles, with animals). Then, each settlement had to be connected to others by means of this single PTTI, or it would be otherwise isolated. This in not the case for later scenarios. Already in the Roman Empire, the most important cities were connected through primary roads, while less important towns and villages were reachable thanks to a dense net of secondary roads and less manufactured pathways (see, for instance, [18]). Such a difference plays a crucial role both in the way the empirical network is constructed and in the modelling approach.

Considering their high construction costs, some PTTIs, such as railway systems or highways, are not designed to directly reach each single town and village in a territory. Instead, many human settlements benefit indirectly from the presence of a station or exit nearby.

Since one of our objectives is to capture the effect of population distribution over the territory on PTTIs structural patterns, the option to discard some groups of inhabitants based on whether their residence place is reached by the infrastructure under study or not must be ruled out.

We had therefore to disregard some of the most usual network representations, such as the so-called space-of-stations for railway systems, in which nodes are stations and links represent physical connections [19].

As an alternative, we sought for a method to integrate information about both population distribution and PTTI's connectivity in a single empirically-based graph (*i.e.*, a PTTI accessibility network). Specifically, we developed a procedure to assign fractions of the population to each train station, highway exit, etc. (hereafter, referred as station for the sake of simplicity) by 'merging' neighbouring localities, while redefining both their coordinates and connectivity.

[1]Such a procedure is equivalent to simply building, at each step, the link to the minimum of the R_{ij} matrix. Nonetheless, the metaphor of the individual priorities that have to be sorted is useful for devising meaningful generalization of the present baseline model.

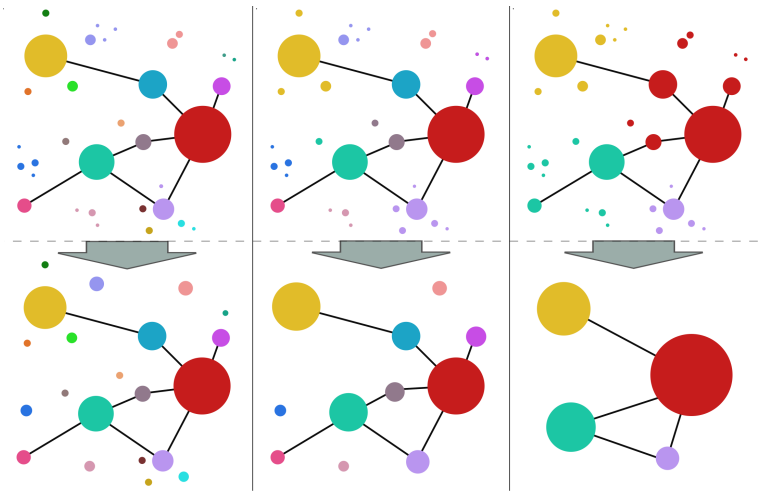


Figure 1: Schematic example of the merging procedure for the same initial network and a small (left), intermediate (centre), and large (right) value of Γ^*. Node size proportional to the population, colors according to merging results.

Our merging procedure grasps information about population distribution from the most fundamental units (*i.e.*, in general and hereafter, the municipalities) and their geographical positions. Initially, each municipality (with or without direct train connection) is represented as a node. The process takes into account population sizes and distance between municipalities to, iteratively, joining two nodes into one. In that way, low populated nodes are likely to be merged with closer, larger ones, in a location somewhere in the middle of both. Along the process, newly merged nodes preserve any connection existing in the previous step.

Formally speaking, the process is represented by an algorithm that gets as input information about the initial set of nodes i (geographic coordinates and population P_i) and the links among them, and executes the following steps:

1. It creates a list with all possible pairs of nodes i, j computing its inter-distances d_{ij}. The list is ranked in ascendant order according to the distances.

2. For the first element in the list, it calculates $\Gamma_{ij} = P_{\min} * d_{ij}$, where $P_{\min} = \min(P_i, P_j)$. This quantity expresses the balance between the importance of the smallest node and the distance that separates it from the largest one.

3. If the condition $\Gamma_{ij} < \Gamma^*$ is fulfilled, both nodes merge into one.

   The new node will have a population $P' = P_i + P_j$ and a new position (Lat', Lon') at the center of mass of i and j according to their former populations:

$$Lon' = \frac{P_i Lon_i + P_j Lon_j}{P_i + P_j} \quad , \quad Lat' = \frac{P_i Lat_i + P_j Lat_j}{P_i + P_j} \tag{5}$$

4. If there existed a link between the two merged nodes, the link disappears. If there were links adjacent to i or j, they are rewired correspondingly to the new node. Notice that those links would also change their length since the position of the merged node is now somewhere between former i and j positions.

5. The four previous steps are repeated until no more pairs of nodes are able to merge.

The parameter Γ^* fixes how lowly populated a node has to be and how close to another one more populated to merge them. It acts as a restrictive merging condition, being more permissive to nodes' merging as it increases its value. By means of the parameter Γ^*, we are considering the finite capacity of a station, allowing for a higher node density in highly populated areas, a necessary feature of almost any PTTI.

This method presents several advantages. It is independent on the territory and it is able to capture potential heterogeneity across a region (*e.g.*, rural and urban). More importantly, this method is able to define the scale of the system that suits our goals. In this sense, it is more flexible and less arbitrary than allocating nodes based, for instance, on administrative divisions of the territory (*e.g.*, provinces, counties...) and allows for an easier comparison between case studies.

Although different solutions are possible, in most of the cases it is sensible to set the value of Γ^* so that in the output the number of links matches, or slightly exceeds, the number of nodes. In principle, this condition allows to obtain a single connected component. Hence, isolates - if any - can be interpreted as representative of the population and geographical position of less favoured communities in terms of accessibility to the PTTI under study. Selecting a lower Γ^* would make the number of nodes to be greater than the number of links, thus forcing some nodes to stay disconnected and making this feature less meaningful. On the contrary, a considerably higher Γ^* might hide the existence of purely connected sub-regions by forcing further node merging.

4 Generalized modelling approach

The main difference between proto-historical scenarios and more recent case studies is the relevance of the context preexisting the construction of the infrastructure under study. The first steps of the algorithms of our baseline models consist in building links between isolated nodes that are completely equal, except for their geographical coordinates. The underlying hypotheses is that differences in power and importance did not preexisted the creation of the very first transportation infrastructure, before which there were just a number of (almost) disconnected settlements. However, if we want to model a PTTI that is not the first ever built communication network in the region, this supposition does not hold. In particular, we must take into account (1) the coexistence of more PTTIs, which implies that there exist alternative ways to reach each location, and (2) the uneven level of agency throughout the different communities in the considered territory. Hence, nodes in the initial state of the models are neither equal, nor disconnected. Here we address such issues keeping the algorithms as well as the input data as simple as possible.

4.1 Generalised Equitable Efficiency Model

Taking Eqs. 4 and 2 as a starting point, we introduce two modifications and define a new, generalized version of the EE model: The *generalized Equitable Efficiency Model* (hereafter referred as gEEM).We are interested in including the preferential attachment mechanism, since it proved to be useful when handling unbalanced settlements in terms of power[10]. However, instead of measuring the relative relevance across nodes using the node weighted degree (as in Sec. 2.2), gEEM uses an attribute of nodes that is now assumed to be a feature previous to the construction of the PTTI. Since our interest in this paper is focused on population distribution, here we are taking nodes' population. Notice, however, that this modelling approach could be applied using other attributes (such as nodes' wealth, for instance) depending on the requirements of each particular application. Thus, a new definition for the normalized distance is proposed:

$$R'_{ij} = \frac{d_{ij}}{L_{ij}}(p_i)^{-a} \quad , \qquad p_i = \frac{P_i}{P_{min}} \tag{6}$$

where P_i is the population (or any other desired attribute) of node i and P_{min} is the smallest population of all nodes.

On the other hand, we redefined the limit taken to evaluate the shortest path between two existing nodes, which the previous models took as $L_{ij} \to \infty$. We propose to make it finite defining it as:

$$L_{ij}^{[d]} = e^{-m} L_{CG} \tag{7}$$

where i, j are nodes in different connected components and L_{CG} is the total link length of a complete graph[2] built on the same set of geolocalized nodes. The parameter m controls the merging mechanism, making it less restrictive for larger values. By redefining this limit, our purpose is to reproduce the overall effect that other infrastructures have in shaping the railway network. Considering that between

[2]A graph is said to be complete when it exists a link between each pair of nodes, creating a fully connected structure.

two disconnected nodes there is a finite existing path, we make them reachable. Thus, this trait allows to consider them not so primordial for the network and to open the possibility to construct other links beforehand (between already connected nodes, for instance).

4.2 Network characterisation by structural metrics

In order to better characterise the empirical network resulting from the merging process and compare it with synthetic counterparts later on, we calculated several network properties that provide information about both the structure of the network and its influence on communication dynamics:

- *Average link length*: It is a useful metric for a basic characterization of the links in a weighted spatial network:

$$\langle l_e \rangle = \frac{1}{N} \sum_{i=1}^{M} (l_e)_i \tag{8}$$

  where M is the number of edges.

  Moreover, in order to get information about link length variability, we computed the standard deviation of this metric σ_{l_e}

- *Standard deviation of node strength*: The average value of the node strength $\langle s \rangle$ is, by construction, the same for the empirical network as for any synthetically generated counterpart, but its standard deviation σ_s can be used as an indicator of how balanced the node connectivity is.

On the other hand, we found reasonable to assess the efficiency of the empirical network in terms of its main functionality, namely the interchange of goods, information and people. Several ways to evaluate such property have been proposed [20, 21]. From the different definitions in this literature, we took one proposed specifically for weighted networks in [21]. This definition compares the existing shortest path between two nodes, L_{ij}, with the geodesic distance that separates them, d_{ij}, assuming that the information flows better when the first value approximates the second. For a single connection, this definition coincides with our normalized distance R and, as previously mentioned, can be seen as the inverse of what is called the route factor or detour index [15]. According to this approach, we calculated the global and local efficiencies of the network as follows:

- *Global efficiency*: For each pair of nodes i and j, it compares the path connecting them through the network (L_{ij}) with the ideal case corresponding to the straight line (d_{ij}). Then, the global efficiency is obtained by averaging this ratio over all pairs of nodes.

$$E_{\text{glob}} = \frac{1}{N(N-1)} \sum_{i \neq j}^{N} \frac{d_{ij}}{L_{ij}} \tag{9}$$

- *Local Efficiency*: For each node i, it evaluates how well information is exchanged between its first neighbours when i is missing. In other words, it assesses the robustness of the network and how it is able to deal with failures (for instance, when the communication through a specific node is not possible).

$$E_{\text{loc}} = \frac{1}{N} \sum_{i=0}^{N} \frac{1}{k_i(k_i-1)} \sum_{j \neq k \in \Gamma_i}^{N} \frac{d_{jk}}{L_{jk/i}} \tag{10}$$

  Where j, k belong to the subgraph of first neighbours of i, Γ_i, and $L_{jk/i}$ is the shortest path connecting j and k when the node i is removed.

  Finally, we defined a generalized version of the Global Efficiency that takes into account the number of inhabitants associated to each node.

- *Passenger Efficiency*: For each pair of passengers p and q that do not belong to the same node, it compares the path connecting them through the network (L_{pq}) with the ideal case corresponding to the straight line (d_{pq}). Then, the passenger efficiency is obtained by averaging this ratio over all pairs of passenger.

$$E_{pass} = \frac{1}{N_F} \sum_{p \neq q}^{P_{tot}} \frac{d_{pq}}{L_{pq}} = \frac{1}{N_F} \sum_{i \neq j}^{N} P_i P_j \frac{d_{ij}}{L_{ij}} \tag{11}$$

Where $N_F = P_{tot}^2 - \sum_i^N P_i^2 = N(N\bar{P}^2 - \overline{P^2})$ is the total number of pairs of individuals discarding those associated to the same node, being P_{tot}, $\bar{P}$, and $\overline{P^2}$ the total population of the system, the mean node population and the mean square node population, respectively. The passenger efficiency is equal to the global efficiency when all the nodes have the same population.

5 Application of our generalised methodology to a present-day case study

This section illustrates the applicability of our generalised methodology (including the merging procedure to obtain the empirical network and gEMM as the model generating synthetic counterparts) to present-day scenarios.

5.1 Empirical accessibility network from node merging

The specific territory used in this study to illustrate the applicability of our generalised methodology was Catalonia, a Mediterranean region in northeast Spain. This region is densely populated and presents heavy unbalance between rural and urban areas, with more than 70% of its population living in cities. It also has a relevant industrial sector and diverse and developed transportation infrastructures. Among them, the railway network. Such infrastructure can be divided into high-speed railroads and regional services. The former is actually part of a larger-scale structure, which has been framed according to a country-wide perspective (*i.e.*, the Spanish high speed rail service *AVE*). Therefore, here we consider only the regional service, which is, composed of two structures managed by two different operators (*ADIF* and *FGC*) owned by the Spanish and Catalan regional governments, respectively.

In Figure 3 (left), we see Catalonia's regional railway networks, together with the municipalities served by them. As already pointed out in Section 3, many localities in Catalonia (and their corresponding share of the population) are not directly served by the railway networks. This justifies obtaining a train accessibility network as our empirical network instead of a physical railway one (*i.e.*, the space-of-stations representation).

In this particular case, we chose to set the threshold parameter Γ^* so that it is potentially possible for the resulting total link length to connect all the nodes in a single connected component. Starting from an initial state in which nodes outnumber links 945 to 205, the merging process progressively reduces the number of nodes, with some sporadic link losses as a collateral consequence. Meanwhile, the total link length of a hypothetical minimum spanning tree built on the same set of geolocated nodes also decreases, and does it faster than the total link length. Depending on the value of Γ, the resulting final network has different number of nodes $N(\Gamma)$ and links $M(\Gamma)$, and different total link length $L_{tot}(\Gamma)$ and minimum spanning tree length $L_{MST}(\Gamma)$. In general, the larger the value of Γ, the smaller the ratios $N(\Gamma)/M(\Gamma)$ and $L_{MST}(\Gamma)/L_{tot}(\Gamma)$, until they both reach a plateau. Once the condition $L_{MST}(\Gamma)/L_{tot}(\Gamma) < 1$ is satisfied, the EE model always build a connected network, and actually if the ratio is equal to 1 it produces the MST. However, in order to allow the present generalized version of the model to generate connected networks different from the MST, we chose a value Γ^* such that the additional condition $M(\Gamma^*)/N(\Gamma^*) < 1$ (*i.e.*, the number of links in the resulting empirical network exceeds the number of nodes) is also satisfied.

This is just a heuristic criterion, but has the clear advantage of not forcing the graph to have isolates while permitting their existence. In this way, the "surviving isolates" in the empirical network represent a meaningful trait of the system under study, not just a mere consequences of a wrong scale choice. We can thus apply our methodology to investigate the reason behind network features such as isolated notes, multiple connected components, and cycles (closed loops).

The empirical network is showed in Figure 3, before (left) and after (right) applying the merging procedure setting $\Gamma^* = 0.82$.

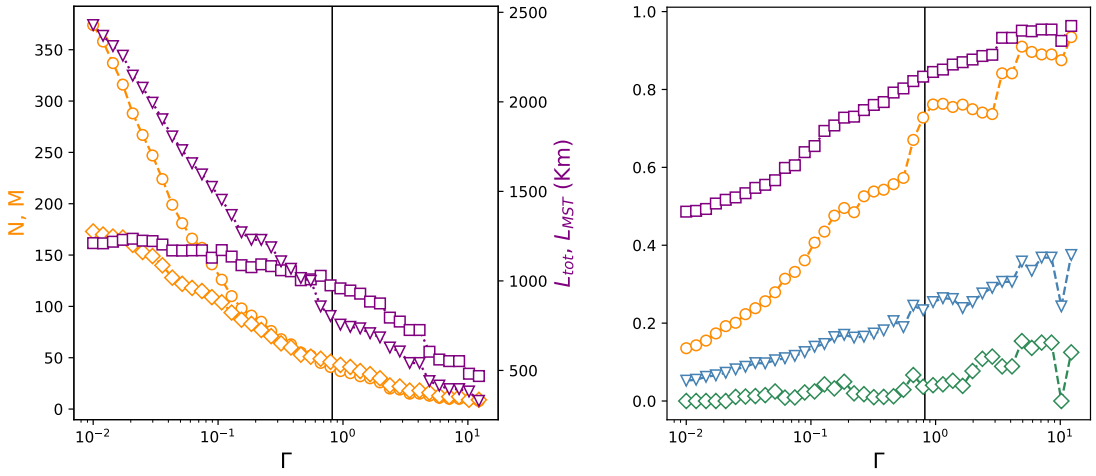


Figure 2: Left Panel. The number of nodes N and edges M in the empirical network (right axis) and the corresponding total link length L_{tot}, together with the length of the minimum spanning tree L_{MST} built on the same set of geolocalized nodes (left axis), are plotted as functions of the parameter Γ. Right Panel. Green diamonds: clustering coefficient; Blue triangles: local efficiency; Orange circles: global efficiency; Purple squares: passenger efficiency. Vertical line: Γ^*

	N	M	$\langle C \rangle$	$\langle l_e \rangle$	σ_{l_e}	σ_s	I	E_{glob}	E_{loc}	E_{pass}
$\bar{x}$	45.8	38.5	0.068	25.5	8.9	34.1	11.9	0.403	0.150	0.469
σ	2.3	3.1	0.034	1.2	1.3	3.0	2.1	0.043	0.042	0.098
P	1	0.035	0.795	1	0	0.655	1	0	0.01	0
e	39	44	0.038	21.7	1.3	33.1	2	0.732	0.245	0.834

Table 1: $\bar{x}$ stands for mean values, σ are the corresponding standard deviations, and P their p-values. In the line labelled **e** we report the values calculated on the empirical network... All metrics have no dimension except for $\langle l_e \rangle$, σ_{l_e} and σ_s, which are expressed in Km. In bold, the highest value for each metric in the table.

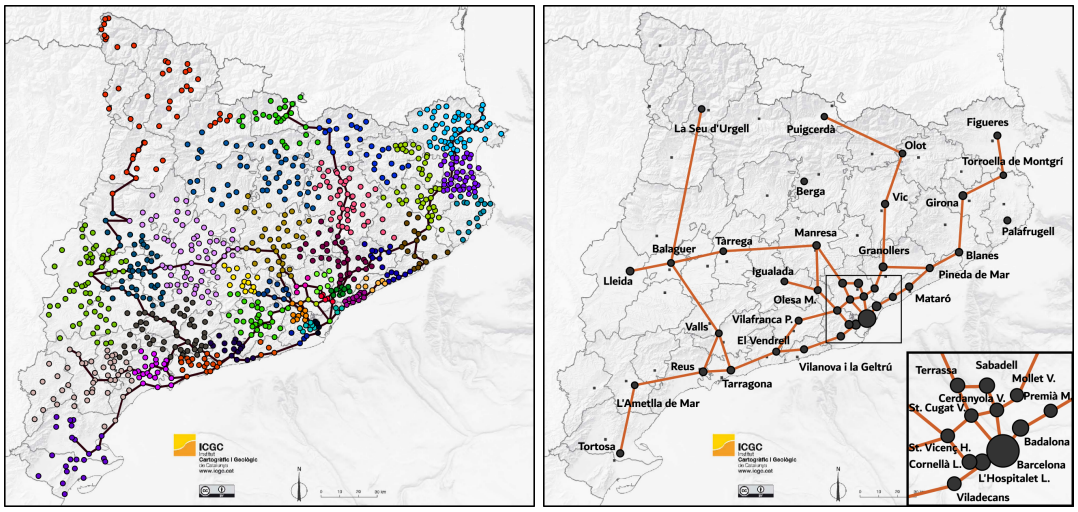


Figure 3: Left: Municipalities ($N = 945$) and train connections ($M = 205$) before the merging procedure. The initial total link length is $L^0_{tot} = 1208.51\,(km)$ and the total length of the corresponding MST is $L^0_{MST} = 3737.26\,(km)$. Colours according to merging results. Right: Empirical network after merging procedure for $\Gamma^* = 0.82$. Labels assigned according to the most populated municipality. A node size is proportional to the total population of the municipalities inside it.

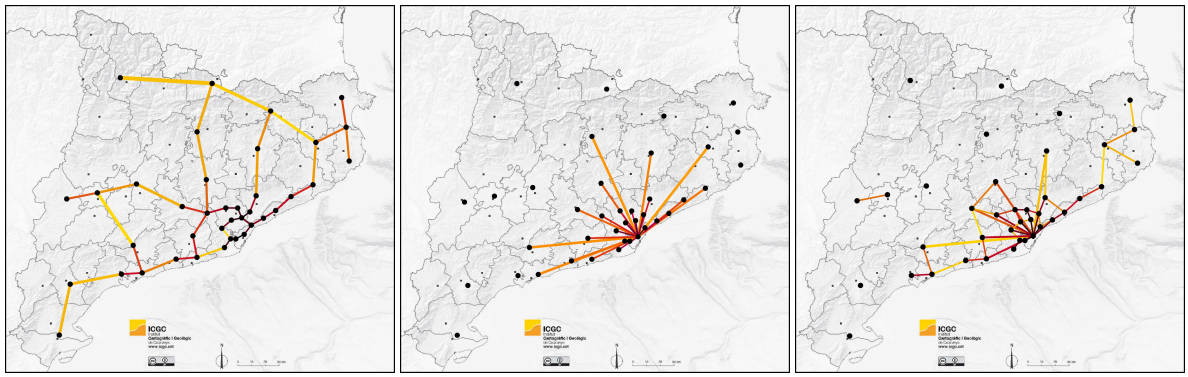
5.2 Synthetic networks generated by the gEMM

Given a node layout $\mathcal{L}$, *i.e.*, set of geolocalized nodes, and a total link length L_{tot} such that $L_{tot} > L_{MST}(\mathcal{L})$, by varying the two parameter a and m, the gEMM is able to produce a great variety of network topologies, ranging from star-like graphs to disconnected cliques (all-to-all sub-graphs) and almost regular lattices, which may include a variable amount of isolates and connected components of different sizes.

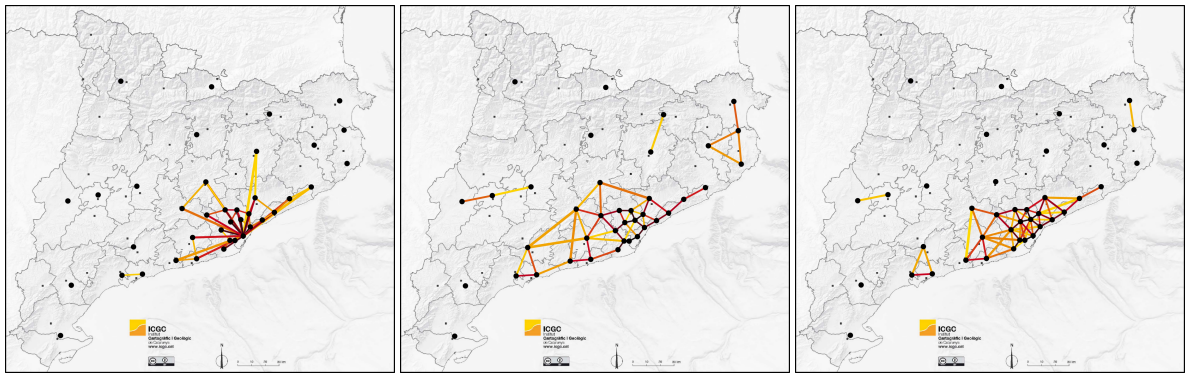
To asses how some basic structural properties of the model's outputs depend on m and a, we applied the gEMM to the node layout and total link length of the empirical network obtained in Sec. 5.1 varying both parameters.

Since the number of different network topologies that can be constructed on a given node layout for a fixed total link length is finite, the parameter plane $m - a$ can be ideally partitioned into regions corresponding to an identical, unique output of the model. However, such an exhaustive exploration is beyond the scope of the present study. At a more coarse-grained scale, we observed that whenever $m \lessapprox 4.5$, this parameter does not affect the output of the model, that is, the output for any (m, a) is the same as for $(4.5, a)$ for any a. Something analogous happens for the region $m \gtrapprox 10$ where the output for any (m, a) is the same as for $(10, a)$ for any a. Similarly, for $a \gtrapprox 1$, the model does not produce any topological novelty and any network created for (m, a), with $a > 1$, is also generated for at least another pair (m', a'), where $m' < m$ and $a' < 1 < a$.

In particular, for $m \lessapprox 4.5$ and $a = 0$, the output of the gEEM is the same as that of the original EE model (Fig. 4a). The first connections to be created belong to the MST, afterwards the algorithm adds few shortcuts until the total link length equals that of the empirical network. When the value of any of the two parameters is increased, some of the links belonging to the MST start to be rewired and some nodes may be left disconnected from the largest component of the resulting network. While the principal effect of increasing the value of m (Figs. 4b,4c,4d,4e,4f) is the appearance of closed triangles, larger values of a (Figs. 4b,4c,4d) force the links to be redirected towards most populated nodes.



(a) $m = 4.0$, $a = 0.0$. (EEM model) (b) $m = 4.0$, $a = 1.0$. (c) $m = 6.7$, $a = 0.41$



(d) $m = 7.2$, $a = 0.71$ (e) $m = 7.2$, $a = 0$ (f) $m = 7.7$, $a = 0$.

Figure 4: Artificial networks examples.
Artificial networks for different pairs of values for the parameters are represented. Edges' color illustrates the order of construction of the links by the algorithm. Darker colors stand for earlier edges while brighter stand for later ones.

5.3 Comparing synthetic and empirical networks

m	a	$\langle C\rangle$	$\langle l_e\rangle$	σ_{l_e}	σ_s	I	E_{glob}	E_{loc}
≤ 4.0	0	0	22.1	13.5	30.5	0	**0.800**	0.192
≤ 4.0	1.0	0	**36.9**	**24.6**	**153.1**	12	0.358	0
6.7	0.41	0.120	29.8	20.8	125.2	7	0.336	0.123
7.2	0.71	**0.431**	21.8	15.2	103.0	14	0.431	0.281
7.2	0	0.208	18.2	9.63	37.4	5	0.208	**0.389**
7.7	0	0.360	15.9	6.67	42.6	11	0.360	0.285
≥ 10.0	0	0.295	16.0	6.45	75.3	**27**	0.295	0.089
≥ 10.0	1.0	0.299	16.7	7.25	77.6	**27**	0.299	0.086
Empirical		0.039	21.7	14.9	33.1	2	0.732	0.245

Table 2: Network metrics values.
For different representative pairs of values for m, a, the corresponding metrics are shown, along with those corresponding to the empirical merging network. All metrics have no dimension except for $\langle l_e\rangle$, σ_{l_e} and σ_s, which are expressed in Km. In bold, the highest value for each metric in the table.

The empirical network presents an almost vanishing clustering coefficient and a considerably high Global efficiency ($E^{emp}_{\text{glob}} = 0.732$), just slightly lower than the EEM network (see Table 2). When comparing with the network topologies generated by the gEEM, we find that these two feature can be observed when $a \lessapprox 0.4$ and $0.55 \lessapprox m \lessapprox 0.65$ (see Figs. 5a and 5b). If we look at other metrics such as the Local efficiency ($E^{emp}_{\text{loc}} = 0.245$), the average link length ($\langle l_e\rangle^{emp} = 21.7\,km$), and the standard deviation of the node strength ($\sigma^{emp}_s = 14.9\,km$), similar values can also be found in the same region of the parameter space.

Additionally, we computed a network distance between each synthetic topology and the empirical network based on the shortest path length between node pairs:

$$D_m = \frac{2}{N(N-1)} \sum_{ij} 2 \frac{|L^{emp}_{ij} - L^{synt}_{ij}|}{L^{emp}_{ij} + L^{synt}_{ij}}, \tag{12}$$

where L_{ij} is the sum of the lengths of the links in the shortest path between i and j if both nodes belong to the same connected component, and $L^{[d]}_{ij} = e^{-m} LCG$ otherwise. For disconnected nodes in the empirical network, we assume the same value of m that generated the synthetic counterpart we are comparing it to.

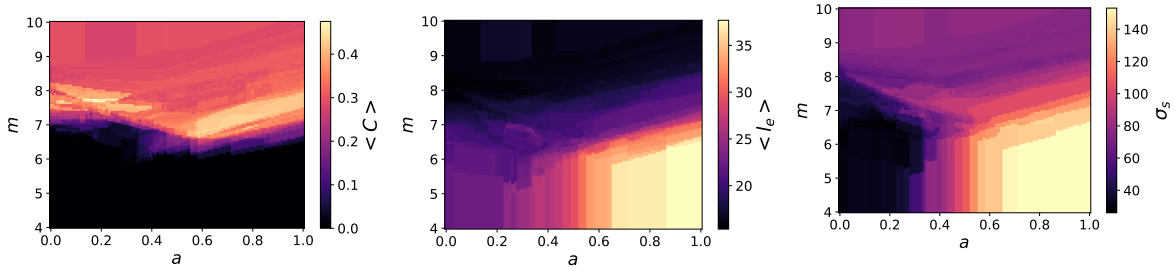
Once more, it is confirmed that the most similar topologies can be obtained by setting $m \in [0.6, 0.65]$ while keeping the value of a small, namely $a \lessapprox 0.35$.

For such values of m, we have $L^{[d]}_{ij} \in [93, 153]\,(km)$, that is, larger than the average (geodesic) distance between node locations $\langle d\rangle = 83.5\,(km)$, and also larger than the largest link in the empirical network $l^{max}_e = 85\,(km)$, but considerably smaller than the maximum distance between nodes $d_{max} = 269.5\,(km)$.

In the system, for the selected value of merging parameter Γ^*, there are 12 nodes (30.7%) whose average geodesic distance to other locations is in the range $[93, 153]\,(km)$ or slightly larger. For these nodes, the underlying assumption of the gEEM that the overall preexisting connectivity allowed them to reach any other place in the system through a path of length $L^{[d]}_{ij}$ is, in this range of the parameter m, incorrect and, from the viewpoint of the infrastructure design, unfair.

In particular, the presence of isolates is a trait that can be observed in synthetic topologies only if $m > 5$ and/or $a > 0.35$.

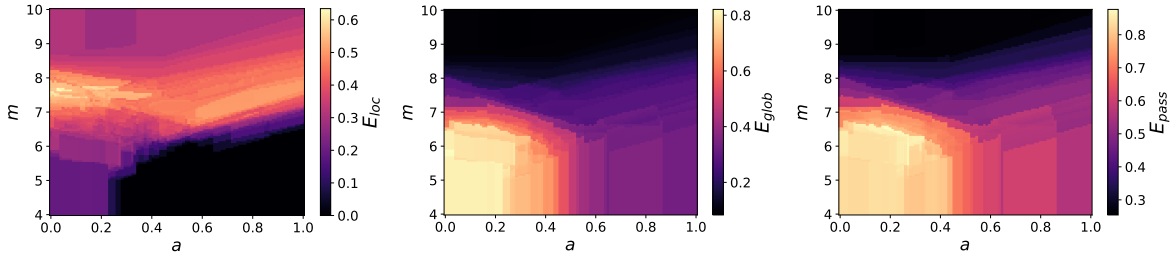
Therefore, we can conclude that for the model to be able to reproduce this and other features of the empirical network, a non negligible share of unfairness towards the periphery of the system is required,



(a) Average clustering, $\langle C \rangle$

(b) Average link length, $\langle l_e \rangle$

(c) Weighted degree (strength) standard deviation, σ_s



(d) Local efficiency, E_{loc}

(e) Global efficiency, E_{glob}

(f) Passenger efficiency, E_{pass}

Figure 5: Network properties in artificial networks.
For different values of a, m artificial networks are created and their properties are computed.

either underestimating geographical distances (m), or explicitly favouring the most populated locations (a), or both. Taking into account the presence of alternative or preexisting TTIs in the same territory in a more realistic way ($m < 6$) does not produce synthetic networks that are close enough to the empirical one.

At a more local scale, the first node to get disconnected in the topologies generated by the gEEM is *La Seu d'Urgell* (see Fig. 3), which has the smallest merged population associated to it ($P = 53154$) and the furthest distance to the nearest neighbor ($d = 66.331\,km$). The isolates in the empirical networks have quite larger populations ($P(Palafrugell) = 113773\ km$ and $P(Berga) = 84238$) and nearer neighbors (at 24.607 km and 35.379 km, respectively). Although they both are far from the top ranking nodes, the features considered by the present version of the model do not make them plausible candidates to be the sole disconnected nodes of an otherwise connected topology. Possibly, the explanation for their specific condition could be found in the details of the historical process that shaped the TTI under study in its current state. For instance, until 1973, there existed a railway connection *Berga-Manresa* [22], a link that the gEEM builds in most of the topologies that are similar to the empirical network.

6 Conclusions

We have presented a methodology to shed light on the mechanisms, power balances, and competing interests that shaped a given TTI into its current configuration. Relying on the analytical and theoretical toolbox provided by network science and adopting an inverse engineering approach, we propose a two-step procedure that allows us to address a broad variety of systems.

The first step consists in mapping the infrastructure under study into a geographic network in a flexible way, adjusting the spatial scale of the representation to the specific situation of the system and its function by tuning a single parameter.

The second step enables us to investigate how different the considered system is from an ideal model of TTIs that has proven to be able to capture most of the relevant features of ancient proto-historical networks of pathways, that is, the EEM. The main characteristics of such a model are (i) the equitable treatment of the necessities and interests of each node-place in the system, regardless of its real power

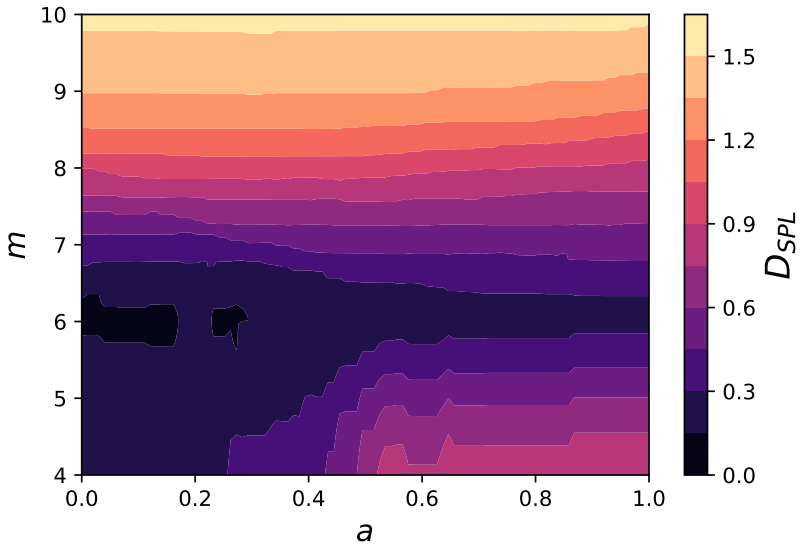


Figure 6: Shortest path length distance between empirical and artificial networks as a function of parameters a and m.

or importance; (ii) the assumption that if in the network there is no path connecting two nodes, then it is impossible to reach one place from the other, that is, the considered network represent the only TTI existing in the territory.

We have devised a mechanistic network model, the gEEM, that using the EEM as a starting point, progressively and independently relaxes both its main assumptions by means of two parameters.

Finally, we have shown how to effectively infer information about a real case-study applying the proposed methodology to an illustrative example, *i.e.*, the regional railway connections in Catalonia. Comparing the output of the gEEM for different values of the parameters to the empirical network of the Catalan regional train service, we evinced that network topologies similar to the empirical one can be generate if the most geographically central and demographically important places are slightly favoured when designing the TTI. These conclusions, which are in line with common knowledge but translate it into more precise quantitative terms, confirm the potential of simple mechanistic models as a powerful explanatory instrument for tackling complex systems.

References

[1] Alfonso Herranz-Loncán and Johan Fourie. "For the public benefit"? Railways in the British Cape Colony. *European Review of Economic History*, 22(1):73–100, 08 2017.

[2] F. Pablo-Martí, Á. Alañón-Pardo, and A. Sánchez. Complex networks to understand the past: the case of roads in Bourbon Spain. *Cliometrica*, 15:477–534, 09 2021.

[3] Raymond Chevallier. *Roman Roads*. University of California Press, Berkeley and Los Angeles, 1976.

[4] Christopher Taylor. *Roads and Tracks of Britain*. Orion, London, 1979.

[5] C. D. Trombold. *Ancient Road networks and settlement hierarchies in the New World*. CUP, New York-Cambridge, 1991.

[6] D. Jenkins. A network analysis of inka roads, administrative centers, and storage facilities. *Ethnohistory*, 48:655–687, 2001.

[7] Monica L. Smith. Networks, territories, and the cartography of ancient states. *Annals of the Association of American Geographers*, 95(4):832–849, 2005.

[8] M. R. Groenhuijzen and P. Verhagen. Testing the robustness of local network metrics in research on archeological local transport networks. *Front. Digit. Humanit.*, 3(6), 2016.

[9] Luce Prignano, Ignacio Morer, Francesca Fulminante, and Sergi Lozano. Modelling terrestrial route networks to understand inter-polity interactions (southern etruria, 950-500 bc). *Journal of Archaeological Science*, 105:46–58, 2019.

[10] Francesca Fulminante, Luce Prignano, Ignacio Morer, and Sergi Lozano. Coordinated and unbalanced powers. how latin cities shaped their terrestrial transportation network. *Front. Digit. Humanit.*, 4(4), 2017.

[11] Francesca Fulminante, Sergi Lozano, Luce Prignano, and Ignacio Morer. Why rome and not veii? analysing geographical networks in etruria and latium vetus between the early iron age and the archaic era. In Helen Dawson and F. Iacono, editors, *Bridging Social and Geographical Space through Networks, Proceedings of the Topoi Conference*. Sidestone Press, Leiden, 2021.

[12] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175 – 308, 2006.

[13] Mark E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45:167–256, 2003.

[14] Luciano da Fontoura Costa, Osvaldo N. Oliveira Jr., Gonzalo Travieso, Francisco Aparecido Rodrigues, Paulino Ribeiro Villas Boas, Lucas Antiqueira, Matheus Palhares Viana, and Luis Enrique Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.

[15] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.

[16] Timothy Earle. Paths and roads in evolutionarly perspective. In Charles D. Trombold, editor, *Ancient road networks and settlement hierarchies in the New World*, pages 10–17. CUP, New York-Cambridge, 1991.

[17] Maxwell Lay. *Ways of the World: A History of the World's Roads and of the Vehicles That Used Them*. Rutgers University Press, Piscataway, New Jersey, 1992.

[18] Cèsar Carreras and Pau De Soto. The roman transport network: A precedent for the integration of the european mobility. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 46(3):117–133, 2013.

[19] Maciej Kurant and Patrick Thiran. Extraction and analysis of traffic and topologies of transportation networks. *Phys. Rev. E*, 74:036114, Sep 2006.

[20] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87:198701, Oct 2001.

[21] I. Vragović, E. Louis, and A. Díaz-Guilera. Efficiency of informational transfer in regular and complex networks. *Phys. Rev. E*, 71:036122, Mar 2005.

[22] Jaume Perarnau i Llorens. Carrilet manresa-berga. aproximació a les influències socioeconòmiques, el. *Dovella*, (3):9–14, 1981.