UNIVERSITATEA BABEŞ-BOLYAI
BABEŞ-BOLYAI TUDOMÁNYEGYETEM
BABEŞ-BOLYAI UNIVERSITÄT
TRADITIO ET EXCELLENTIA

Statistical Physics Methods for understanding Complex Networks

a dissertation presented
by
István Papp
to
The Department of Physics

Scientific supervisor
Professor Dr. Zoltán Néda

Babeş-Bolyai University
Cluj-Napoca, Romania

April 2020

Statistical Physics Methods for understanding Complex Networks

Abstract

Computers went through a fast evolution in the last few decades and this also led to big advancements in data based sciences. While databases grew quickly the scientific community learned to handle easier large amount of data. This phenomenon also affected network sciences. Networks grew with the amount of information gathered and completely changed how the scientific community thinks about networks. Complex systems nowadays commonly have network descriptions, while in almost all domains from biology to computer science, engineering, economics, politics, etc. rises at least one question, where the answer is included in finding the community structure. Traditional definition of these communities are based on their connectivity, members of a community have more connections within the community than with the rest of the network. We implemented the concept of Voronoi diagrams used mainly to divide geometric space onto network partitioning. We defined a metric system using the edge clustering coefficient as distances between nodes, and introduced a local density measure based on these distances to identify the Voronoi generator nodes. Then we updated this method with a generalization of the seeds by selecting them randomly and perform the Voronoi partitioning multiple times with different sets of seeds, giving the ability of the algorithm to reach a fuzzy clustering property. We also studied human and information mobility through various experiments with some human travelling modes (road and air transit) and data transferring measurements on the Internet. We examined the average speed as function of the geodesic distance and transmission times of messages depending on the distance. The results suggested a sub-linear trend in all cases. The cause of these trends is not only due to the networks' structure, delays also exist on individual nodes including the start and ending points. For a better understanding of the Internet network's features we introduced a model which was capable of reproducing both the structure and the observed dynamic scaling characteristics. The thesis contains three chapters: a motivation and introduction to the field; a theoretical and mathematical description of network clustering, arguing the benefits of networks embedded in space; and the last chapter contains experiments and modelling on mobility networks and the Internet.

Table of Contents of the Thesis

Table of Contents of this Summary

Keywords

complex networks, random graphs, scale-free networks, network topology, complex systems, community detection, Voronoi diagrams, stochastic graph Voronoi diagrams, scaling laws, universalities, Internet

1

Introduction

Social science, information theory, technology, biology, neuro-science, etc., all study systems that can be represented as networks, and graph analysis has become crucial in understanding the features of these systems [1, 2].

In almost all systems with graph representations rises at least one problem, where finding the community structure of the network might play a key role in understanding it. An immediate problem is distinguishing importance of publications in a citation network with overlapping disciplines. Communities also exist in several networked systems from biology to computer science, engineering, economics, politics, etc.

Usually communities are defined as vertices connected more densely within their group than compared to their average connectivity in the graph. The challenge in identifying any community structure could be in the nature of its description, as it is qualitative, no widely accepted mathematical definition has been developed yet [3]. While a large variety of community definitions and detection methods exist [3], combining meaningful mathematical definitions with computationally efficient algorithms remains a problem. Similar clustering problems also occur in data mining, pattern recognition, machine learning and statistical data analysis [4, 5]. However, they are defined in continuous metric spaces, leading to a simpler formulation. Voronoi diagrams [6] are used commonly to divide metric space into subsections, called Voronoi cells.

Working with networks embedded in metric spaces are very important, not only for good clustering, but they surround us in nature. Even human mobility and our communication channels (as an immediate example the Internet) are networks embedded in metric space. From our everyday experience we have learned, that the travelling time does not scale linearly with the travel distance[7–9]. The cause of this is more intricate and includes various effects. To travel greater distances we use highways, but most of the time is taken up by getting in and out of cities, hence

closer cities might be connected also by increased travelling time. Similar situation is observable on air traffic, large portion of the time is spent on taking off, landing and parking. The Internet on the router level is also a complex network embedded in geographical space. Beside it's topological scaling properties (scale-free degree distribution) [10, 11], it also exhibits a dynamical scaling, similar to the human mobility. Here we investigate these effects more extensively and search into their observable causes using a variety of online databases and GPS tracking, and a few experiments based on Internet control message protocol: Ping and Traceroute.

To explain this novel scaling law and other measurable topological properties of the Internet a realistic model has to be built. Such a model must be based on realistic assumptions regarding the wiring process and has to reproduce the measured topological properties of the Internet, including the observed scaling of the communication speed with the distance.

The thesis contains two main chapters. The first contribution deals with mathematical description and application of our network clustering method. The second chapter contains experiments and modeling regarding the scaling properties of travel networks and the Internet.

2

Community detection in complex networks

2.1 Detection by graph Voronoi diagrams

To present the reasoning behind our clustering technique, we first demonstrate its essential aspects on a community detection problem defined in two-dimensional Euclidean space (see figure 2.1(a)).

Let us select $G = (V, E)$ as a weighted, directed graph with V of N vertices and a set E of M links. We denote by $l(\mu, \nu) > 0$ the weight of a link connecting vertex μ and ν. The length of a path is acquired by adding up the weights of links constructing the path. We denote the distance linking two vertices ν_i and ν_j as $d(\nu_i, \nu_j)$, which is the length of shortest path connecting them. This definition of a link length ensures that the network can be embedded into a metric space. Naturally, the more straightforward selection of $l(\mu, \nu)$ is 0 when μ and ν are not connected; 1 when they are directly joined. We select a group of $S \equiv (\gamma_1, \gamma_2, \ldots, \gamma_g) \subset V$ generator vertices. The resulting Voronoi partitioning of the network G respecting S will be the splitting of V into vertex groups $V_1, V_2, \ldots, V_g \subset V$, where each group (Voronoi cell) belongs to a generator and satisfy: i) Network G contains all Voronoi cells without overlapping; ii) All vertices in a cell are closer to its generator vertex than any other seed. Detailed mathematical characteristics of Voronoi graph diagrams are described in [6] along with distinct identification techniques and their corresponding computational complexity.

To make efficient gain of our geometric approach, we need appropriate definition for distance measurement that transforms vertex membership into segregation in metric space. Here we have

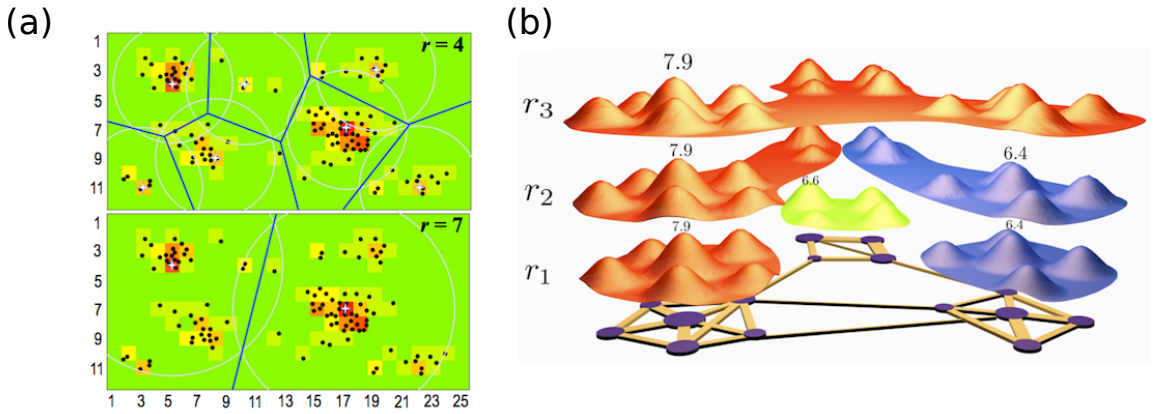

Figure 2.1: Locating Voronoi cell centers. (a) In order to find the best partitioning of the black points in clusters we divide the plane in 12×25 squares and estimate the local density (number of points) in every square. The squares are colored according to their density from green (0 dots) to red (largest 6 dots). The square with the highest density in its vicinity within a radius (gray circles, $r = 4$ up and $r = 7$ down in square units) becomes a generator square (marked white star). The Voronoi tessellation is symbolized with blue lines. (b) Illustrating the mapping of algorithm on networks: every node will have a local density, shown by the size of the vertices and the mountains in the community labels. Lengths of connections are proportional with inverse edge-clustering coefficient (marked by width of connections). Seed nodes are indicated by their local density having the largest in their neighborhoods with radius $r_1 < r_2 < r_3$. Clusters sequentially merge together as r radius increases.

chosen to adjust one of the easiest, generally accepted and computationally very efficient measure: the Edge Clustering Coefficient (ECC) proposed by [12]. The ECC of a link connecting vertex i and vertex j is defined as

$$C_{i,j} = \frac{z_{i,j}}{min[(k_i - 1), (k_j - 1)]} \tag{2.1}$$

where k_i, k_j are the degrees of these two vertices, $z_{i,j}$ is the number of triangles to which the link belongs and $min[(k_i-1), (k_j-1)]$ is the number of possible triangles to which it could belong, since it is the lower value of the degrees of the two adjacent vertices, minus one (the link examined). The lower the ECC, the more probable it is to connect nodes in distinct clusters. Therefore, in our graph Voronoi partitioning method, we specify connection length (weight) as inverse of the ECC.

Additional objective is to pick one seed vertex in each community. We have chosen a generator node selection procedure using relative local density of the vertices [3], defined as:

$$\rho_i = \frac{m}{m + k} \tag{2.2}$$

where m is the internal degree (number of incoming links) of the neighbourhood a sub-network containing first neighbours of vertex i, and k is the external degree (number of outgoing links) of the neighbourhood. This density is higher for nodes inside the center, dense part of communities, as illustrated in figure 2.1(b). Just like in 2D space, generator vertices on the network will be selected as the nodes with largest local density within the region of radius r. Overall, the detection complexity of generator vertices remains much below $O(Mg)$ (for g Voronoi seeds), and partition-

ing has complexity of $O(N \log N)$ [6], for detailed computational efficiency check the thesis.

As the publication by Fortunato et al. reasoned in [3], it is necessary to test community detection methods both on benchmark graphs (produced with predetermined communities, ground-truth) and on real-world networks. For benchmark testing please take a look at the thesis.

Detecting the community structure of real-world graphs however is not trivial. In addition, efficiency of any method is more difficult to evaluate as the information of ground-truth is not available. New algorithms are therefore tested in comparison with previously acknowledged methods through a quality measure. To be more specific, we have selected the modularity. The modularity is defined as follows: $Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(V_i, V_j)$, here summing up over all pairs of nodes. A is the adjacency matrix, m the total number of links in the network, and P_{ij} stands for the expected number of links between nodes i and j in the null model. Here meaning, the random graph model. The δ function returns 1 if i and j nodes are in the same community that is $V_i = V_j$, returns 0 otherwise [13]. We evaluated our algorithm on several real-world networks that are regularly used in the literature, structurally and originally they are very distinct:

1. The well known network of friends inside the group of 34 members in the Zachary karate club connected with 78 links [14].

2. Network of neurons from the nematode Caenorhabditis elegans [15, 16] containing 297 vertices and 2359 connections.

3. Protein-protein interaction network of yeast [17] comprising of 1845 nodes connected by 4405 links.

4. A revised version of the collaboration network between scientists on condensed matter archive at www.arxiv.org. This network consists of 39576 vertices and 175692 edges, constructed on preprints which were published in the archive during 1 January 1995 – 31 March 2005 period [18].

5. Web of connections between American political blogs [19] including 1223 vertices and 19087 links.

We compared our algorithm with five commonly used methods (for more information see Addendum): 1) the Louvain algorithm optimizing modularity [20]; 2) the label propagation algorithm (LPA) [21]; 3) GANXiS or SLPA (speaker-listener label propagation algorithm) [22, 23]; 4) link-communtiy detection [24] and 5) Infomap (IM) [25, 26]. Our approach performed as the second best, the modularity it reaches is always above average. The only algorithm that has better performance is Louvain, it was also expected, since it optimizes the very same quality function we use to evaluate performance.

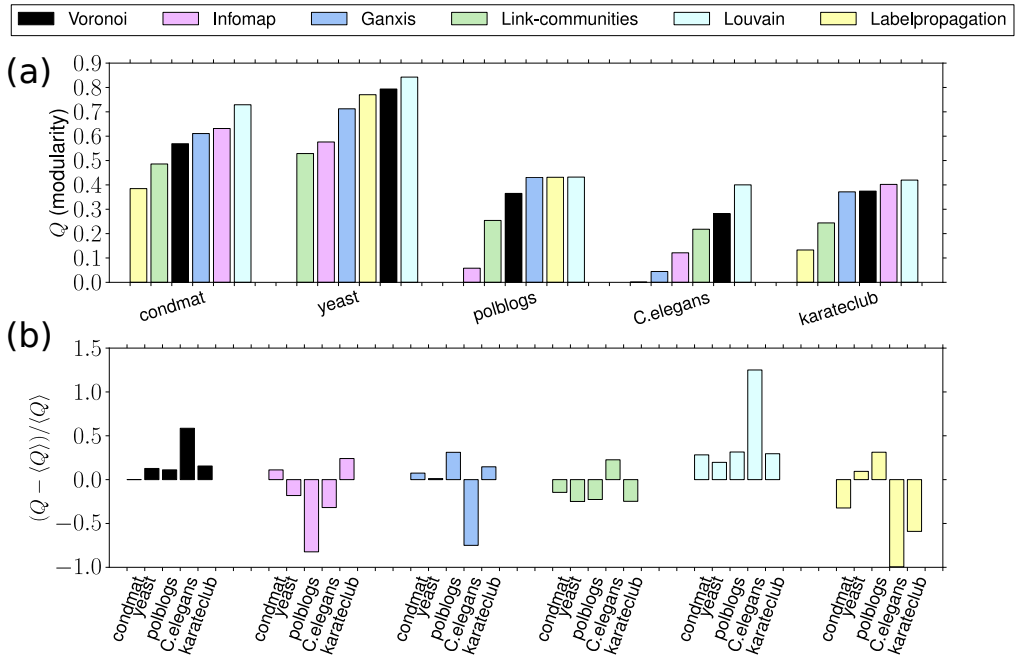

Figure 2.2: Real-world network testing. Modularity obtained on all five real-world networks with distinct methods (specified in the legend). (a) Q modularities in increasing order on every tested network.(b) Shows the relative error $(Q - \langle Q \rangle)/\langle Q \rangle$, here $\langle Q \rangle$ represents the average of results for all algorithms computed apart for every network.

2.2 Detection by stochastic Voronoi diagrams

For a given undirected network with N vertices and M edges, we neglect the careful selection of Voronoi seeds, instead we use a set of Voroni cells, where each set is obtained from arbitrarily picked generators [27], with the following main steps: i) We randomly select a number of g vertices from the network and use them as Voronoi seeds to perform a graph-Voronoi partitioning. ii) We calculate the Voronoi cohesion matrix or the co-location probabilities from averaging over the co-location matrix by repeating the tessellation R times (see figure 2.3).

For testing the method we took large non-directed random graphs with $N \to \infty$ vertices organized in m non-overlapping but connected Erdős-Rényi (ER) type communities of size $N_i \equiv \alpha_i N,\ 0 < \alpha_i < 1$ and connectivity featured by the link density matrix $q_{ij} = M_{ij}/(N_i N_j)$, $i, j = \overline{1, m}$, meaning the probability of having a connection with endpoints in i and j modules. M_{ij} represents the number of links connecting the two modules, as a rule, an $\mathcal{O}(N^2)$ dependence on network size. We denote with q_i the matrix's diagonal elements, and we will refer to it as intra-module link densities while off-diagonal elements will be called as inter-module (bridge) link densities. The number of bridge nodes in module i forming a bridge with module j is B_{ij}. Let us define the following events: i) X_{ij} - two vertices from communities i and j, respectively, $i, j = \overline{1, m}$, are assigned to the same Vronoi-cell; ii) $G_{n_1 n_2 \ldots n_m},\ \sum_{i=1}^{m} n_i = g$ - the g seed vertices are shared between m communities so that n_i seeds will get in module i. $\{G_{n_1 n_2 \ldots n_m}\}$ is a complete set of C_{g+m-1}^{m-1} events. It is equivalent to the amount of realizations for separating a linear chain of g equal balls by $m - 1$

arbitrarily positioned barriers. C_n^k indicates the binomial coefficient for the k-combinations of n events. Accordingly:

$$X_{ij} = \sum_{n_1 n_2 \ldots n_g} X_{ij} \cdot G_{n_1 n_2 \ldots n_g} \, . \tag{2.3}$$

Then the probability of one vertex from community i and one from community j, belonging to the same Voronoi-cell or shortly Voronoi-cohesion is:

$$c_{ij} \equiv P(X_{ij}) = \sum_{n_1 n_2 \ldots n_g} P(X_{ij}|G_{n_1 n_2 \ldots n_g}) P(G_{n_1 n_2 \ldots n_g}) \, ,$$

which can be given in the form of:

$$\mathbf{c} = \mathbf{V} \cdot \mathbf{g} \, , \tag{2.4}$$

where $\mathbf{c}, \mathbf{V}$ and $\mathbf{g}$ are matrices of size $N(N+1)/2$-by-1, $N(N+1)/2$-by-C_{g+m-1}^{m-1} and C_{g+m-1}^{m-1}-by-1, respectively. For the case study of extreme modularity please take a look at the thesis.

Let us examine a case of two equivalent modules, both with size $N \gg 1$ and connection density q, and bridge density b. Next we have summarized some meaningful quantities of a vertex pair as a function of their location (in module 1. or 2.). The first table collects the probability distribution of their common distances. Presuming that they are both generator vertices the second table contains the relative sizes of the matching Voronoi-cells in all of the modules:

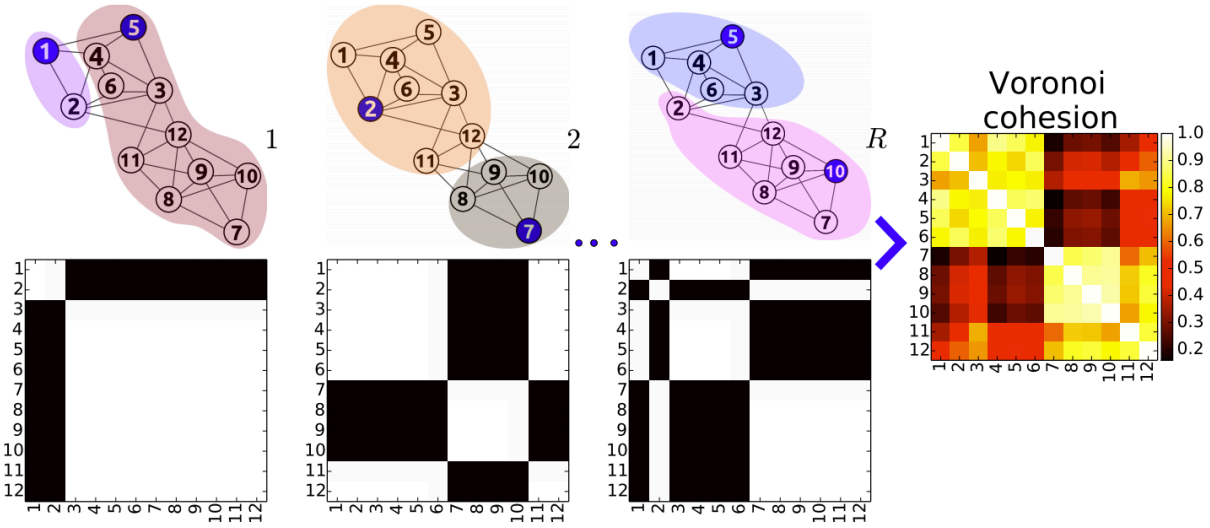

Figure 2.3: The basic idea for stochastic graph Voronoi diagrams. Figures show how the Voronoi cohesion map is calculated using R number of tessellations. Every sub-figure shows a realization of the binary colocation matrix generated with 2 seed nodes randomly chosen each time for example (1, 5), (2, 7), ..., (5, 10). The values of the colocation matrix are $1 = white$ for nodes in same cell, $0 = black$ if nodes are in different cells. The average over all R matrices reveals the cohesion map in other words the probabilities of nodes being in the same community.

loc. in mod.	$d=1$	$d=2$
1,1	q	$1-q$
1,2	b	$1-b$
2,2	q	$1-q$

module 1		module 2		loc. in mod.
gen.1	gen.2	gen.1	gen.2	
1/2	1/2	1/2	1/2	1,1
s	$1-s$	$1-s$	s	1,2
1/2	1/2	1/2	1/2	2,2

where s is the relative size of Voronoi-cell 1 in module 1 when the generator vertex 2 is located in module 2. The value of s can be calculated by summing up the following contributions (figure 2.4(a)):

1. bq fraction of module 1 has direct connections to both of the generator nodes and is evenly shared among the two Voronoi-cells;
2. $q(1-b)$ fraction has direct connections only to generator 1, as a result it contributes fully to cell 1;
3. $(1-q)b$ has direct connection only to seed 2 therefore is not contributing to cell 1;
4. $(1-b)(1-q)$ is not connected directly to both generators. It will be divided between the two generators proportionally to the number of shortest paths. The "modus operandi" node to generator 1 can be from module 1 with a probability of q^2 and b^2 from module 2. Similar values for the distant generator 2 are both bq, therefore, a fraction of $(q^2+b^2)/(q+b)^2$ of this domain corresponds to cell 1.

Adding up the contributions results:

$$s(q,b) = \frac{qb}{2} + \frac{1-b}{(q+b)^2}\left[q^2+b^2+2q^2b\right] \ . \tag{2.5}$$

Constructing matrix **c** from equation (2.4):

$$\mathbf{V} = \begin{pmatrix} 1/2 & 1-2s(1-s) & 1/2 \\ 1/2 & 2s(1-s) & 1/2 \\ 1/2 & 1-2s(1-s) & 1/2 \end{pmatrix} \ , \quad \mathbf{g} = \begin{pmatrix} 1/4 \\ 1/2 \\ 1/4 \end{pmatrix} \ .$$

Therefore

$$c_{11} = c_{22} = \frac{3}{4} - s(1-s) \ , \tag{2.6}$$

$$c_{12} = \frac{1}{4} + s(1-s) \ . \tag{2.7}$$

whence the connection density dependent contrast is

$$\gamma = \frac{1}{2}\left[1-4s(1-s)\right] \ . \tag{2.8}$$

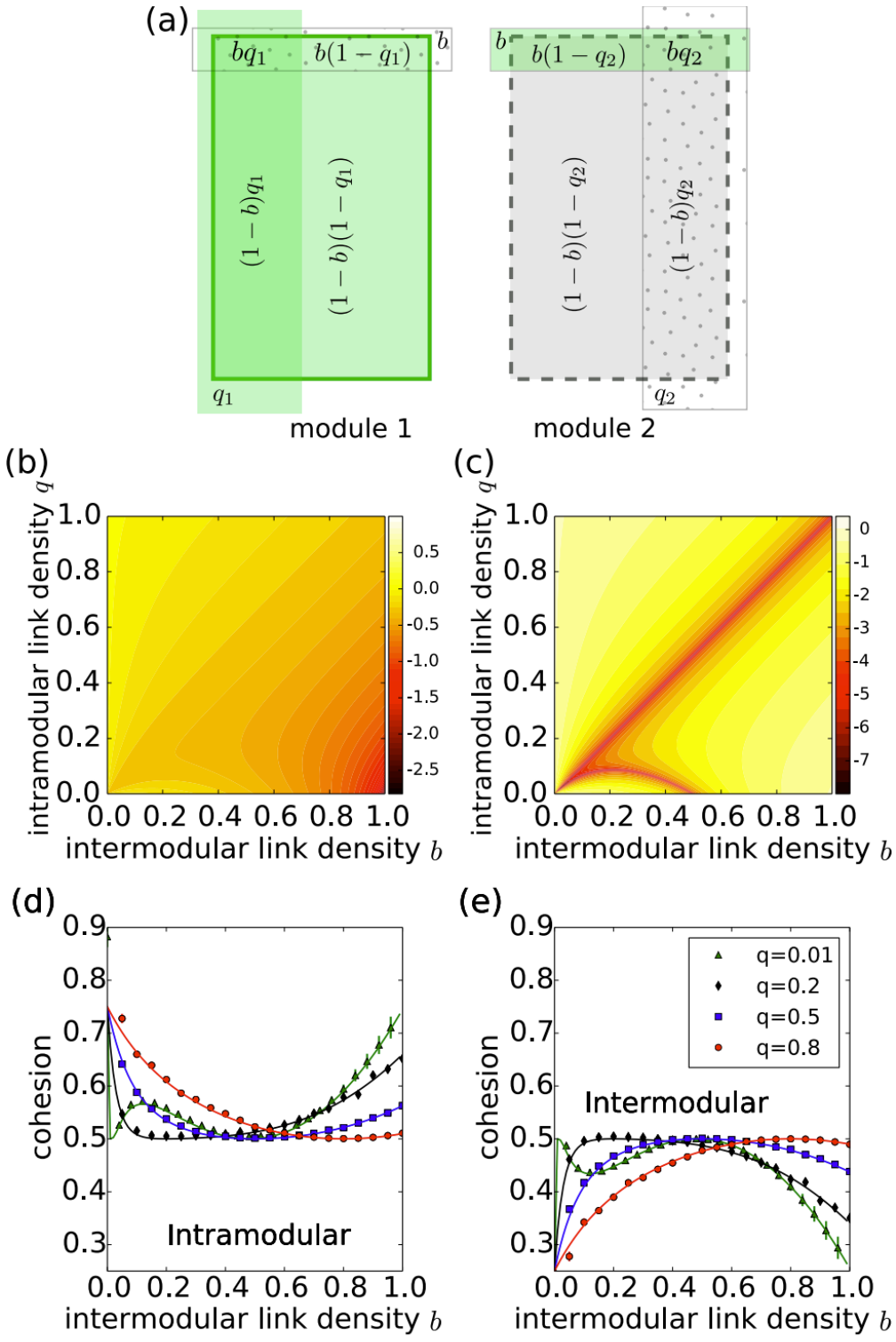

Figure 2.4: Non-infinitesimal bimodular connection density. Illustration of how two Voronoi regions share two equal in size ER-type modules, with non-infinitesimal inter-modular link density, b. (b) Shows how the intramodular link densities, $q_1 = q_2 = q$, and intermodular link density, b influence the relative size of cell 1 in module 1 (described by equation (2.5)). (c) The corresponding contrast calculated from equation (2.8). (d, e) Presents the intra- and intermodular Voronoi cohesions depending on the intermodular link density, b. Analytical results for several intramudular connection densities from equations (2.6) and (2.7) are shown in comparison with simulations obtained from large ensemble of 5000 tessellations, meaning 10 topologies and 500 generator node sets on a network of $N = 2 \times 800$ verices.

If $b = q$, then the intra- and intermodular connection densities equate, $\gamma = 0$, this means the network is no longer modular. The the maximum contrast is $\gamma = 1/2$, when the bridge size is negligible. The striking feature of this technique is that the theory can be confirmed even by a comparatively low network and ensemble size (figure 2.4).

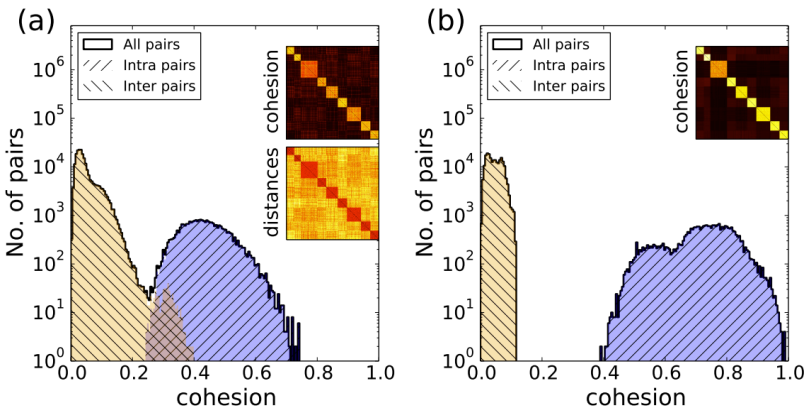

Figure 2.5: Benchmark network cohesion matrix. (a) Cohesion histogram for every vertex pair from the cohesion map of a benchmark network consisting of $N = 500$ nodes, $M = 5000$ connections, and $m = 9$ modules [28, 29]. For the cohesion calculation $g = 15$ generators, and $R = 3000$ repeats were implicated. Intra- and intermodular node pairs are colored based on the ground-truth. The insets are cohesion and the inter-node distance matrices, ordered based on truth information and here the larger values are closer to the main diagonal line. (b) Cohesion histogram calculated for the exact same network but with the contrast boosting method(for more detail please check section 2.2.6 in the thesis).

After attaining effective convergence, the network's community structure can be extracted from the cohesion matrix. The easiest method is to put a threshold in the cohesion histogram (see figure 2.5 (a)). In figure 2.5(a) we plot the distribution of the Voronoi cohesion matrix elements for all node pairs. For further optimization, after every 200 repeats we relocate a small percentage of low correlated nodes' connections to highly correlated unconnected node pairs. This way we increase the intra-inter gap while keeping the community structure unharmed 2.5(b).

Communities can then be found the following way: *1.* initially every node gets a separate community label; *2.* in a loop over all vertices the label of a node is changed to another node's label, which has an unchanged label and has a similar cohesion value within a predefined threshold. While the gap in the histogram appears to be a clear modular separation, these methods rarely work optimally for community detection in real-world networks. For detailed analysis on real-world networks please take a look at the thesis.

3

Scaling in real-world networks

3.1 Further we travel the faster we go: a general rule for human mobility networks

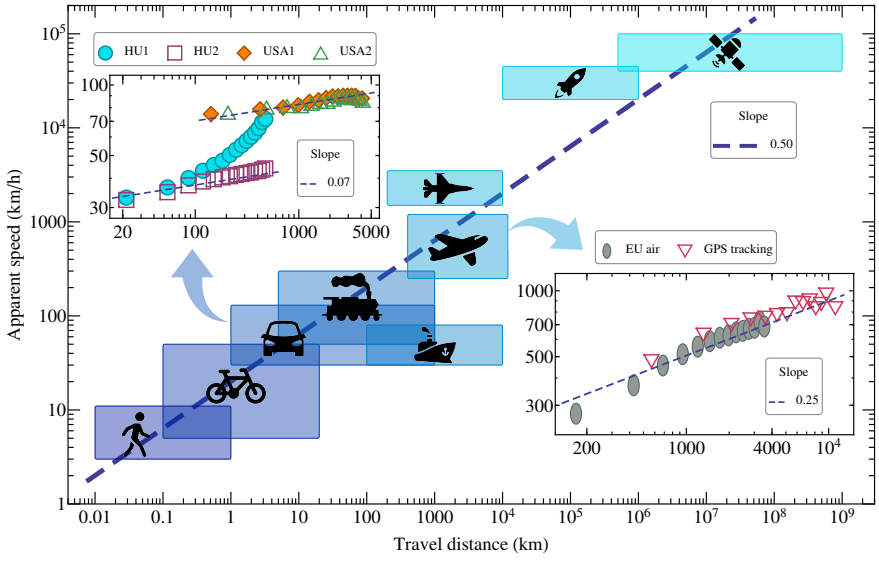

Figure 3.1: The apparent travel speed for all travel modes (estimated from averaging travel time on the geodesic line) as a function of the travelling distance. Boxes show speed and distance intervals on which the specified mode is used. Dashed line indicates a power-law trend. The two inset figures present some averaged results broken down on the two most popular travelling modes: car and air travel.

Firstly, for all human travel modes, extending from walking to cosmic transport, we determine

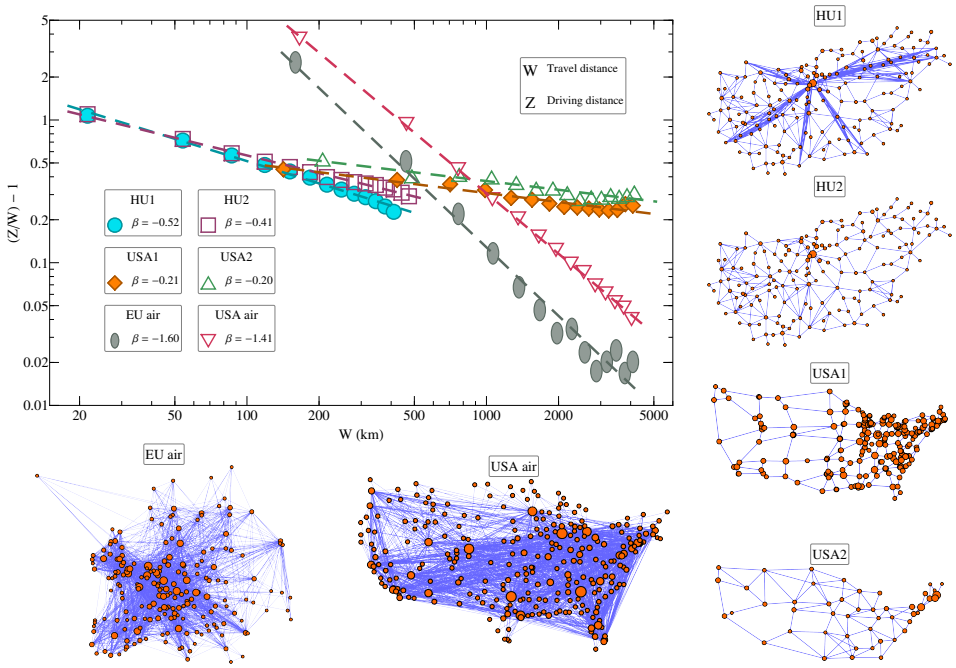

Figure 3.2: Topology of travel networks. Scaling of the cruising distance (z) and geodesic distance (w) on distinct travel networks. The lines indicate a valid $\frac{z}{w} = 1 + C \cdot w^{-\beta}$ relationship. β values are shown in the legend. We also show here the network topology of some human mobilities used in this study.

approximately the distance and velocity magnitudes (boxes on 3.1). The first impressive outcome we can observe is that the boxes follow a power-law trend with an exponent of roughly 0.5 as the distance increases. We find from these outcomes that when we consider roadways with similar ranks, with roughly the same speed limits (HU2, USA1, USA2), a power-law trend is attained with not so different scaling exponents. Regarding air transit, only direct flight information between airports have been used and the results show similar scaling with exponent of about 0.25.

The topology of the networks is the first obvious cause, on which the transportation is taking place. Vertices are not necessarily connected by straight roads and there are typically no direct routes between them [30]. To travel from one vertex to another, the commuters follow a path on the graph with the shortest road-length, noted here as commuting distance and labelled with z. As the w geodesic distance of the transit increases, the z commuting distance comes closer in approximation to w. Evaluating the topology for certain transport networks in this study, we notice a rather general scaling relation (see figure 3.2):

$$\frac{z}{w} = 1 + C \cdot w^{-\beta} \tag{3.1}$$

The proper β exponents visibly separate the trends for road- and air transit networks. For road travel $\beta \approx 0.2 - 0.3$ was fitted, and for the air travel $\beta \approx 1.4 - 1.5$ was found. Meaning that the z

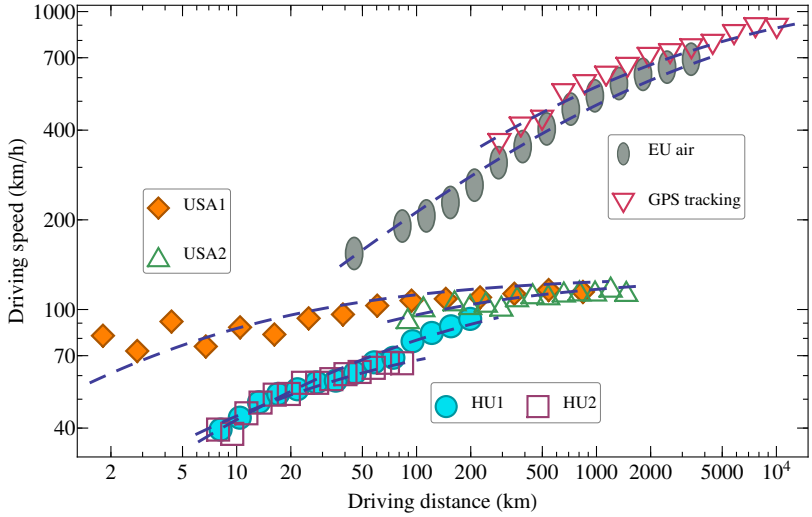

Figure 3.3: The travelling speed. The average apparent speed, u, depending on the cruising distance, z. Results are plotted for road travel and air transit together, both having a successful fit using the formula in equation (3.2) with $\alpha = 0.5$.

value for the case of road travel networks converges slower to w than for the case of air transit.

In the study of air commuting the increase of apparent speed as a function of the geodesic distance was still present, despite the fact that only direct (presumably in straight line) flights were taken into account between airports, for which supposedly: $z = w$. This is the same result when we take into consideration just the direct links between vertices in the road network. In this case the average driving or cruising speed on a direct connection is defined as $u = z/t$ and also increases with the commuting distance, z. Figure 3.3 shows the results in this context.

It can be argued that the source and the target vertices cause delays on each connection. Furthermore, there are also delays outside the vertices for each linear segment. This is because the traffic flow is different from the ideal, but the increase is not inherently linear. When we presume there is a limited travelling speed, u_0, in a particular segment and the delays increase with a power-law as function of the segment's length, z: $t_{delay} = K \cdot z^{\alpha}$, we get:

$$u = \frac{z}{\left(\frac{z}{u_0}\right) + t_{delay}} = \frac{1}{\left(\frac{1}{u_0}\right) + K \cdot z^{\alpha - 1}} \tag{3.2}$$

We have an increasing trend when $\alpha < 1$. On figure 3.3 we show that the equation (3.2) fits quite successfully for every experimental data from road to air travel. For national road data (HU2) the seed limit was $u_0 = 90$ km/h, while for highway and interstate data (HU1, USA1, USA2) we used $u_0 = 130$ km/h. For air travel we used $u_0 = 1200$ km/h ≈ 1 Mach. A fixed $\alpha = 0.5$ exponent gives a roughly suitable fit for all the data, hinting to a universal convergence with the u_0 speed limit.

Finally, the limiting speed value also increases with the length of the road- or flight-segment. As the road-segment is longer, the speed limit is usually increased. Highways have longer segments and increased speed limit compared to national roads. On larger air travel segments, usually faster airplanes are cruising, and similar effect is true for rail-travel. All of these effects explain well the non-trivial scaling observed in human mobility.

3.2 Scaling on the Internet, experiment and model

In the previous section we revealed an interesting scaling between travelling time and travel distance (measured on the geodesic lines), valid on 10 orders of magnitudes in space for all human transportation modes [9]. Here we show that a similar scaling can be found for data transmission on the Internet [31]. In order to understand it, we propose a model, which can return not only this scaling but also the measured topological properties.

The experiments regarding the dynamics of data transfer on the Internet were based on echo request packet sending and receiving with Internet Control Message Protocol (ICMP) [32]. We used the very popular “ping” command [33] to test whether the source computer could reach a designated target computer. This most common time measurement unit is ms. A total of 24700 target computers were chosen from diverse locations on the globe. Their global positions (GPS coordinates) have been determined from an IP address table, using the web page of IP2LOCATION [34]. We determined the geodesic distance d between the origin and destination routers by using the GPS coordinates. The ”traceroute” uses the same basic principle as ”ping” does. The ICMP echo request gets the RTT in this situation too, but also receiving the intermediate hops’ addresses, revealing the router-level topology of the network. Both the large-scale ”ping” experiment and the freely accessible results of the CAIDA UCSD IPv4 Routed/24 Topology Dataset [35] attained with ”traceroute” are taken into account.

We examined the findings with a fit of RTT $= a \cdot d^{1/2}$, and the R^2 determination coefficients obtained were $R^2 = 0.98$ for the data on ping and $R^2 = 0.88$ in the case of the traceroute (see figure 3.4).

3.2.1 The Internet model and results

We consider a network model with a basic wiring rule that defines the connection between nodes in order to explain this non-trivial scaling law discovered in the previous part. In the model the cities are represented by the nodes of the graph and the connections between them correspond to the network cables or wiring channel. N nodes are spread evenly in the Euclidean space in the most basic approximation. The territory in question is a square with unit size edge.

We chose the population of cities, W_i, to determine its ”connectivity radius”, ω_i, as $\omega_i = \beta\sqrt{W_i}$.

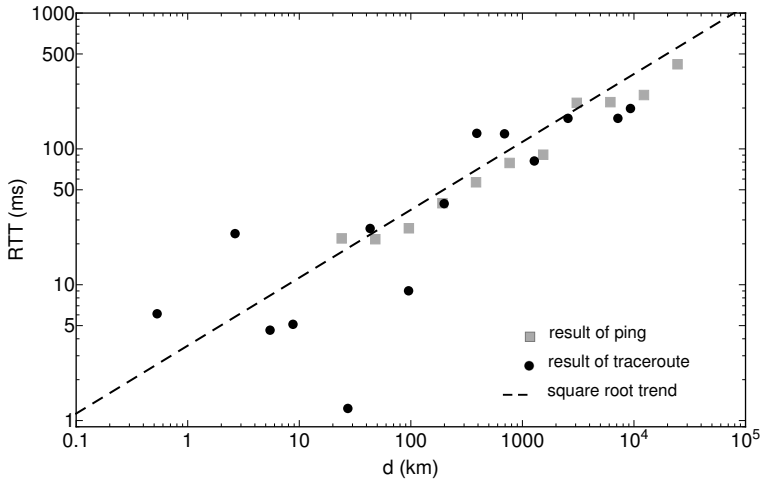

Figure 3.4: Ping and Traceroute experimetns. Round-trip time of both experiments as function of the distance, using the same logarithmic binning mentioned in [9]. The dashed line indicates a power-law trend with exponent $1/2$.

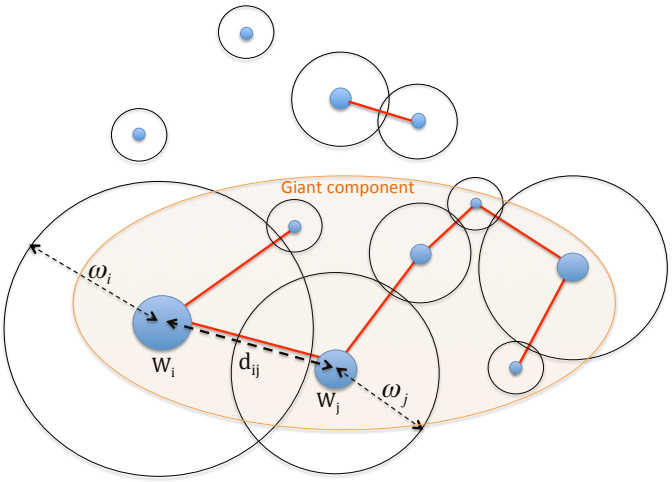

Figure 3.5: Connection rules. Main aspects and the linking rule of the model.

W_i are values allocated according to a Tsallis-Pareto type distribution of exponent $\alpha = 1$, proved to be adequate for large settlements [36, 37].

In the wiring procedure, we calculate the following ratio for every node pair:

$$f_{ij} = \frac{\omega_i + \omega_j}{d_{ij}} \tag{3.3}$$

Where d_{ij} is the Euclidean distance (or length of the geodesic line) between two cities. IF $f_{ij} > 1$, we link the two nodes, else they stay unconnected (see 3.5).

We were searching for the optimal β value, that was capable of reproducing the scaling laws and the average degree (links/node) of 8.68 found in the experimental observations. For every parameter group we averaged the simulation results over 100 independent versions of the network achieved with fixed parameters of N and β .

For $N = 2400$ we reached to a conclusion that the observed properties by the experiments are best replicated by the model for a proportionality factor of $\beta \approx 0.4$. Network constructions for $N = 8000$ and $\beta = 0.4$ result in graphs with statistically similar giant components.

The average degree of the giant component given by the traceroute is $\langle k \rangle = 8.68$. The degree-distribution is fitted with a Paretto-Tsallis (or Lomax II) distribution:

$$p(k) = \frac{\alpha}{(\alpha - 1)\langle k \rangle} \left(1 + \frac{k}{(\alpha - 1)\langle k \rangle}\right)^{-1-\alpha} \tag{3.4}$$

The degree distribution for the experimental results is shown with black dots in figure 3.6, and the Tsallis-Paretto fit (3.4) with $\alpha = 1.23$ is presented with red dashed line. We remind here that similar degree distributions including a scale-free tail is usually received in the core of exponentially diluted growth models with preferential attachment [37, 38]. The measured degree distribution and the result of the model in comparison (figure 3.6)shows agreement also with the topological properties found in [39].

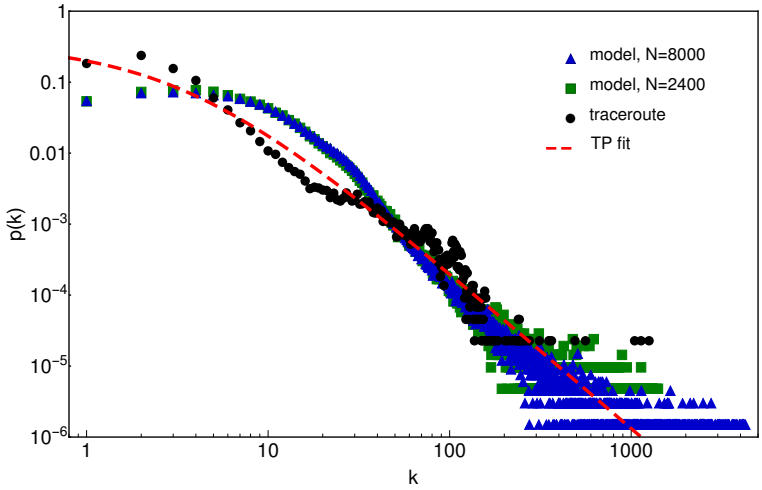

Figure 3.6: Degree distribution. The resulted network's degree distribution from the traceroute experiment marked with black points and the same of the network produced by the model with green squares for $N = 2400$ and blue triangles for $N = 8000$. The dashed red line indicates a Tsallis-Pareto distribution fit (3.4) with $\alpha = 1.23$ and $\langle k \rangle = 8.68$. The determination coefficient calculated for the experiment is $R^2 = 0.85$, and for the results of the model $R^2 > 0.9$ always.

The biggest contribution to the round-trip time is connected to the waiting periods suffered at the routers. Thus we could assume that the measured average time should increase with the amount of routers (hops) H met until the target is reached. Indeed, the experiments suggests a RTT$\propto H^{\gamma}$ relation (see figure 3.7) with $\gamma \approx 3/4$. By taking into consideration a $\gamma = 3/4$, the fit provides a coefficient determination of $R^2 = 0.98$. Consequently, we also point out that the amount of visited hops is increasing with the distance with a scaling exponent of: $(1/2)/(3/4) = 2/3$. The results shown in figure 3.8 validate this scaling.

For determining the topological shortest paths between vertices we used the Breadth-first

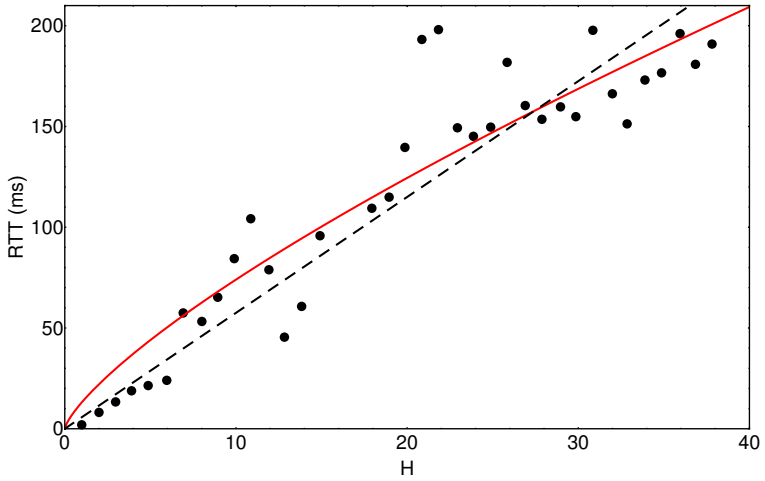

Figure 3.7: RTT vs Hops. Averaged RTT as a function of the hops count established from traceroute experiment. The dashed line shows a linear proportionality with a determination coefficient of $R^2 = 0.96$, while the continuous red line indicates a power-law fit with an exponent of $3/4$. The coefficient of determination for the latter fit is $R^2 = 0.98$.

search method implemented in the Python Igraph package [40], this is justified by the ICMP protocol, also visiting the minimum amount of hops. It is apparent that the trend is a power-law, both from experimental outcome and the one yielded by our model, consistent with the exponent $2/3$ and noticeably distinctive from $1/2$. This similarity between the experiments and model's results show that a simple model is capable of capturing the essence of such non-trivial scaling.

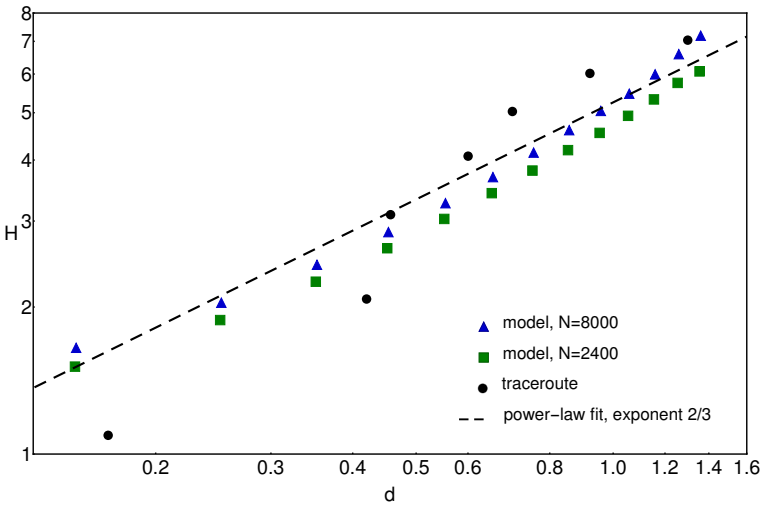

Figure 3.8: Hop vs distance. Number of hops as a function of the distance. The black points represent the data obtained from the traceroute experiments. Green squares ($N = 2400$) and blue triangles ($N = 8000$) represent the model's results. The dashed line suggests a power-law trend having an exponent of $2/3$. The determination coefficient of this fit ($y = a \cdot x^{2/3}$) for the traceroute data is $R^2 = 0.98$, while for the data points given by the model is over $R^2 = 0.99$. The geodesic distances for the traceroute experiment's network is rescaled in the $(0, 1) \times (0, 1)$ square.

4

Conclusions

In the present thesis we used the methods of statistical and computational physics for investigating two modern topics in the field of complex networks. We have reached to the following conclusions:

4.1 Community detection in complex networks

We proved that the graph version of Voronoi diagrams are suitable for detecting clusters on complex networks. In order to proceed in this manner we identified two major requirements: (1) definition of an appropriate distance metric between the nodes, (2) identifying the generator vertices which indicate the community centres. For the distance measure we chose the inverse of link (edge) clustering coefficient, $1/ECC$. Nodes were selected as generators when their relative local densities were the highest among the neighboring vertices within a radius r. With both of these measures, we have demonstrated that our technique can compete with all other methods. Only one of five popular algorithms, which by the way is programmed for optimizing the quality function used in our comparisons, managed to outperform our methods. We suggest that Voronoi tessellation with increasing radius r values is the best strategy when using our method.

We discussed theoretical and practical aspects of how stochastic graph Voronoi tesselations work. Contrary to other node-similarity definitions [3, 41] the Voronoi cohesion returns local node information taking into account the network's global structure. We have demonstrated that this kind of information overlaps with community relationships. In the analytical section of the thesis we demonstrated how a two-module graph can be partitioned and what characteristics the bridge nodes introduce to the cohesion matrix. These analytical findings were backed up by

simulations. In the thesis we presented a procedure that translates the information found in the cohesion matrix to community information for every node.

4.2 Scaling laws

We proved our initial hyphothesis according to which on every transport mode the travelling time has a non-trivial scaling as function of distance. The average apparent speed (geodesic distance/-travelling time) increases with a power-law trend as function of the distance. We have demonstrated that both the structure of the road network's topology and the factors which lead to deviation from ideal conditions (the delaying effects of the nodes, the increasing speed limits of longer travel-segments) contribute to this universal effect. For increasing the apparent speed, firstly, the road networks (or air connections) have to be optimized in such way that the topological β exponents are maximized. In this case the driving distance could reach rapidly the traveling distance, decreasing to a transit path length as short as possible. Planning the geometry of the road networks between large cities seems a bit problematic task, due to their given spatial distribution which determines the topology of the road network. However, in air travel collaborations among different airlines can cope with this situation.

The experimental study of the Internet leads to similar conclusions. Comparing the model with the experimental results brings to a conclusion that this simple one parameter wiring model embedded in geometric space is able to qualitatively reproduce the discovered statistical features of the Internet's network on router level. In this sense the resulted nontrivial scaling of the average round trip time of an echo request as function of the geodesic distance is due to the specific topology of the network. On the other hand, the difference between these scaling exponents, namely the trend of the hop number versus the distance ($\approx 2/3$), and the scaling exponent of the round-trip time dependence on the distance ($\approx 1/2$) indicates that on the routers a simple constant average delay could not be totally accounted for the scaling. Along with network topology, most likely certain factors have to be addressed in order to build a more realistic model. This is somewhat similar to the characteristics learned in the investigations of human mobility networks.

5

Publications

Publications relevant to the thesis

Scientific papers

- I. Papp, L. Varga, M. Afifi, I. Gere, and Z. Néda, Scaling in the space-time of the Internet. Scientific Reports, 9(1):9734, 2019. ISSN 2045-2322. (IF: 4.122)
- Z. I. Lázár, I. Papp, L. Varga, F. Járai-Szabó, D. Deritei, M. Ercsey-Ravasz, Stochastic graph Voronoi tessellation reveals community structure, Phys. Rev. E, 95, 2017 (IF: 2.353)
- Varga L, Kovács A, Tóth G, Papp I, Néda Z, Further We Travel the Faster We Go, PLoS ONE 11(2): e0148913, 2016 (IF: 2.776)
- D. Deritei, Z.I. Lázár, I. Papp, F. Jarai-Szabo, R. Sumi , L. Varga, E. Regan, M. Ercsey-Ravasz, Community detection by graph Voronoi diagrams, New Journal of Physics, 2014 (IF: 3.773)

Conference presentations

- I.Papp, M. Afifi, L. Varga, I. Gere and Z. Néda, Nontrivial dynamical scaling in the Internet: experiments and a simple model, talk, 12th Joint Conference on Mathematics and Computer Science Cluj-Napoca, June 14 – 17, 2018.

- I. Papp, M. Afifi, L. Varga, I. Gere, Z. Néda, Scaling in the space-time of the Internet, Poster 55, MECO43, 2018 May
- Varga L, Kovács A, Tóth G, Papp I, Néda Z, Velocity versus distance scaling in human travel, poster, XXXVI Dynamics Days Europe, Corfu, Greece, (2016)
- Z. I. Lázár, I. Papp, L. Varga, F. Járai-Szabó, D. Deritei, M. Ercsey-Ravasz, Community detection using stochastic graph Voronoi tesselation, poster, NetSci, (2015)
- D. Deritei, Z. I. Lázár, I. Papp, F. Jarai-Szabo, R. Sumi, M. Ercsey-Ravasz; Community detection by graph Voronoi diagrams; NetSci 2013, International School and Conference on Network Science, June 3-7 2013, Copenhagen, Denmark

Other publications

Scientific papers

- L.P Csernai, M. Csete, I.N. Mishustin, A. Motornenko, I. Papp, L.M. Starov, H. Stöcker, N. Kroó, Radiation dominated implosion with flat target, Physics of Wave Phenomena, February 2020, accepted for publication (IF:0.641)
- L.P Csernai, N. Kroó and I. Papp, Radiation dominated implosion with nano-plasmonics, Laser and Particle Beams, Volume 36, Issue 2, June 2018 , pp. 171-178 (IF: 1.194)
- K.Z. Rónai, A.Szentkirályi A. S. Lázár Z. I. Lázár I. Papp, F. Gombos, R. Zoller, M. E. Czira, A. V. Lindner, I. Mucsi, R. Bodizs, M. Z. Molnar, M. Novák, Association of symptoms of insomnia and sleep parameters among kidney transplant recipients, Journal of Psychosomatic Research, 95-104, 2017 (IF: 2.722)
- L.P. Csernai, I. Papp, S.F. Spinnangr and Y. Xie, Physical Basis of Suatainable Development, Int. J. of Central European Green Innovation 42, 2016, 39-50.

Conference presentations

- I. Papp, L.P. Csernai, S.F. Spinnangr and Y. Xie, Some number crunching on Earth's energy forecast, talk, September 18, 2019, IWoC, Kőszeg
- L.P. Csernai, N. Kroó and I. Papp, Applications of Advances in Relativistic Fluid Dynamics to Laser Fusion, 6th International Conference on New Frontiers in Physics (ICNFP 2017)

- D. Deritei, L. Varga, I. Papp, F. Járai-Szabó, Z. I. Lázár, R. V. Florian and M. Ercsey-Ravasz, Genetic-like algorithm applied on citation networks for evaluating scientific publications, poster, Conference on Complex Systems, Amsterdam, Netherlands, (2016)
- I. Papp and L.P. Csernai, Inertial Confinement Fusion: Radiation Dominated Implosion, talk, Exploring the Shores of Fundamental Matter: Advances around the Northern Seas (NorSAC-2015), Bergen, Jul. 31, 2015
- L. Varga, F. Jarai-Szabo, D. Deritei, Z. I. Lázár, I. Papp, R. Florian, M. Ercsey-Ravasz , Local Cluster Detection Method for Normalizing Scientometric Indicators, poster, NetSci, (2015)
- I. Papp, M. Ercsey-Ravasz, D. Deritei, R. Sumi, F. Jarai-Szabo, R. V. Florian, A. I. Cabuz, Z. I. Lázár; The PH-Index: Hirsch Index of Individual Publications, poster; ISSI 2013, 14th International Society of Scientometrics and Infometrics Conference, 15-19 July 2013, Vienna, Austria

Patent

- L. P. Csernai, N. Kroo and I. Papp, Procedure to improve the stability and efficiency of laser-fusion by nano-plasmonics method, Patent No. P1700278/3 of the Hungarian Intellectual Property Office.

Selected references

[1] Newman, MEJ. Networks: An Introduction. Oxford University Press, 2010.

[2] Barabasi, AL. Network science. 2012.

[3] Santo Fortunato. Community detection in graphs. Physics Reports, 486(3):75 – 174, 2010. ISSN 0370-1573.

[4] Han, J. Kamber, M. Pei, J. Data mining: Concepts and techniques. 2011.

[5] Theodoridis, S. Koutroumbas, K. Pattern recognition. 2008.

[6] Aurenhammer, F. Voronoi diagrams - a survey of a fundamental geometric data structure. ACM Computing Surveys, 23(3):345–405, 1991.

[7] Kung, K.S., Greco, K., Sobolevsky, S., Ratti, C. Exploring universal patterns in human home-work commuting from mobile phone data. PLoS ONE, 9(6), 2014.

[8] Gallotti, R. & Barthélemy, M. Anatomy and efficiency of urban multimodal mobility. Sci. Rep., 4:6911, 2014.

[9] Levente Varga, András Kovács, Géza Tóth, István Papp, and Zoltán Néda. Further we travel the faster we go. PLOS ONE, 11(2):1–9, 02 2016.

[10] Barabasi AL and Albert R. Emergence of scaling in random networks. Science, 286:509–512, 1999.

[11] SH. Yook, H. Jeong, and AL. Barabasi. Modeling the Internet's large-scale topology. PNAS, 99:13382–13386, 2002.

[12] F. Castellano, C. Cecconi, F. Loreto, V. Parisi, D. Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA, 101:2658, 2004.

[13] Newman, MEJ. Modularity and community structure in networks. PNAS, 103(23):8577–8582, 2006.

[14] Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33:452–473, 1977.

[15] White, JG. Southgate, E. Thompson, JN. and Brenner, S. . The structure of the nervous system of the nematode caenorhabditis elegans. Phil. Trans. R. Soc. London, 314, 1986.

[16] Watts, DJ and Strogatz, SH. Collective dynamics of 'small-world' networks. Nature, 393: 440–442, 1998.

[17] Jeong, H. Mason, S. Barabási, AL and Oltvai, ZN. Lethality and centrality in protein networks. Nature, 411:41–42, 2001.

[18] Newman, MEJ. The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA, 98:404–409, 2001.

[19] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD '05, pages 36–43, New York, NY, USA, 2005. ACM. ISBN 1-59593-215-1.

[20] Blondel, VD. Guillaume, JL. Lambiotte, R. Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), 2008.

[21] Raghavan, UN. Albert, R. Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E, 76:036106, 2007.

[22] Jierui Xie and Boleslaw K. Szymanski. Towards linear time overlapping community detection in social networks. In Proceedings of the 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part II, PAKDD'12, pages 25–36, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-30219-0.

[23] J. Xie, B. K. Szymanski, and X. Liu. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In 2011 IEEE 11th International Conference on Data Mining Workshops, pages 344–349, Dec 2011.

[24] Ahn, YY. Bagrow, JP and Lehmann, S. Link communities reveal multiscale complexity in networks. Nature, 466:761–764, 2010.

[25] Rosvall, M and Bergstrom, CT. Maps of random walks on complex networks reveal community structure. PNAS, 105:1118–1123, 2008.

[26] Rosvall, M. Axelsson, D. and Bergstrom, CT. The map equation. Eur. Phys. J. Special Topics, 178:13–23, 2009.

[27] Zsolt I. Lázár, István Papp, Levente Varga, Ferenc Járai-Szabó, Dávid Deritei, and Mária Ercsey-Ravasz. Stochastic graph Voronoi tessellation reveals community structure. Phys. Rev. E, 95:022306, Feb 2017.

[28] Lancichinetti, A. Fortunato, S. Radicchi F. Benchmarks graphs for testing community detection algorithms. Phys. Rev. E, 78, 2008.

[29] Lancichinetti, A. Fortunato, S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E, 80, 2009.

[30] Marc Barthelemy. Spatial networks. Physics Reports, 499(1):1 – 101, 2011. ISSN 0370-1573.

[31] István Papp, Levente Varga, Mounir Afifi, István Gere, and Zoltán Néda. Scaling in the space-time of the Internet. Scientific Reports, 9(1):9734, 2019. ISSN 2045-2322.

[32] J. Postel. Internet Control Message Protocol. RFC, 777:1–14, 1981.

[33] Linux man page for ping command. https://linux.die.net/man/8/ping.

[34] The ip adress databases. http://www.ip2location.com.

[35] The CAIDA UCSD IPv4 Routed /24 Topology Dataset - 2017. http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

[36] M.E. Moses E.H. Decker, A.J. Kerkhoff. Global patterns of city size distributions and their fundamental drivers. Plos One, 22:e934, 2007.

[37] T-S. Biro and Z. Neda. Unidirectional random growth with resetting. Physica A, 499:335–361, 2018.

[38] Barabasi AL and Albert R. Emergence of scaling in random networks. Science, 286:509–512, 1999.

[39] Mahadevan, P. and Krioukov, D. and Fomenkov, M. and Huffaker, B. and Dimitropoulos, X. and Claffy, K. and Vahdat, A. The Internet as-level topology: Three data sources and one definitive metric. ACM SIGCOMM Computer Communication Review (CCR), 36(1):17–26, 2006.

[40] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. InterJournal, Complex Systems, 1695(5):1–9, 2006.

[41] M. E. J. Newman. Detecting community structure in networks. The European Physical Journal B, 38(2):321–330, Mar 2004.