Information and Dynamics in Urban Traffic Networks

Giovanni Petri

May 18, 2012

Submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy of Imperial College London and the Diploma of Imperial College London
Declaration

I herewith certify that all material in this dissertation which is not my own work has been properly acknowledged.

Giovanni Petri
Abstract

The study of complex systems has intensified in recent years. Researchers from many different disciplines have realised that the study of systems possessing a large number of degrees of freedom interacting in a non-linear way can offer insights into problems in engineering, biology, economics and many other fields besides. Among the themes in complexity, we focus here the issues of congestion and congestion emergence in the context of urban networks, with particular reference to the effects of dissemination of information about the system’s status. This topic is of great relevance today, due to the increasing availability of real-time information about traffic conditions and the large diffusion of personal devices that allow travellers to access such information.

Through the analysis of a few simple models of information propagation in urban environment, we uncover that, contrarily to the naïve expectation, complete information is often detrimental to the global performance of the urban traffic network. Indeed, global or long-range dissemination induces correlations in the systems that become a source for spatial disorder, making the system more prone to the emergence of congested states and pushing it away from its Wardrop equilibrium. The models we study range from simple flow models on network to complete agent-based simulations on real-world networks with interacting agents and dynamical information.

We then analyse real data, coming from London’s network of traffic detectors. We confirm that the heterogeneity in the distribution of traffic flow and occupancies across the network reduces its performances, consistently with the results obtained for the information propagation models. In addition, we find a rich phenomenology strikingly similar to the one found in critical self-organised systems. Indeed, we measure power-law correlation
functions and 1/f power spectra, hinting to long spatial and temporal effects in the traffic flow, and confirm this result through the community detection analysis of the detectors’ correlation network, which showing that the whole urban area behaves as a single large chunk. We conclude discussing the origin of these features and how they can be used to improve the network performances.
To my family

perché ogni scarafione è bello a mamma voja
Contents

Abstract IV
Table of Contents VIII
List of Publications IX
List of Tables X
List of Figures XVIII
List of Notations XX
Acknowledgements XXII

1. Introduction 1
   1.1. Outline .............................................. 5
   1.2. Main contributions ................................. 6

2. Relevant Background and Methods 8
   2.1. Networks ............................................. 9
       2.1.1. Elements of Complex Network Theory .......... 9
       2.1.2. Community Detection Methods ................. 14
   2.2. Self Organised Criticality ........................ 22
       2.2.1. SOC and Cascades .............................. 25
   2.3. Fluctuation Scaling ................................ 26
       2.3.1. Meaning of the scaling exponents .......... 29

   3.1. Background and Aim ............................... 31
   3.2. The Freeflow Project and its Dataset ........... 34
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.</td>
<td>London’s Macroscopic Fundamental Diagram</td>
<td>40</td>
</tr>
<tr>
<td>3.3.1.</td>
<td>Current Understanding</td>
<td>40</td>
</tr>
<tr>
<td>3.3.2.</td>
<td>The Role of Spatial Heterogeneity</td>
<td>44</td>
</tr>
<tr>
<td>3.3.3.</td>
<td>Flow heterogeneity and Hysteresis</td>
<td>47</td>
</tr>
<tr>
<td>3.4.</td>
<td>Correlation Networks and Congestion</td>
<td>55</td>
</tr>
<tr>
<td>3.4.1.</td>
<td>Results</td>
<td>58</td>
</tr>
<tr>
<td>3.5.</td>
<td>Power-law correlation length and power spectra</td>
<td>65</td>
</tr>
<tr>
<td>3.5.1.</td>
<td>Origin of Criticality</td>
<td>72</td>
</tr>
<tr>
<td>3.6.</td>
<td>Summary of Chapter</td>
<td>75</td>
</tr>
<tr>
<td>4.</td>
<td>Models of Information Propagation</td>
<td>78</td>
</tr>
<tr>
<td>4.1.</td>
<td>Background and Aim</td>
<td>78</td>
</tr>
<tr>
<td>4.2.</td>
<td>Minimal Flow Model</td>
<td>84</td>
</tr>
<tr>
<td>4.2.1.</td>
<td>Description</td>
<td>84</td>
</tr>
<tr>
<td>4.2.2.</td>
<td>Model setup</td>
<td>86</td>
</tr>
<tr>
<td>4.2.3.</td>
<td>Results</td>
<td>88</td>
</tr>
<tr>
<td>4.2.4.</td>
<td>Rate Model</td>
<td>91</td>
</tr>
<tr>
<td>4.2.5.</td>
<td>Conclusions</td>
<td>96</td>
</tr>
<tr>
<td>4.3.</td>
<td>Information propagation with self-organized propagation</td>
<td>97</td>
</tr>
<tr>
<td>4.3.1.</td>
<td>Model setup</td>
<td>99</td>
</tr>
<tr>
<td>4.3.2.</td>
<td>Discussion</td>
<td>103</td>
</tr>
<tr>
<td>4.3.3.</td>
<td>Conclusions</td>
<td>107</td>
</tr>
<tr>
<td>4.4.</td>
<td>Self-organised propagation on realistic networks</td>
<td>108</td>
</tr>
<tr>
<td>4.4.1.</td>
<td>Model Setup</td>
<td>108</td>
</tr>
<tr>
<td>4.4.2.</td>
<td>Simulation Results and Discussion</td>
<td>112</td>
</tr>
<tr>
<td>4.4.3.</td>
<td>Conclusions</td>
<td>122</td>
</tr>
<tr>
<td>4.5.</td>
<td>Model of Aging and Propagating Information</td>
<td>125</td>
</tr>
<tr>
<td>4.5.1.</td>
<td>Description of the model and setup</td>
<td>126</td>
</tr>
<tr>
<td>4.5.2.</td>
<td>Results</td>
<td>131</td>
</tr>
<tr>
<td>4.5.3.</td>
<td>Discussion</td>
<td>136</td>
</tr>
<tr>
<td>4.5.4.</td>
<td>Conclusions</td>
<td>138</td>
</tr>
<tr>
<td>4.6.</td>
<td>Summary of Chapter</td>
<td>140</td>
</tr>
<tr>
<td>5.</td>
<td>Discussion and Conclusions</td>
<td>143</td>
</tr>
<tr>
<td>5.1.</td>
<td>Macroscopic Fundamental Diagram and Urban Dynamics</td>
<td>143</td>
</tr>
<tr>
<td>5.1.1.</td>
<td>Findings</td>
<td>143</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>5.1.2. Conclusions</td>
<td>144</td>
<td></td>
</tr>
<tr>
<td>5.1.3. Implications and Future Work</td>
<td>145</td>
<td></td>
</tr>
<tr>
<td>5.2. Information Propagation and its Effects</td>
<td>146</td>
<td></td>
</tr>
<tr>
<td>5.2.1. Findings</td>
<td>146</td>
<td></td>
</tr>
<tr>
<td>5.2.2. Conclusions</td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>5.2.3. Implications and Future Work</td>
<td>148</td>
<td></td>
</tr>
<tr>
<td><strong>Bibliography</strong></td>
<td>175</td>
<td></td>
</tr>
<tr>
<td>A. Notes on self-organized propagation model</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>A.0.4. Topologies</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>A.0.5. OD matrices</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>A.0.6. Compliance</td>
<td>177</td>
<td></td>
</tr>
<tr>
<td>B. Freeflow data</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>B.1. Correlation matrices</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>B.1.1. Modularity trick</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>B.1.2. Geodesic and Euclidean Distances</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td>B.2. Coefficient of MFD fits</td>
<td>183</td>
<td></td>
</tr>
</tbody>
</table>
List of Publications


List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1.</td>
<td>Coefficient of the quadratic polynomial fit of $F$ versus $\Sigma_{O}$</td>
<td>183</td>
</tr>
<tr>
<td>B.2.</td>
<td>Coefficients of the quadratic polynomial in $O$ and $\Sigma_{O}$ fitting the Freeflow data</td>
<td>183</td>
</tr>
</tbody>
</table>
List of Figures

3.1. The map of the part of Central London covered by the Freeflow sensors. There are in excess of 3000 ILDs, stretching from Wimbledon and Brixton in the south to Tottenham Hale and Walthamstow in the north, from Stratford in the east all the way to Kensington and Chiswick in the west, for a total catchment area of about 400 km$^2$.

3.2. Map of the time-averaged occupancy $\langle o_i \rangle$ for sensors (top). The sensors’ size and colour both represent the average value of the occupancy for the corresponding sensor. There is a large spatial disorder in the distribution of traffic with the most central sensors experiencing the largest densities. The values (bottom) are distributed around a well defined peak, which can be fit well by a normal distribution with average 8.6 and width 12.9, which accounts for 95% of the total sensors. The remaining 5% lies in the peak (close to occupancy 92%) for very high densities, corresponding to the red, large dots in the picture.

3.3. The distribution of occupancies for the Freeflow sensors separated for periods of the day. The single sensor occupancy distributions show an initial plateau followed by a power-law decay.

3.4. The distribution of single sensor occupancies for the Freeflow sensors clustered according to the total network instantaneous occupancy $O(t)$.

3.5. An example of the flow and occupancy profile (sensor no. n00/005g1). The errors are the standard deviation of the flow/occupancy values at given time.
3.6. The Macroscopic Fundamental Diagram for London ($F(t)$ vs $O(t)$) where data points are coloured according to the value of $\Sigma O(t)$. ......................................... 45

3.7. The relationship between $F$ and $\Sigma O$ for the London data. The best fit is obtained with a quadratic polynomial (adj-$R^2 = 0.9862$, RMSE=12.85). .................................. 46

3.8. The Freeflow MFD and the surface obtained through the polynomial fit ................................................................. 48

3.9. The continuation of the surface obtained for wider intervals of $O$ and $\Sigma O$. The MFD surface unrealistically grows for small occupancies. Negative flows have been set to 0 in this visualization ................................................................. 48

3.10. The Freeflow MFD and the surface obtained through the polynomial fit with the addition of the origin as an additional data point ................................................................. 49

3.11. The continuation of the surface obtained from the Freeflow data with the addition of the origin. At variance with Fig 3.9, the new surface is concave and produces a reasonable behaviour for the MFDs. Negative flows have been set to 0 in this visualization ................................................................. 49

3.12. The Macroscopic Fundamental Diagram again where colours correspond to $\sigma_F(t)/F(t)$. ............................................. 50

3.13. Main plot, the flow spatial heterogeneity $\Sigma F(t)$ versus the network occupancy $O(t)$. The line represents the temporal succession of data points. A clear clockwise hysteresis loop is present: during the first part of the day the system evolves along the upper branch (larger disorder), until it reaches the minimum $\Sigma F$. Successively, it unloads along the lower branch (smaller disorder). Inset, $\Sigma F(t)$ versus $\Sigma O(t)$ displaying an anticlockwise hysteresis loop. ............................................. 51
3.14. Top, anticlockwise hysteresis in the Macroscopic Fundamental Diagram for the first week of recording. Bottom, the typical hysteresis loop: each points has been obtained averaging over the data points corresponding to a certain time of the day (the resolution is 15 minutes, so daily there are 96 measurements). The loop is anticlockwise: during the morning rush period, the network follows the lower branch in the plot, while the unloading in the afternoon happens along the upper branch.

3.15. Top, visualization of the distance matrix between pairs of sensors. Bottom, the correlation matrix $C_f$. Already at a visual inspection it is possible to see, from the groups of sensors at small distances, that the network is spatially clustered as one would expect in a street network. On the contrary, the flow correlation matrix does not show any evident community structure.

3.16. Results for the Girvan-Newman community detection on the correlation network $C_f$. The Girvan-Newman modularity is optimised by a partition of the network containing 4 communities. The two largest ones (red and black) account for about 85% of the sensors and correspond respectively to the northern central area and the southern peripheral areas of London. The third (yellow) appears to be spatially scattered across the central area of London, but has no evident functional structure. The fourth is almost invisible in the network. All communities show a large degree of spatial overlap and it is not easy to identify units allowing to decompose the traffic in different dynamical units.

3.17. Results for the map-equation community detection on the correlation network $C_f$. The optimal compression is obtained by a partition of the network containing 5 communities. The largest one (black) accounts for about 95% of the sensors and covers the whole city. The second largest (red) is itself scattered across the central area of London and has no evident functional structure.
3.18. Results for the spatial modularity community detection on
the correlation network $C_f$. $Q^{spa}$ is optimised by a partition
containing 78 communities. The largest community covers
the whole catchment area and is followed by a series of com-
munities with power law distributed sizes that that do not
provide much functional information (see main text and Fig.
3.19). .................................................. 64
3.19. From top to bottom and left to right, the 2nd, 3rd, 4th and
5th largest communities obtained from the spatial modularity. 65
3.20. Some examples of the networks we can construct through
the linking probability $p(d_{ij},d_0)$ (eq. (3.15)), for the cases
$d_0 = 50, 100, 300, 400m$ ................................. 67
3.21. Plot of the spatial correlation function $C(r,\tau)$ as function of
the spatial distance $r$ for delay $\tau = 0$ (blue) and $\tau = 30 min-
utes$ (gray). The slope is $\beta = -0.26 \pm 0.01$ implying a very
slow decay of correlations over a large interval of distances
(<200 meters). The correlations for $\tau = 30$ minutes have been
lowered for visibility, since the two curves showed remark-
able overlap. ................................. 68
3.22. Top, the size of the largest connected component of the spa-
tial network defined by eq. (3.15) as a function of the link-
ing distance $d_0$. Bottom, $C(r,\tau = 0)$ again as a function of the
distance. Note that the size largest connected component
starts to increase steeply at the same value of distance for
which the correlation function start displaying a power law
dependence. ................................. 69
3.23. Power spectra for the traffic flow (purple) and density (green),
showing very clear power-law scaling behaviour with exponent
$\alpha_f \approx -0.9$ and $\alpha_o \approx -1$ respectively. The long-range
memory effects span a wide range of temporal scales, per-
sisting over weeks. The observable peak at $\omega_{day} \approx 10^{-5}Hz$
corresponds to the daily correlations. The flow data have
been lowered for visibility................. 71
3.24. EFS scaling: $\bar{\sigma}_0$ versus $f_o$. ................................. 73
4.1. $N_J$ vs $\beta$ plot for global and local models on SF (full line) and RG (dashed line) networks ($N = 1000$) with seeded and uniform (inset) initial conditions. Error bars are standard deviations of fluctuations of $N_J$ around its mean value, averaged over the realizations at a fixed $\beta$. Data shown are obtained over 100 realizations of the RG and SF networks.

4.2. $N_J$ as a function of time for global (green) and local (black) models in a RG network with $\beta = 0.9$, averaged over realizations.

4.3. Jammed component expansion mechanism.

4.4. An example of fluctuations PDF for RG for $\beta \approx 0.91 (N = 1000)$. 

4.5. Comparison of the simulation data with the analytical solution of the rate model. The plot refers to the SF networks: simulated data (markers) and fit (line) obtained using Eq. 4.17 for local (green and blue) and global (violet and red) models.

4.6. Example of timeseries of the total number of particles on a Barabasi-Albert scale-free network for different values of $\beta$. From the bottom up, respectively $\beta = 0.60, 0.90, 0.59$. It is evident the sharp transition to a congested state for $\beta = 0.59$, where the nodes are effectively frozen as shown by the linear population increase. For $\beta_c = 0.6$ instead the system manages to deliver many more particles, entering a quasi-stationary state. Interestingly, increasing $\beta > \beta_c$ does decrease performances significantly.

4.7. Stationary timeseries of secondary critical nodes for $\beta \approx \beta_c$ on a scale-free network for $\beta$ near the transition.

4.8. Power spectra near the transition for the ER (red), scale-free networks (green) and lattice (blue). ER and SF networks (see arrows) exhibits powerlaw $\nu^{-l}$ with $l \approx 1$. The dotted lines are a guide for the eye.
4.9. Final particle density $n_f$ vs $\beta$ for Erdös-Renyi graphs (red), Barabasi-Albert networks (green) and square lattice with periodic boundary conditions (blue). A transition from a completely congested state to a highly optimised state clearly appears in all three topologies. 106

4.10. Critical ($\beta = \beta_c$) avalanche size distribution showing the different shapes for ER (red), BA(green) and square lattice (blue). 107

4.11. NYC Manhattan. The red highlighted section is composed of 426 junctions and 1426 one-way links. Two one-way links connecting the same pair of junctions represent a two-way street. 109

4.12. The primal graph built from the red section of figure 4.11. The picture shows how the graphs were obtained from map images of the cities. Junctions were detected and the distances were reconstructed walking along the links. 110

4.13. a) Distribution of capacities $\bar{n}_{ij}$ for single links. The capacity $\bar{n}_{ij}$ of link $ij$ corresponds to the maximum number of cars it can hold. $\bar{n}_{ij}$'s were approximated by the ratio $l_{ij}/h$ between the link length $l_{ij}$ and a standard length $h \approx 5\text{mt}$, representing a standard measure for car length plus headway.

b) Number of delivered cars $D_{tot}$ as function of the forcing rate $N_c$ (akin to traffic demand): the shortest-path routing (red) systematically delivers less cars than the routing with nonlocal (green) and local (blue) information dissemination. For local and nonlocal dissemination, the plotted points correspond to the maximum value of $D_{tot}(\tau)$ at a given $N_c$ (see Fig. 4.14 for details). 116

4.14. The total number of delivered cars $D_{tot}$ as function of the forcing rate $N_c$ and the signalling threshold $\beta$ for the Manhattan section. The top plot refers to the non-local information spreading mechanism, the bottom plot to the local one. 117
4.15. Under local information dissemination (top), a traveller continues until she encounters congestion and then continues following the shortest deviation around the congested area. Under nonlocal information (bottom), a traveller might instead reroute across a trajectory that is further away from the congested area and thus involves areas which were not affected by the emerging congestion. The extra length (or travel time) in the second case can eventually stack up at the network level, making the network more unstable and prone to develop network-wide congestion.

4.16. FD relative to SP (red), NonLocal (green) vs Local (blue) at the peak of delivery section of downtown Manhattan. The slope of the shortest path curve is $\approx -0.1$, while for the local and nonlocal curves is $\approx -0.2$. It might be argued that the local/nonlocal routing only allow for delivery of a large number of cars in the beginning but break down suddenly. Actually the FD show that they both permit to exploit a larger amount of the theoretical flow at the same density.

4.17. Results of limited propagation dynamics for a 10x10 lattice grid. The plots shows represent the number of delivered cars as a function of the signalling threshold $\beta$ and the maximal critical fraction $\epsilon$. The forcing $N_c$ increase from left to right, top to bottom.

4.18. Probability of congestion development as function of the cutoff of nodes able to signal. The nodes highlighted in red are those that are able to signal, selected on the basis of their betweenness centrality.

4.19. $T_c$ versus $N_c$ for fixed (full line) and daily (dashed line) profile forcing for the shortest path routing (red), local congestion-aware routing (green) and decentralised information routing (blue).

4.20. The normalized time to congestion $T_{cong}/T_{max}$ as a function of $\lambda$ for different values of the forcing rate $N_c$. The full lines indicate the expectation values, the dashed the standard deviation values.
4.21. The quantities $H$ (full lines) and $\sigma^2$ (dashed lines) as functions of $\lambda$ for different values of the forcing rate $N_c = 110$ (blue), $120$ (green), $130$ (red). .................................................. 134

4.22. Distribution of travel times for various values of $\lambda$ at a forcing $N_c = 150$. ................................. 135

4.23. Macroscopic Fundamental Diagram for the cases of dynamical aging information (blue squares for $\lambda = 0.07$, red dots for $\lambda = 0.1$ and green upside down triangles for $\lambda = 0.5$), local information only (blue stars), shortest path routing (light grey diamonds). .................................................. 137

A.1. The total number of delivered cars $D_{tot}$ as function of the forcing rate $N_c$ and the signalling threshold $\beta$ for the San Francisco network section. The top plot refers to the non-local information spreading mechanism, the bottom plot to the local one. ................................. 178

A.2. $N_c-\beta$ plot for nonlocal (top) and local (bottom) information dissemination strategies for $c_A = 1$. .................................................. 179

B.1. Details of $C^o$ (left) and $C^f$ (right) for comparison. .................. 180

B.2. Travel time between couple of sensors (from the Google Maps Api) versus geographic distance, calculated from their coordinates. The line is a linear fit with slope $\sim 410^{-3} \text{km/s}$ (adj-$R^2=0.98$). .................................................. 182
## List of Notations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{V}(G), \mathcal{E}(G)$</td>
<td>Sets of nodes and edges in graph $G$, respectively</td>
</tr>
<tr>
<td>$A$</td>
<td>Adjacency matrix of a graph, describing which edges are present</td>
</tr>
<tr>
<td>$a_{ij}$</td>
<td>Element $ij$ of the adjacency matrix $A$, usually either 1 or 0</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>The adjacency matrix for a weighted graph</td>
</tr>
<tr>
<td>$k_i, k_i^{in}, k_i^{out}$</td>
<td>Degree, ingoing and outgoing degree of node $i$</td>
</tr>
<tr>
<td>$\Gamma_i$</td>
<td>Neighbourhood of node $i$</td>
</tr>
<tr>
<td>$\omega_{ij}$</td>
<td>Element $ij$ of the adjacency matrix $\Omega$, with values in $\mathbb{R}$</td>
</tr>
<tr>
<td>$K$</td>
<td>Total degree of a graph</td>
</tr>
<tr>
<td>$M$</td>
<td>Total number of edges in a graph</td>
</tr>
<tr>
<td>$\langle k \rangle$</td>
<td>Average (expected) degree for nodes of a given network</td>
</tr>
<tr>
<td>$P(k)$</td>
<td>Probability distribution function of degrees over the nodes of a network</td>
</tr>
<tr>
<td>$p_{ij}$</td>
<td>Probability for link $(i,j)$ to be present in a given network</td>
</tr>
<tr>
<td>$\kappa_1$</td>
<td>Leading eigenvalue of $A$, used in measuring centralities on a network</td>
</tr>
<tr>
<td>$Q$</td>
<td>Modularity function</td>
</tr>
<tr>
<td>$B_{ij}$</td>
<td>Modularity matrix</td>
</tr>
<tr>
<td>$L(M)$</td>
<td>Minimal length description of a random walker on a network with partition $M$</td>
</tr>
<tr>
<td>$V_{i,n}(t)$</td>
<td>Contribution of component $n$ to the average activity of a node $i$</td>
</tr>
<tr>
<td>$\alpha_T, \alpha_E$</td>
<td>Temporal and ensemble scaling exponents for the Taylor’s fluctuation scaling</td>
</tr>
<tr>
<td>$H_V$</td>
<td>Hurst exponent for the random components of the measured activity (in the context of Taylor’s law)</td>
</tr>
<tr>
<td>$\Theta(\ldots)$</td>
<td>The Heaviside step function</td>
</tr>
<tr>
<td>$\bar{n}_i$</td>
<td>Capacity of node $i$</td>
</tr>
<tr>
<td>$\bar{n}_{ij}$</td>
<td>Capacity of the link from $i$ to $j$</td>
</tr>
<tr>
<td>$J_i(t)$</td>
<td>Congestion state as experienced by node $i$ at time $t$ on a given network</td>
</tr>
<tr>
<td>$F_{ij}$</td>
<td>Flow from node $i$ to node $j$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Fraction of the capacity at which a node will start signalling to its neighbours about its congestion state</td>
</tr>
<tr>
<td>$S_i(t)$</td>
<td>Binary state of node $i$, congested or uncongested, at time $t$</td>
</tr>
<tr>
<td>$L_i(t), n_i(t)$</td>
<td>Load (continuous) and number (discrete) of travellers on node $i$ at time $t$</td>
</tr>
<tr>
<td>Symbol</td>
<td>Definition</td>
</tr>
<tr>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>$d_{ij}$</td>
<td>Graph distance between node $i$ and $j$</td>
</tr>
<tr>
<td>$l_{ijd}$</td>
<td>Length of the shortest path from $i$ to $d$ which crosses $j$</td>
</tr>
<tr>
<td>$p_{ij}$</td>
<td>Probability for a traveller to proceed from node $i$ to node $j$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Rate of information removal</td>
</tr>
<tr>
<td>$N_c$</td>
<td>Driving (number of cars/travellers) introduced at each iteration in the models of information dissemination</td>
</tr>
<tr>
<td>$f_i(t)$</td>
<td>Measured flow by sensor $i$ at time $t$ in the Freeflow dataset</td>
</tr>
<tr>
<td>$o_i(t)$</td>
<td>Measured occupancy by sensor $i$ at time $t$ in the Freeflow dataset</td>
</tr>
<tr>
<td>$\Pi_{f,o}^i$</td>
<td>Average flow or occupancy daily profile for sensor $i$</td>
</tr>
<tr>
<td>$F(t)$</td>
<td>Average flow across sensors (in London) at time $t$</td>
</tr>
<tr>
<td>$O(t)$</td>
<td>Average occupancy across sensors (in London) at time $t$</td>
</tr>
<tr>
<td>$\Sigma_F(t)$</td>
<td>Spatial heterogeneity of flows</td>
</tr>
<tr>
<td>$\Sigma_O(t)$</td>
<td>Spatial heterogeneity of occupancies</td>
</tr>
<tr>
<td>$C_{ij}(\tau)$</td>
<td>Correlation between sensors $i$ and $j$ at a delay $\tau$</td>
</tr>
<tr>
<td>$C(r, \tau)$</td>
<td>Expected correlation at distance $r$ and delay $\tau$</td>
</tr>
</tbody>
</table>
Acknowledgements

The path that brought me to finally complete my PhD has been an incredible and adventurous journey, personally and professionally.
In addition to London, the PhD work brought me to Zürich, Catania, Vienna and Santa Fe. Even the thesis itself has been compiled in many different places: Pisa, London, Sao Paulo, Tamriel, Rio de Janeiro...
It started in the wake of personal disappointment with something I loved dearly and, through weird and unexpected roads, it now seems to have led me to a new perspective and outlook on the things to come.

It is then due to acknowledge the importance of the people that came along for this trip, without whom likely this whole business would not have happened.

First and foremost, my supervisors, Prof. Henrik J. Jeldtoft and Prof. John W. Polak, who can be considered guilty of a number of crimes: they gave me the chance and the trust to explore a previously untrodden path, which skimmed the edges between their respective domains, while at the same time providing me with the information and guidance I needed at the right points down the line. Moreover, they let the leash loose enough on me to allow me to take risks, make errors and eventually grow into –I hope– an independent, if somewhat impulsive, young researcher. I hope the result of this experiment is of their liking.
In consideration of my chronic disorganisation, my kudos go to Ms Jackie Sime, who traded so many jokes with me, but –more importantly– has
probably been as instrumental as my supervisors in getting me to the end of this, patching up my organisational failures.

Along the way, I shared thoughts, discussions, pints, snippets of code, papers and generally time with a very funny and nicely geeky ensemble of people, the (present and past) Complexity and Networks crew, who created the great environment within which this thesis took shape, and each provided a different input: Dr. Tim Evans, Prof. Renaud Lambiotte, Dr. Michael Gastner, Miss Fatimah Abdul Razak, Mr. Eduardo Viegas, Dr. Ignacio Gallo and Dr. Tomas Alarcon (in random order). Moreover, I had the fortune to meet and collaborate with a number of people, who in different ways shaped my PhD work and approach to research: (quasi- and soon-to-be-Drs.) Michael Szell and Roberta Sinatra, with whom I shared a great time in Santa Fe and Vienna, and Prof. Vito Latora, who hosted me in Catania and supplied loads of horse meat among many other inputs; Sam Scarpino, who is responsible for some of the beautiful images of NYC and San Francisco in this thesis, and for a lot of fun in Santa Fe with Felix Hold and Drew Levin; Dr. Pietro Panzarasa, who gave me the chance to teach and work with him, and has been of great support in many ways through the last two years; Prof. Dirk Helbing, who during my stay at ETH Zürich taught me a lot, but mainly that "the hard part is tailoring the math to the problem, not the opposite", and finally Dr. Tomaso Aste and Dr. Tiziana di Matteo, who shared some crazy ideas with me and brought that Italian touch into my time here, something that is easily lost in the Big Smoke.

Indeed, my life in Ankh-Morpork has been at the same time trying and rewarding. Money was short. Tea, rain and tennis were abundant. I was lucky though. I met a kaleidoscopic bunch of friends that redefined and confused my concepts of family and home.

With some I shared a house and food, with others science and squash, with even others wine and theatre\[1\]. Their names –once again in shuffled order– are Dominic Jones, Daniel Seaton, Nicky Sugar Zachariou, (Dr!) Paul Expert, Arne Pohlmeier, Patrizia Baldini, Nico and Racheli Rouge-berg. With Dom and Dan I shared a strongly entangled superposition of house, strong friendship and science; with Nicky, Nico and Racheli an across-the-

\[1\]Or any other permutation of the previous.
Channel fellowship of spirit which overcame languages and cultures; Arne and Patrizia, with their decadent influence on my liver and uplifting effect on my morale, arrived in the second age of this all, but became quickly very much part of it.

A particular note must be made for Paul, for we worked and drank together as friends since the beginning of our PhDs.

Finally, I thank Joanna, to whom—in words better than mine—"I owe the leaping delight" of the life she gives me, which "no peevish winter wind shall chill".

This thesis is dedicated to my generalised family, or social capital as someone likes to call it, because without polpette al sugo (si, nonna, proprio quelle...) and their unyielding support it would have been much harder.

---

2For Imperial’s copyright hounds, this is T.S. Eliot, not me clearly.
1. Introduction

"... the totality is not, as it were, a mere heap, but the whole is something besides the parts ..."

*Metaphysics*,
*Aristotle*

Almost forty years have passed since the day P.W. Anderson published a paper titled "More is different" [1]:

*The main fallacy in this kind of thinking is that the reductionist hypothesis does not by any means imply a "constructionist" one: The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe"

He advocated (or maybe heralded) the change in attitude that followed suit, slowly but surely pushing the dominant reductionist approach toward a more nuanced, holistic approach to scientific research. Indeed, neurones are by no means conscious on their own, one car does not produce congestion patterns, spontaneous magnetisation does not happen to the single atomic spin, society and markets do not exist if humans do not interact, and humans do not work in turn if the web of functional connections among their genes is disrupted. The same argument could be repeated for psychology, systems engineering, architecture, urban planning, infrastructures, astrophysics, conflict resolution, artificial intelligence and many others.

The simultaneous realisation across different fields that monistic approaches were not sufficient anymore to address the arising scientific challenges propelled the paradigm shift and coined a new concept, *emergence*, the idea that large interacting systems can display phenomena that are absent from
the microscopic dynamics of the elements.

A wealth of research resulted. Percolation theory [2], perhaps the most incensed and iconic theory in complex systems nowadays, has found applications spanning the whole spectrum from oil extraction (and coffee making) to epidemic spreading [3]. More importantly, it has brought forward important concepts of statistical physics that now constitute the pillars of complexity research: scaling and critical phenomena seen as main examples of emergent properties.


Cellular automata [7], fractals [8], computational complexity theory [9] tried to quantify and describe the patterns of algorithms, theories and – ultimately – physical systems.

But, what is a complex system exactly? Despite all the research and attention, the community (or rather, communities) has not been able to agree on one definition yet. Many exist, usually demanding the system to display a combination of non-linearity (heritage of Prigogine’s), closeness to the "edge of chaos", self-organisation or emergent properties [10]. However, most agree one fact:

\[
\text{a complex system is composed by many units equipped with non-trivial, usually strong, interactions.}
\]

where non-trivial usually implies a certain degree of non-linearity in the mathematical description of the system.

Urban traffic networks certainly fit among them, spanning many length and temporal scales [11] [12]. From crowds to trains, urban mobility is a fundamental topic for the future development of the human environment. However, despite their centrality and importance for continued urban development and economy, they are just beginning to be understood. Their complexity lie in the tower of interacting networks that compose cities. Vehicular traffic on street networks alone is the convolution of the internal traffic flow dynamics with the socio-economical conditions that define where people live and where they go, all mixed up by complex street topologies and traffic control schemes.
If there is one lesson to be learned from complex networks theory is that, in networks things get complicated (and interesting). Unfortunately, so does traffic. It is then no wonder that a solid theory for urban traffic networks, able to couple spatial and temporal dynamical traffic variables with the classical equilibrium traffic assignment theories, is still missing.

The iconic tool of traffic flow theory, the Fundamental Diagram, that relates density, speed and flow in traffic, is at present ill-defined for networks.

Most traffic theory developed so far focused on the dynamics endogenous to the flow. In a traffic network instead, the constraints come from the structure, not necessarily from the flow. Thinking about dense urban environments, it is easy to convince oneself that, due to the low speed and short links, most of the interesting dynamics will be in the queues, rather than in the density profiles [13–18].

At the same time, it is hard to believe that the network dynamics of a whole city could be described by a relation between average flows and densities alone, neglecting the effect of the distribution of densities. Unfortunately, although the first ideas about urban Fundamental Diagrams go back to Godfrey, Ardekani and Herman and Olszewski et al. [19–22], there is still little (and somewhat contradictory) empirical evidence and only very recently Geroliminis and Daganzo were able to give evidence in favour of the existence of a Macroscopic Fundamental Diagram based on dynamical arguments [23].

In absence of dynamical principles to study network traffic, most urban traffic planning is based on demand models [24, 25], econometric models in which travellers essentially trade routes in a competition to minimise their travel times. They proceed along the classical four steps [26]:

1. Trip generation, which establishes the matrix of origins or destinations of trips across areas and by trip purpose (the OD demand matrix), usually as a function of socio-economical factors;

2. Trip distribution, which assigns origins to destinations, usually through some variation of a gravity model [27];

3. Mode choice, which calculates the fraction of trips using a particular
transportation mode;

4. Traffic (or route) assignment, which allocates trips to routes, usually according to instances of Nash’s equilibrium, the Wardrop’s principle [28] of user equilibrium.

The last step has become critical nowadays. The availability of new information sources and communication devices forced Dynamic Traffic Assignment models [29,30] to evolve toward hybrid assignment systems, in order to include the effects of new information. Even more recently, with the advent of traffic information available to the driver in real-time, Advanced Travel Information Systems (ATIS) required to develop traffic assignments schemes that could take into account pre-trip or en-route travel information [31–37]. Although their accuracy and effects on route behaviour have been questioned [38–41], ATIS are generally considered to improve network performances [42] and the dominant paradigm assumes that (a proper use of) complete information will reduce congestion and lower travel times. This idea, however, has not been proven to be general, but rather disproven in some cases [43–45]. In striking contrast, congestion-aware routing on complex networks [46] based on local information was shown to increase significantly the navigability and capacity of networks. Inspired by these results, Scellato et al. [47] showed in realistic urban networks that routing based on limited knowledge of the local network state allows travellers to better organise their movements, achieving global optimisation and outperforming schemes based on complete information. Despite these contradictory results, there seems to be little thought given to the question of how much information and how far it should be propagated.

In the light of the above considerations, this thesis will address two main questions.
The first, subject of chapter 3 is

What is the macroscopic role of traffic disorder in urban traffic networks?

where by disorder we intend the spatial variability of traffic densities. The analysis will highlight a strong role of traffic heterogeneity in determining
overall network performances and surprising features of the large scale structure of flows. These results suggest that traffic information dissemination can play a role in mitigating such phenomena and, thus, motivate the modelling work of chapter 4 which focuses on the question:

Does an optimal information horizon exist for decentralised information systems in urban traffic networks?

where by optimal length we intend one that is able to effectively smooth the traffic density landscape.

1.1. Outline

The thesis is structured along two main branches, data analysis (chapter 3) and modelling (chapter 4). Because of the differences between them and the range of touched topics, each of these two chapters is organised as an independent entity, containing a motivation, a short introduction to the subject matter and the relevant literature. Also, each of the main chapters ends with a discussion of the relevant results and recommendations for future work. When necessary, additional details on the methods are described in detail before the relevant sections, and an attempt to bridge the linguistic differences between the physics and traffic literatures is made throughout.

The thesis outline is as follows:

Chapter 2 introduces the background concepts and methods, mostly from statistical physics, that will be used in the following two chapters: networks and community detection methods, self-organised criticality and fluctuation scaling.

Chapter 3 focuses on the analysis of traffic data coming from London's street network. The analysis reveals a clear role of the spatial traffic heterogeneity in defining the macroscopic fundamental diagram of the network. In addition, using community detection methods and notions from the theory of self-organised criticality, we show that the city dynamically behaves as a indivisible unit and give evidence in support of an underlying
1.2. MAIN CONTRIBUTIONS

The first main contribution of this thesis is the empirical observation that the spatial heterogeneity of traffic has an importance comparable to the average network density itself in determining the network flow (or accumulation). That is, for a given traffic density a very inhomogenous distribution reduces the overall network flow drastically. This leads naturally to the proposal of generalising the network (or macroscopic) Fundamental Diagram to a 2D surface indexed by first and second moments of the traffic density distribution.

Moreover, we find the presence of long range correlations in the fluctuations, exhibiting an algebraic decay in space. Relating the exponent of the power law to the Taylor’s law for fluctuation scaling, it was possible to interpret the fluctuations as originating from an underlying spatial synchronisation among traffic flows.

Chapter 4 addresses the question posed by Arnott, de Palma and Lyndsay in 1984 [48]: does providing information to drivers reduce traffic congestion? We analyse models of decentralised non-predictive information dissemination coupled to the traffic dynamics: we start with very simple models of flow and increase the complexity up to realistic models, in order to isolate the robust features of the different information dissemination mechanisms. In particular, we concentrate on the information regime (local, non-local, self-organised), finally linking the modelling results to the heterogeneity of the traffic distributions.

Chapter 5 summarises the thesis and discusses the joint results obtained in the two previous chapters, highlighting their relevance for the understanding of traffic dynamics and for potential changes in traffic policies. Further research ideas are exposed too.

A small Appendix follows containing details that were left out of the main text for clarity of exposition or because redundant.

1.2. Main contributions
Along the line of the previous results, we find that the information horizon in a decentralised, non-predictive information dissemination system strongly affects network performances. Scellato et al. [47] had previously shown that local information dissemination performs better than global information.

This thesis extends those results by investigating the interval between strictly local and strictly global information dissemination, and concludes that, in the context of the model of decentralised routing studied, local information dissemination provides the best network congestion resilience.

Finally, our approach naturally links the effects of the range of information dissemination to the network traffic distribution, providing a physical picture to understand the worse performances of long range dissemination. Indeed, longer range dissemination is more likely to produce heterogeneous traffic distributions, which in turn are more likely to seed congested areas and thus make the network unstable.
2. Relevant Background and Methods

In consideration of the branching structure of this thesis this chapter is concerned with presenting an overview of and introduction to the tools used across the thesis; it will also try to put forward convincing reasons for considering concepts and methods borrowed from the field of complex systems.

This chapter is divided in three sections: notions of network theory (section 2.1.1), self-organised criticality (section 2.2) and fluctuation scaling (section 2.3). The need for these notions comes from the approach we will follow in this thesis. In particular, chapter 3 will analyse data coming from London’s street traffic network from the perspective of statistical mechanics with particular reference to the features of systems that self-organise to special states and display properties typical of phase transitions. Furthermore, the analysis of the fluctuation scaling will allow to identify the dynamical origin of such states in the context of street traffic. Moreover, notions from network theory is used throughout the thesis and in particular to detect the existence of dynamical traffic communities in London’s street traffic.
2.1. Networks

Network science over the last ten years underwent a sudden and very fast development. Barabasi's [49] and Watts and Strogatz's seminal papers [50] sparked an incredible interest, which in turn generated a huge amount of work in disciplines as diverse as field theory on networks and the spread of obesity in the American teen population. Due to the sheer volume of work (Barabasi’s paper alone counts 12406 citations at today), this section is not intended as a review of complex networks theory, most of which is rather irrelevant to this thesis. It is instead intended as an introduction to the fundamental concepts that we will be used in the rest of the thesis. The analysis will require some network notions from network theory and of centrality measures, which we introduce in next section. Building on that, the fundamentals of community detection methods (needed for chapter 3) will be presented in section 2.1.2.

2.1.1. Elements of Complex Network Theory

The very first notion needed in discussing networks is their definition. A network is a collection of vertices and links, joining the vertices. For a graph $G$, the sets of nodes of nodes and edges belonging to $G$ are usually denoted as $\mathcal{V}(G)$ and $\mathcal{E}(G)$. The nodes are usually thought of as the atoms of the system under study, while the links (or edges) their interaction. In a social network, people are usually the nodes while the edges represent their ties [51]. Similarly, in a gene-regulatory network, each gene is a node and the influence toward expression or remission of that gene is described by the presence of an edge. Network can be connected, if there exist a path from each node to every other, or disconnected. When the network is disconnected, the connected parts are called components. The terms largest connected component (LCC) refers to the one containing the largest number of nodes. We will see later that in particular network models, the appearance of the LCC signals a significant change in the network’s structure.
2.1. NETWORKS

Adjacency Matrix

The most natural form of investigating a network is drawing it on a piece of paper and looking at it. However, this solution becomes very rapidly impracticable when the number of nodes grows. Moreover, it would be very hard to measure network-related quantities. Instead, let us label the \( n \) nodes in \( V(G) \), where \( n = |V(G)| \) with a label \( i = 1 \ldots n \). \( n \) is usually referred to as the order or size of the network. Edges are naturally labeled by the pair of nodes at their ends, that is \((i, j)\) is the edge connecting nodes \( i \) and \( j \). Consider now a quantity \( a_{ij} \) that is 1 if \((i, j) \in E(G)\) and 0 otherwise, defined for all pairs of nodes in \( V(G) \). It is easy to see that we have just defined a matrix \( A \), whose elements are the \( a_{ij} \), that contains the whole information about the network topology. The matrix \( A \) is called the adjacency matrix of the network \( G \) and is the fundamental block upon which all graph and network theory is built [51]. Using simple properties of the adjacency matrix, it becomes straightforward to identify different type of networks.

A graph that has no self loops, that is links that connect a node to itself, and multiple links between the same pair of nodes, are called simple, implying \( a_{ii} = 0 \) \( \forall i \) and \( a_{ij} \in \{0, 1\} \). Networks displaying self-loops or multiple links are called pseudographs and multigraphs. Undirected networks are networks for whom edges have no direction; this means that if \((i, j) \in E(G)\) then \((j, i) \in E(G)\), which makes the adjacency matrix symmetric,

\[ A = A^T. \] (2.1)

Networks whose adjacency matrix violates the condition above instead are called directed. Indeed, if the link \((i, j)\) belongs to \( G \) while \((j, i)\) does not, then one can say that there is an edge from \( i \) to \( j \) but not the reverse, hence the directionality of the network. We can rephrase this in stricter mathematical terms by saying that for undirected network \( E(G) \) is the set of unordered pairs of elements of \( V(G) \), while for directed networks instead \( E(G) \) is the set of ordered pairs of elements of \( V(G) \). In some cases, one is interested also in describing networks where the strength of the link is important, not only its presence or absence, e.g. the network of passengers travelling between cities [52]. We can account for this by allowing the elements \( a_{ij} \) to take values different from 0 and 1. The resulting network is
called a *weighted* network, because of the weighting on its edges. To distinguish the weighted cases, usually a weight matrix is used instead of the adjacency matrix and is denoted as $\omega_{ij}$ to avoid confusion. In some occasions, a multigraph can be described as a weighted graph. For example, if, in the network of phone calls between people, we associate each call to a link between two nodes, the resulting graph is a multigraph. However, what we would be probably more interested in knowing is how many calls occurred between two persons and this is well captured by a weighted adjacency matrix, with weights proportional to the number of calls.

### Degree and Degree distribution

The simplest property of a node $i$ is its degree, $k_i$, which is the number of edges that link to $i$ or equivalently the number of nodes belonging to the neighbourhood of $i$, $\Gamma_G(i)$. For undirected networks, the adjacency matrix provides a simple and elegant expression for $k_i$:

$$k_i = \sum_j a_{ij}.$$  \hspace{1cm} (2.2)

In directed networks instead, nodes have incoming and outgoing edges. One then needs to define an in-degree and an out-degree, which are given by $k_i^{in} = \sum_j a_{ji}$ and $k_i^{out} = \sum_i a_{ij}$, with the total degree defined as $k_i = k_i^{in} + k_i^{out}$. This carries over to weighted networks,

$$s_i = \sum_j \omega_{ij}$$ \hspace{1cm} (2.3)

where $s_i$ is called the *strength* of a node, rather than its degree. Several global features of a network can easily be calculated from the adjacency matrix. Summing over the nodes’ degrees, one obtains the degree of the whole network, $K$. For a simple network, $K$ takes the form:

$$K = \sum_i \sum_j a_{ij} = 2M$$ \hspace{1cm} (2.4)

where $M$ is the number of edges in $G$, $M = |E(G)|$. The factor 2 accounts for the double counting of edges, since in the sum over the nodes each edge is
counted once for each end. This quantity is not very interesting however because it depends strongly on the number of nodes in $G$. However, it allows us to separate networks in two classes: sparse networks, for which $M$ is of the order $O(N)$, and dense network, for which $M \sim O(N^2)$. Furthermore, if we divide $K$ by the number of nodes $N$, one obtains the average degree of nodes:

$$
\langle k \rangle = \frac{2M}{N} = \frac{1}{N} \sum_i k_i = \frac{1}{N} \sum_i \sum_j a_{ij}.
$$

(2.5)

A natural property yet fundamental based on the degree is the degree distribution $P(k)$ of a network, that is the probability that randomly picking a node from the network it will have degree $k$, with the normalization $\sum_{k=1}^{k_{max}} P(k) = 1$ where $k_{max}$ is the maximum degree in $G$). It turns out that $P(k)$, or more precisely, its functional shape already captures most of the network properties of a network and is the main tool for their classification.

Centralities

Centrality metrics are used to establish the relative importance of a node in a network relatively to other nodes. In a social network, a person with many friends is well connected and is intuitively more central than a person with only one link –physicists for example–.

Degree is then an apparent measure of importance in a network. The larger the degree of a node, the more influence it will be able to exercise on the remaining nodes. The degree centrality of a node is then trivially equal to the node’s degree.

A natural extension of degree centrality is the eigenvector centrality, which measure not only the importance of the node, but also the importance of the node’s neighbours. Going back to the social network example, between two people with the same number of friends, the one with better connected friends will be in a more central position as compared to the other. Eigenvector centrality translates this by giving to each node a score proportional to the score of centrality of its neighbours. It starts giving the same score
2.1. NETWORKS

To all nodes \( x_i = 1 \forall i \) and calculates a new score for each node summing the centralities of the neighbours, \( x'_i = \sum_j a_{ij} \), which can be also put in matrix form as \( x' = Ax' \). Iterating this process, at iteration \( t \) one has \( x(t) = A^t x(0) \), which converges to \[ x \propto \kappa_1^t v \] (2.6)

where \( \kappa_1 \) is the largest eigenvalue of the adjacency matrix \( A \) and \( v \) the corresponding eigenvector. Therefore, the (limit) eigenvector centrality vector is equivalent to the leading eigenvector of \( A \).

**Closeness centrality** takes a different approach. It measures how close one node is from all the other nodes. Denoting \( d_{ij} \) as the geodesic distance between \( i \) and \( j \) on the network – the length of the shortest path from \( i \) to \( j \) –, the mean distance of \( i \) from the rest of the network is

\[
l_i = \frac{1}{n} \sum_j d_{ij}.
\] (2.7)

and the closeness centrality of a node \( i \) is given by \( C_i = \frac{1}{l_i} \). In a disconnected network, all nodes will be at distance \( d = \infty \) from at least one another node, leading to \( C_i = 0 \ \forall i \). In those case, it is customary to take into account only the LCC.

Finally, the most relevant for dynamics on networks, the **betweenness centrality** [54], focuses on the number of paths passing through a given node. Imagine to have a network made of two components, completely disconnected but for one link, say \((l,m)\). It is clear that such link acts similarly to a bridge between the two components and thus \( i \) and \( j \) must be central in the network. Betweenness centrality measure this by asking the following question: if messages were exchanged between all pairs of nodes, what fraction would have to pass through a given node \( i \)?

Denote \( n^s_i \) the number of geodesic (shortest) paths that connect \( s \) and \( t \) and pass through \( i \), and \( g_{st} \) the total number of geodesic paths from \( s \) to \( t \).

\[ ^1 \text{Convergence is assured by the positive semidefinitess of symmetric matrices} \]
Then, the betweenness centrality of $i$ is given by:

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}.$$ (2.8)

This measure of centrality focuses on how much "in between" the network a node is. Moreover, in principle betweenness centrality is unrelated to other measures, for example in a graph composed of a chain of nodes (all with degree 2) will have very low degree but high betweenness centrality. The concept of betweenness centrality is very important in network traffic. Although not all traffic happens along geodesic paths, the betweenness centrality has been found to be a good representation of flows across networks. For example, Newman [55] managed to define betweenness centrality through the movement of random walkers, an idea extended by Borgatti [56] to match different type of walks to the corresponding centrality.

2.1.2. Community Detection Methods

Girvan-Newman Modularity

Community detection is the collective name for a large set of methods that aim to uncover, when it exists, a natural division of a network in a set of non-overlapping communities of nodes. The simplest case is the division of the network in two communities. Intuitively, if we want the two communities to have a meaning, we want to maximise the number of links inside the communities and minimise the number of links running between different communities. This problem is known as "minimum cut" [57] in the graph-theoretic and computer science community. In general however, a network might contain more than two communities and the minimum cut does not allow for more communities. Moreover, the absolute number of links between communities is not a good indicator of the quality of a partition. A good partition is characterised by a larger number of links than expected among nodes belonging to the same communities and less than expected between different communities. The rationale behind this is that a true community structure should be a "statistically surprising arrangement of edges" [58]. It is possible to quantify this by using the modularity [51], which gives positive values to networks with potential non-trivial community structure. In the simplest case, division of the network in two
2.1. NETWORKS

communities-, one usually uses a variable, \( s_i \), to identify to which of the two communities nodes belong to. For a node \( i \), \( s_i = 1 \) if the node belongs to community 1, \( s_i = -1 \) if it belongs to group 2. Using this convention, the quantity \( 1/2(s_is_j + 1) \) is 1 when \( i \) and \( j \) belong to the same community, 0 otherwise. The modularity is defined then as [57]:

\[
Q = \frac{1}{4M} \sum_{ij} \left( a_{ij} - \frac{k_ik_j}{2M} \right) (s_is_j + 1) = \frac{1}{4M} \sum_{ij} \left( a_{ij} - \frac{k_ik_j}{2M} \right) s_is_j,
\]

where \( a_{ij} \) is the adjacency matrix of the network and \( k_ik_j/2M \) is the expected number of links between \( i \) and \( j \). Eq. (2.9) is then summing the expected number of links over all the couples of nodes. Then, maximising \( Q \) gives exactly the most statistically significant partition in two communities. The generalisation to a larger number of communities can be done by repeating the community detection, within the two groups. To do this, it is more comfortable to define the matrix

\[
B_{ij} = a_{ij} - \frac{k_ik_j}{2M}
\]

and the vector of \( s_i \) as \( s \). Using these, the modularity \( Q \) can be written as the bilinear form:

\[
Q = \frac{1}{4M} s^T B s.
\]

After the first splitting of the network in two communities, a natural step would be to remove the links present between the two communities and repeat the process directly. However, this would modify the degrees of the nodes and therefore change the modularity previously calculated. Instead, one can try to quantify how much the modularity would increase or decrease in the case that one of the communities was further divided in two communities. This correction to the modularity obtained by splitting a community \( g \) containing \( n_g \) nodes into two smaller groups can be expressed as:

\[
\Delta Q = \frac{1}{2M} \left[ \frac{1}{2} \sum_{i,j \in g} B_{ij}(s_is_j + 1) - \sum_{i,j \in g} B_{ij} \right] = \frac{1}{4M} s^T B^{(g)} s,
\]

where \( B^{(g)} \) is the matrix of the \( g \)-th community.
where $B^{(g)}$ is a square matrix of dimension $n_g$ whose elements are given by

$$B^{(g)}_{ij} = B_{ij} - \delta_{ij} \sum_{k \in g} B_{ik}.$$  \hfill (2.13)

Naturally, finding the best partition in practice is a very complicated problem. Over the last decade, a number of different technique have been developed, ranging from simulated annealing \[59\] to spectral methods \[60, 61\], passing by synchronisation \[62\] and other folding heuristics \[63\]. A recent review of current methods can be find in \[64\] together with a comparison on a set of benchmark networks \[65\]. It is beyond the purpose of this literature review to describe in detail these techniques. For the results of section \[5.4.1\] we use a particular mechanics based on spectral tri-partitioning \[66\]: the splitting of communities is very similar to the one described above; however, instead of considering only candidate divisions in two groups, also all possible divisions in three groups are considered. This extension allows to reduce the chances (in a similar way to simulated annealing) for the optimisation process to get stuck in a local maximum, instead of reaching the global maximum of the modularity.

**The Map Equation**

Community detection algorithms based on modularity focus on the assignment of links to groups on the basis of pair-wise interactions. This means that modularity probes the network formation process but is not able to map networks flows to communities. However, most real-world networks exist exactly to support some kind flow -money flow among companies, passengers in transport networks, information or signals in information networks-. To understand globally a network and how the flow dynamics on it emerges, one needs to be able to analyse separately from the network formation process. For example, suppose we want to visualise the path of a random walker on a large network. Because of the network’s size, the visualisation would quickly become very unclear. One way to overcome this might be to coarse-grain the network into super-nodes, corresponding to regions of the network where the random walker spends some time before jumping to a new one. This would solve the visualisation problem, but also requires us to be able to find and define such regions beforehand. The
map equation, first proposed by Rosvall et al. [67], is an elegant method to produce the coarse-grained regions, i.e. find the hierarchical community structure by exploiting the "duality between the problem of compressing a data set, and the problem of detecting and extracting significant patterns or structures within those data" [67]. It was recently used to map changes in the patterns of money flow among associates after Lehman’s Brother collapse and to identify trends in the current scientific literature [68]. The idea behind it is simple: suppose a sender wants to communicate efficiently (i.e. with the minimum of bits) to a receiver the position of a walker moving on a given network. The message can simply be the list of successive positions of the walker. However, we want to minimise the length of the message. This problem is the subject of study of Minimum Description Length statistics, an important branch of information theory grounded in Shannon’s coding theorems [69], which pose limits to how much a message can be compressed. A more instructive way of looking at the maximal compression problem is to invert it: instead of thinking how to shorten a list of given node labels, let us ask what is the way to assign labels to nodes in order minimise the length of the message by exploiting the regularities in the walker’s trajectory. One standard way to label the nodes is by using a Huffman code [70], which provides optimally efficient symbol-to-symbol codewords. In practice, it assigns short codewords to frequent symbols, reserving the longer codewords for rarer symbols, thus saving length. For the purposes of encoding the walker’s trajectory, one needs a codebook, that is a set of different codewords for the nodes. However, if the network has a modular structure, the walker will spend long periods of time inside a community before jumping to a new one. This is due to the higher density of links inside a community than between communities and is also the basis of some methods for optimising modularity. It is possible then to improve this description by reusing the same codebook in different regions of the network. The regions are by definition smaller than the whole network and therefore the codewords in their codebooks will be shorter than the ones one would have to use in case of a single codebook. To specify in which region is the walker one uses an additional codebook, called the index codebook. The encoding process finally works as follows: the index codebook is used to identify the region (also called module), then the region codebook is used to specify the walker’s trajectory; when the walker exits
the region, an exit codeword is used and a new region specified through the index codebook. In the region and index codebooks, the length of the region and index codewords are associated to the frequency at which a random walker enters the corresponding node or module. Through the use of multiple codebooks links the minimal description length for the nodes on the trajectory to the problem of finding the best partition with regard to flow, not just topology. This link can be made quantitative by analysing the optimal length of the code associated with a certain candidate partition of the network. Consider a partition $M$ of the network composed of $m$ communities. The lower bound on the code length $L(M)$ is given by [67]:

$$L(M) = q \bowtie H(Q) + \sum_{i}^{m} p_{i} \bowtie H(P_{i}).$$

(2.14)

where $p_{i} \bowtie$ is the probability of staying in the same community, while $q \bowtie$ is the probability of leaving it. Here $H$ is the Shannon entropy [69] and represents the lower bound on the length of the codewords in a codebook. $H$ on a random variable $X$ with state probabilities $p_{i}$ is given by:

$$H(X) = -\sum_{i} p_{i} \log(p_{i})$$

(2.15)

where in our case the $p_{i}$ are the visitation frequencies of the random walker on the node (region), obtained from the stationary random walker distribution [71]. Therefore, $H(Q)$ is the frequency-averaged length of the codewords belonging to the index codebook, while the $H(P_{i})$ are the ones relative to the single region codewords. The first term, the probability of using the index codebook, is weighted by $q \bowtie = \sum_{i=1}^{m} q_{i} \bowtie$ with $q_{i} \bowtie$ being the probability of leaving module $i$. Using the relation $H(Q)$ is given by:

$$H(Q) = -\sum_{i=1}^{m} \frac{q_{i} \bowtie}{\sum_{j=1}^{m} q_{j} \bowtie} \log \left( \frac{q_{i} \bowtie}{\sum_{j=1}^{m} q_{j} \bowtie} \right).$$

(2.16)

Similarly, the module codebook $i$ is used with rate $p_{i} \bowtie = \sum_{\alpha \in i} p_{\alpha} \bowtie + q_{i} \bowtie$, with $p_{\alpha}$ the random walker stationary probability for visiting node $\alpha$ in region $i$, and account for the time the random walker spends in $i$ plus the exit
message. Using this the second term in equation eq.(2.14) becomes:

\[ H(P^i) = -\frac{q_{i\cap}}{q_{i\cap} + \sum_{\beta \in i} P_{\beta}} \log \left( \frac{q_{i\cap}}{q_{i\cap} + \sum_{\beta \in i} P_{\beta}} \right) - \frac{p_a}{q_{i\cap} + \sum_{\beta \in i} P_{\beta}} \log \left( \frac{p_a}{q_{i\cap} + \sum_{\beta \in i} P_{\beta}} \right) \] (2.17)

The lower bound on code length for \( M \) (2.14) can be rewritten as:

\[ L(M) = \left( \sum_{i=1}^{m} q_{i\cap} \right) \log \left( \sum_{i=1}^{m} q_{i\cap} \right) - 2 \sum_{i=1}^{m} q_{i\cap} \log(q_{i\cap}) \] (2.18)

\[ - \sum_{a=1}^{n} p_a \log(p_a) + \sum_{i=1}^{m} \left( q_{i\cap} + \sum_{\alpha \in i} p_{\alpha} \right) \log \left( q_{i\cap} + \sum_{\alpha \in i} p_{\alpha} \right) \]

Upon inspection one sees that \( \sum_{a=1}^{n} p_a \log(p_a) \) does not depend on the partition and that in the other terms containing \( p_{a} \), there is always a sum that washes away the dependence on \( p_{a} \). It is therefore sufficient to track the changes in the \( q_{i\cap} \), which can be obtained very simply from any candidate partition. Moreover, we will be interested in undirected weighted networks and in this case equation (2.18) takes a simpler form:

\[ L(M) = \omega_{\cap} \log(\omega_{\cap}) - 2 \sum_{i=1}^{m} \omega_{i\cap} \log(\omega_{i\cap}) \] (2.19)

\[ + \sum_{i=1}^{m} (\omega_{i\cap} + \omega_{i}) \log(\omega_{i\cap} + \omega_{i}) \]

where \( \omega_{a} \) is the relative node strength, \( \omega_{\cap} \) the total relative weight of links between modules, \( \omega_{i} \) the relative module weight and \( \omega_{i\cap} \) the relative weight of links outgoing from module \( i \).

This same mechanism can be extended to incorporate more than one level of description, by having more than one index codebooks nested inside each other, effectively devising a hierarchy of nested modules. The mechanism is exactly the same as before, the process is simply extended by repeating the indexing step and comparing the lengths of the new descriptions, thus producing a hierarchy of communities \[72\]. In practice, the algorithm used to find the best partition is based on a mixture of greedy and recursive programming\[73, 74\], with modules being built by adjoin-
ing nodes into larger and larger supernodes. For the community detection on the traffic correlation network, we used the code developed by Rosvall and Bergstrom [72] and available at www.mapequation.org.

Spatial Modularity

The Girvan-Newman modularity described in section 2.1.2 is the prominent tool for community detection. Its power comes from considering solely the pair-wise interactions, in other words the elements of the adjacency matrix of the network. The map equation identifies communities by focusing on the information content of the trajectory of a random walker. While both these methods have proven extremely useful [64], there are cases in which this approach overlooks important features of the network as, for example, when nodes and links have properties that are not encoded in the topology. One emblematic case is provided by spatial networks. Nodes and edges are embedded in space, putting severe constraints on many network properties: the presence of hubs is reduced by the cost involved in maintaining long links, while the clustering is increased by spatial proximity [75–81]. Therefore, looking at a spatial network through the goggles of Girvan-Newman modularity, potential anomalies or dynamical correlations on large scales tend to be obscured by the effects of spatial proximity and so the resulting community structure yields little novel information. At the same time flows tend to be stronger between nodes closer in space and this would bear on the efficiency of the map equation. Here we introduce an extension of the Girvan-Newman model that allows to take into account the effects of distance between nodes. As mentioned before in this chapter, modularity tries to identify statistically surprising arrangement of nodes. This implies a statement about what is to be expected. So, the crucial point of the modularity function $Q$ is the null model $P_{ij}$, which contains the assumptions we are making. Let us first consider an unweighted undirected network, characterised by its binary adjacency matrix $A$. In the standard Girvan-Newman modularity (Eq. (2.9)) the assumptions made are that the degrees of the nodes are the most important features and that the degree of a certain node is uncorrelated from the degree of its neighbouring nodes [82]. Under these assumptions, the probability of finding a link between a node $i$ with degree $k_i$ and $j$ with degree
2.1. NETWORKS

\( j \) is simply given by the number of ways they can connect \((k_i, k_j)\) divided by twice the total number of links \((2M)\), therefore

\[
P_{ij}^{GN} = \frac{k_i k_j}{2M}. \tag{2.20}
\]

where \( m = \sum_{i<j} a_{ij} f \) is the number of distinct links in the network. For a weighted undirected network, one needs to use a weighted adjacency matrix \( \Omega_{ij} \), where the equivalent of the degree for a node \( i \) is the so-called strength \( \omega_i = \sum_j \Omega_{ij} \) and the null weighted model becomes:

\[
P_{ij}^{GN} = \frac{\omega_i \omega_j}{2\omega}. \tag{2.21}
\]

where, in analogy to before, \( \omega = \sum_{i<j} \Omega_{ij} \) is the total network strength.

In case one wants to include different properties of the network, one has then to extend \( P_{ij} \) to include those quantities. A well known example that is very relevant to spatial networks is the gravity model (GM) [27, 83]. The idea in such models is that the flow or interaction (the \( \Omega_{ij} \) previously defined) between two nodes (or locations) \( i \) and \( j \) can be effectively modelled by a distance-dependent term and by a second term that is proportional to some node-related quantity \( N_i (N_j) \): in the case of passenger flow between cities, one reasonable quantity is the cities’ populations; for money flows between banks, the total bank social capital. So, for the gravity null model we write:

\[
P_{ij}^{Spa} = N_i N_j f (d_{ij}) \tag{2.22}
\]

where \( f \) can be any (generally decreasing) function of the distance \( d_{ij} \) between \( i \) and \( j \) and the \( \omega_{i,j} \) play a role similar to the mass. The function \( f \) is defined as:

\[
f (d_{ij}) = \frac{\sum_{i,j \mid d_{ij} = d} \Omega_{ij}}{\sum_{i,j \mid d_{ij} = d} \omega_i \omega_j}, \tag{2.23}
\]

The functions defined above, usually called deterrence functions, can be determined directly from the data and so does not require a fit to some functional form [27], thus making the method independent of external hypothesis. At this point, one can use the modularity machinery to optimise the modularity defined in terms of \( P_{ij}^{Spa} \) instead of \( P_{ij}^{GN} \), thus overcoming the limitations of community detection based only on network topological
Finally, the complete expression for the null models used are:

\[
P_{ij}^{Spa,o} = \frac{N_i N_j \sum_{i,j \mid d_{ij} = d} C_{ij}^o}{\sum_{i,j \mid d_{ij} = d} \langle o_i \rangle_t \langle o_j \rangle_t},
\]

(2.24)

\[
P_{ij}^{Spa,f} = \frac{N_i N_j \sum_{i,j \mid d_{ij} = d} C_{ij}^f}{\sum_{i,j \mid d_{ij} = d} \langle f_i \rangle_t \langle f_j \rangle_t},
\]

(2.25)

We refer to the expressions above as *spatial null model* and we denote the corresponding modularity function to be maximised as \(Q^{Spa}\).

### 2.2. Self Organised Criticality

The seminal paper by Bak, Tang and Wiesenfeld (BTW) [84] introduced the concept self-organised criticality (SOC), merging the concepts from criticality and self-organisation. Criticality in itself was born in the field of equilibrium phase transitions: systems near a second-order phase transition were found to display scale invariant behaviour [2,85]. The observation of power-law phenomena in many natural systems, like the Gutenberg-Richter law for earthquakes [86] or Zipf’s law of ranked word frequency distribution [87], prompted the application of the ideas of criticality to non-equilibrium dissipative systems, by linking their spatial and temporal degrees of freedom and so providing a seductive explanation for the two ubiquitous phenomena that had no clear explanation until then, 1/f noise and spatial self-similarity.

Bak’s classical model was a pile of sand [88], driven by the addition of...
further sand grains, but slowly enough for eventual avalanches of grains to stop before new grains were added. On a one-dimensional lattice with open boundary on the right, the dynamics is very simple. The numbers $z_n$ represent the difference of the number of grains $h$ on successive sites, $z_n = h(n) - h(n + 1)$ where $n$ labels the site. The dynamics then takes the form:

$$z_n \rightarrow z_n + 1 \quad (2.26)$$
$$z_{n-1} \rightarrow z_{n-1} + 1 \quad (2.27)$$

When the slope $z_n$ at site $n$ exceed a threshold $z_c$, grains would topple according to:

$$z_n \rightarrow z_n - 2 \quad (2.28)$$
$$z_{n\pm 1} \rightarrow z_{n\pm 1} + 1 \text{ for } z_n > z_c \quad (2.29)$$

with the boundary conditions $z_0 = 0$, $z_N \rightarrow z_N - 1$ and $z_{N-1} \rightarrow z_{N-1} + 1$ for $z_N > z_c$, where $N$ is the size of the lattice. The observable quantities was the distribution of avalanche lifetimes, that is the number of successive topplings before an avalanche stopped, and their distribution of size, the total number of toppled grains. Bak et al. then interpreted the height at a site as its energy and thus the size of an avalanche $s$ corresponded to the total dissipated energy

$$s = \int F(t)dt = \int dt dx f(x, t) \quad (2.30)$$

where $F(t)$ is the total dissipated energy at time $t$ and $f(x, t)$ the local dissipation. They then looked for the relation between the spatial and the temporal scaling exponents. The one-dimensional sandpile however turns out not to be critical, due to the effect of the boundary, which effectively pinches the avalanches. However, in higher dimensions Bak was able to show, numerically and analytically, that a power-law distribution of lifetimes $D(t) \simeq t^{-\alpha}$ and sizes $D(s) \simeq s^{-\tau}$ leads to a power spectrum for the dissipation $F(t)$ of the form

$$S(f) = \int \langle F(t_0 + t)F(t)\rangle \exp(2\pi if t)dt \sim f^{-2+\alpha}. \quad (2.31)$$
2.2. SELF ORGANISED CRITICALITY

This can be done by relating the sizes and lifetimes of avalanches. The argument goes as follows: at the critical point, the systems will present minimally stable clusters of all sizes, which implies that local perturbations would propagate on all scales. If one assumes that an exponent $\gamma$ for their growth within the cluster, it is possible to relate the lifetime of a cluster to its size, $\tau^{1+\gamma} \sim s$. The distribution of lifetimes is recovered, by noting that:

$$t^\alpha = D(t) = \frac{s}{t}D(s(t)) \frac{ds}{dt} \sim t^{-(\gamma+1)\tau+2\gamma}. \quad (2.32)$$

This yields the connection between the size and lifetime of the avalanche and finally the spectral property of the fluctuations, therefore relating the the spatial self-similar properties with the temporal features, the flicker noise.

The BTW model just described is just one of numerous SOC models (see for example [89, 91]), which have been studied on a number of different topologies. For example, Goh and Lee et al. [92] studied a variation of the BTW model on scale-free uncorrelated networks, where the threshold for toppling depended on the node degree, and observed the existence of two different scaling regimes depending on the value of the degree distribution of the network.

Vespignani and Zapperi [93] and Dickman et al. [94] described the dynamics of the two most well known SOC models, sandpiles and forest fire models [95, 96], relating them to absorbing phase transitions through a unified mean-field master-equation approach, based on the single site approximation to the master-equation. Criticality was identified in terms of the singularities of the zero-field susceptibility, showing that a stationary state with scaling characteristics of self-organised criticality emerges in the limit of vanishing driving rates. In particular, SOC emerged from the competition of driving forces and dissipation at the boundaries, in such a way as to let the system hover around their critical point. Prüssner and Peters [97] challenged the universality of these models, suggesting that the description based on absorbing state phase transitions is incomplete. Despite the lack of accepted theoretical framework and continuing debate [98, 99], the concept intrigued the scientific community because it promised to provide a natural explanation for a number of different phenomena observed
across a wide range disciplines [100], among the many brain dynamics [101], neuronal networks [102], non equilibrium system [103, 104], solar flares [105] and plasma physics [106], neural networks [107, 108], optimisation [109], economy [110, 111], evolution [112–114] and, naturally, traffic [115–119].

2.2.1. SOC and Cascades

Important phenomena associated with SOC and sandpiles models are the so-called cascading failures. For example, power blackouts are among the most disruptive large-scale accidents in power grids, where the blackout is the result of an avalanche of overload failures that eventually conducts to a network-wide, rather than local, blackout. Many models of cascading failures exist, treating subject as diverse as power grids and earthquakes [120, 121]. On networks, a very simple though representative model, proposed in [122], considers a network, in which each vertices has a load represented by the vertex’s betweenness centrality. In addition, each vertex $i$ has a maximal load—capacity—:

$$c_i = (1 + \alpha)b_{0i}, \quad (2.33)$$

where $b_{0i}$ is the vertex’s initial betweenness centrality. The constant $\alpha \geq 0$ accounts for the tolerance, showing how much an initial load can be exceeded before failure. A cascading failure in this model looks as follows:

(i) Remove a vertex. This causes a redistribution of loads of the other vertices: $b_{0i} \rightarrow b'_{0i}$.

(ii) Delete all overloaded vertices, that is the vertices with $b'_{0i} > c_i$.

(iii) Repeat the process until no overloaded vertices remain. The number of the deleted vertices is the size of the cascade.

Motter and Lai measured the ratio $G = N_{after}/N$, where $N$ and $N_{after}$ are, respectively, the original number of connected vertices in a network and the size of remaining largest connected component after the cascading failure. The resulting $G(\alpha)$ depends on the topology of a network, $\alpha$, and the characteristics of the first vertex to fail. For instance, a highly connected node
2.3. FLUCTUATION SCALING

will generally have a greater effect on the whole network than a low degree one.

In a random regular graph \( (k_i = \langle k \rangle \forall i) \), for \( \alpha > 0 \), \( G = 1 \), and only if \( \alpha = 0 \), the network will be completely fail, \( G = 0 \). On the contrary, in networks with heavy-tailed degree distributions (thus heterogeneous structure), \( G \) depends on the degree or betweenness of the first removed vertex. Scale-free networks \(^3\) with \( \gamma = 3 \) \( \alpha = 0 \) gives \( G = 0 \) for any starting vertex, while \( \alpha \to \infty \) results in \( G = 1 \) (see also [123]). What is more interesting however is the form of \( G \): when the first removed vertex is chosen at random, the cascade is large (\( G \) strongly differs from 1) only at small \( \alpha \), and \( G(\alpha) \) rapidly grows from 0 to 1. If the first vertex is chosen from ones of the highest degrees, then \( G \) gently rises with \( \alpha \), and cascades may be giant even at rather large \( \alpha \).

Lee et al. [124] numerically studied the statistics of cascades in this model defined on scale-free networks with \( 2 < \gamma \leq 3 \) and found the critical point at \( \alpha = \alpha_c \approx 0.15 \). For \( \alpha < \alpha_c \), the avalanches cover a finite fraction of the network, spanning the whole system, while for \( \alpha > \alpha_c \) the avalanches are finite. These authors observed that at the critical point, the size distribution of avalanches has a power-law form, \( \mathcal{P}(s) \sim s^{-\tau} \), where exponent \( \tau \approx 2.1(1) \) in the whole range \( 2 < \gamma \leq 3 \).

These models will be the inspiration for the first two models of the information propagation in chapter 4. The complication in those cases will be that the dynamics will not belong only to the network topology, but to the interaction between the network structure and capacities, and the quantities flowing on the network.

2.3. Fluctuation Scaling

One of the reasons for the great interest raised by SOC was its apparent ubiquity, turning up in contest as different as astrophysics and evolution. With the growing availability of large datasets regarding very different fields, another apparently ubiquitous phenomenon has gathered the attention of the complex systems community. Taylor’s law, from the influential paper [125] by the ecologist L. R. Taylor in 1961, or in modern terms Fluctuation Scaling (FS), has affirmed itself as a tool able to shed light—at least

\(^3\)That is networks characterised by a degree distribution \( P(k) \sim k^{-\gamma} \).
in some cases on the origin of the scaling observed in a very wide range of physical, technological and biological systems.

In its simplest form, the law relates the fluctuations of a positive extensive (i.e. additive) quantity to its average value through a power law relation. The first evidence for this law is generally held to be a paper [126] by H. Fairfield Smith about the yield of crop fields, but many followed. Fluctuation scaling was reported in animal populations [127][129], cells growth [130], heavy ion collisions [131], epidemiology [132][133], protein expression [134], river basins [135], economic growth [136][137], traffic on the Internet [138][139] and in street networks [140]. A comprehensive review can be found in [141].

There are two types of fluctuations that can be studied, temporal and ensemble fluctuations.

Assume we can measure a quantity \( f_i \), additive and positive, on the nodes of a system (labeled by \( i \)). For an interval \([t, t + \Delta t]\) the signal can written as the sum of the contribution of its constituents:

\[
\begin{align*}
 f_{i}^{\Delta t}(t) &= \sum_{n=1}^{N_{i}^{\Delta t}(t)} V_{i,n}^{\Delta t}(t).
\end{align*}
\]  

(2.34)

Here, \( N_{i}^{\Delta t}(t) \) is the number of constituents of node \( i \) during the considered interval, and \( V_{i,n}^{\Delta t}(t) \geq 0 \), so that the temporal average of \( f_{i}^{\Delta t} \) is strictly positive. A good example is thinking about the number of cars passing over a traffic sensor: during \([t, t + \Delta t]\) there are \( N_{i}^{\Delta t}(t) \) cars that pass over the sensor \( i \), each spending \( V_{i,n}^{\Delta t}(t) \) time over the sensor, then total time spent covered by the sensor \( i \) is given by this formula.

The average of Eq. (2.34), denoted as \( \langle f_{i}^{\Delta t} \rangle \) is

\[
\langle f_{i}^{\Delta t} \rangle = \frac{1}{Q} \sum_{q=0}^{Q-1} f_{i}^{\Delta t}(q\Delta t) = \frac{1}{Q} \sum_{q=0}^{Q-1} N_{i}^{\Delta t}(q\Delta t) V_{i,n}^{\Delta t}(q\Delta t),
\]

(2.35)

with \( Q = T/\Delta t \) and \( T \) the measurement total time. The variance is then calculated as usual:

\[
\sigma_{i}^{2}(\Delta t) = \langle [f_{i}^{\Delta t}]^{2} \rangle - \langle f_{i}^{\Delta t} \rangle^{2},
\]
2.3. FLUCTUATION SCALING

where the brackets stand for a temporal average. The notation is unfortunately cumbersome, but it is needed to avoid confusion between the sensors’ signals \( f_i \) and the components of the sensors’ signals \( V_{i,n} \).

It follows from the definitions that \( \langle f_i^{\Delta t} \rangle = \Delta t \langle f_{i+1}^{\Delta t} \rangle \) and \( \langle f_i^{\Delta t} \rangle = \Delta t \langle f_i \rangle \).

One can then drop the temporal indices. Following the notation of [141], we use \( \langle f_i \rangle \) without the upper index to identify the variables corresponding to \( \Delta t = 1 \).

In these variables, Taylor’s law states that the relationship between the temporal average and variance is:

\[
\sigma_i(\Delta t) \propto \langle f_i \rangle^{\alpha_T},
\]  

(2.36)

where the node \( i \) is the varying variable and \( \alpha_T \) is usually in the range \([1/2, 1]\) ([141] and references therein). The index \( T \) indicates that the quantities are temporal averages as in Eq. (2.35).

The second type of scaling, ensemble fluctuation scaling (EFS), relies on a second quantity, according to which it is possible to group the sensors. For example, in the case of Smith’s study [126] of crop yield, this variable was the size of the fields. In general, the requirement is that the elements of a group share similar statistical properties, for example the same number of constituents \( N \), the energy (for ion collision or cosmic rays [131] or size-like parameters. In addition to the temporal averages for the nodes, it is possible to define the average within each of these groups. This is the reason of the term ensemble. It is defined as:

\[
\overline{f_S} = \frac{1}{M_S} \sum_{i:S_i=S} f_i^{\Delta t}(t)
\]  

(2.37)

\[
\overline{\sigma_S^2}(\Delta t) = [\overline{f_S^{\Delta t}}]^2 - \overline{f_S^{\Delta t}}^2.
\]  

(2.38)

where \( S \) is the size-like quantity. The temporal quantity \( t \) and \( \Delta t \) are fixed, and the summation includes the nodes \( i \) with equal size \( S_i = S \), and \( M_S \) is the number of such nodes. Also in this case, fluctuation scaling is of the form:

\[
\overline{\sigma_S}(\Delta t) \propto \overline{f_S}^{\alpha_E},
\]  

(2.39)

where we compare different groups by varying \( S \), while \( \Delta t \) is kept constant.
When we do not want to distinguish between the two scalings (since most expressions can be written for both), we just use $\alpha$.

### 2.3.1. Meaning of the scaling exponents

It is possible to show that for $\sigma_i^2 = \left\langle \left( f_i(t) - \langle f_i(t) \rangle \right)^2 \right\rangle$, can be written in term of the single constituent’s signals as:

$$
\sigma_i^2 = \Sigma_{V_i}^2 \left\langle N_i^{2H_{V_i}} \right\rangle + \Sigma_{N_i}^2 \langle V_i \rangle^2,
$$

where $\langle V_i \rangle$ and $\langle N_i \rangle$ are the mean for of $V_{i,n}$ and $N_i$ and $\Sigma_{V_i}$ and $\Sigma_{N_i}$ the respective variances. The symbol $H_{V_i}$ is the Hurst exponent of the constituents, which is a measure of their correlation.

Two special (limiting) cases for the exponents were pointed out by De Menezes and Barabasi: $\alpha = 1/2$ and $\alpha = 1$. Indeed it is possible to construct simple and general scenarios that produce these values. Consider a system in which every node $i$ is formed by a constant number $N_i(t) = N_i$ of constituents. Assume the constituent’s signal $V_{i,n}(t)$ to be i.i.d. for all $i$, $n$ and $t$, with the same average $\langle V \rangle$ and variance $\Sigma_{V_i}^2$. Eq. (2.40) implies

$$
\sigma_i^2 = \langle f_i \rangle = \Sigma_{V_i} \langle V_i \rangle = \Sigma_{V_i} \langle f_i \rangle,
$$

where we used $\langle f_i \rangle = N_i \langle V \rangle$. This gives immediately $\alpha = 1/2$. Consider now the opposite case: a system with a constant number of identical and synchronized constituents for each node ($N_i(t) = N_i$, $V_{i,n}(t) \equiv V_i(t) \forall i$). Eq. (2.40) becomes

$$
f_i(t) = \sum_{n=1}^{N_i} V_{i,n}(t) = N_i V_i(t).
$$

which imply $\langle f_i(t) \rangle = N_i \langle V_i(t) \rangle$, and $\sigma_i = N_i \Sigma_{V_i}$, finally giving:

$$
\sigma_i = \frac{\Sigma_{V_i}}{\langle V_i \rangle} \langle f_i \rangle \propto \langle f_i \rangle,
$$

(2.41)
and $\alpha = 1$.

It is easy to see why these two classes were associated to case corresponding to the different origins for fluctuations in the system. De Menezes and Barabasi [138] argues that for systems in dominated by their internal dynamics, i.e. with none or weak external coupling, $\alpha$ should be 1/2. Strong external driving instead would produce system-wide coordination that would result in $\alpha = 1$.

This was however disproven, since the value of $\alpha$ was found to be dependent on various other effects: the impact inhomogeneity (i.e. components very different contributions to the sensors’ signals) [144–147], the size of the temporal window $\Delta t$ [139] and correlations between the constituents [148]. All of these mechanisms allow to produce values of $\alpha$ in the range $1/2 \leq \alpha \leq 1$.

For this thesis, the correlations between constituents are of particular interest as we will see in section 3.5.1.

The main idea is that spatial correlations can have a role on the emergence of fluctuation scaling, as for example in the context of ecology [149, 150]. In particular, we consider correlations between constituents. For easiness of exposition, one can assume that the elements are placed on a one-dimensional lattice, with activity that is correlated over long range in space, with a power law decaying correlation function,

$$C(\Delta n) \propto \langle V_{i,n} V_{i,n+\Delta n} \rangle - \langle V_{i,n} \rangle^2 \propto \Delta n^{2H_V-2}. \tag{2.42}$$

The Hurst exponent associates values $H_V > 1/2$ to positively correlated patterns, $H_V < 2$ to anticorrelated pattern, while uncorrelated (random) patterns give $H_V = 1/2$. From Eq. (2.40) it is possible to write the expression for the fluctuations of the combined activity of all constituents is:

$$\sigma_i^2 = \Sigma_{V} N_i^{2H_V} \propto \langle f_i \rangle^\alpha,$$

that implies

$$\alpha = H_V. \tag{2.43}$$

effectively equating the effects of spatial correlation with the emergence of anomalous scaling relationships.

It is a capital mistake to theorise before one has data.

_Sherlock Holmes, Sir Arthur Conan Doyle_

3.1. Background and Aim

In the eighty years since Greenshields’s seminal paper on the Fundamental Diagram [151], traffic flow theory made great leaps in the comprehension of the complicated non-linear phenomena that precede and follow the break down of free flow giving way to a zoo of different congested states: stop-and-go waves, wide moving jams and extended jams [152–155]. Most of the empirical confirmation of flow models, both microscopic and macroscopic, have come from work on highways (in particular continuous stretches) and arterial roads [156–158]. This bias was mostly due to the difficulties inherent in recording and managing large quantities of data. Moreover, these type of roads lent themselves better to theoretical understanding, thanks to the simple structure, which could be decomposed in uniform sections with the addition of ramps for inflow and outflow. From this perspective, urban networks are very different: the structure of the street network and the traffic control strategy are dominant, making less than clear what is the role and importance of traffic flow models and driving behaviour. Also, the amount of data necessary to study in detail large urban areas was very hard to obtain, much less analyse, for a long time.
Very recently this tendency has reversed due to data slowly becoming available about large urban areas. However, dense, complex environments like modern world cities have not been studied in depth yet. An example of this is the current discussion about the existence and definition of a meaningful Macroscopic Fundamental Diagram (MFD) for urban areas. Different proposals have been put forth that try to build on the consolidated knowledge of single link Fundamental Diagram.

In this regard, London offers an unique opportunity. It has grown inexorably over the last few hundred years by progressively incorporating villages and little towns. This resulted in a complex multi-core topology that cannot be easily reproduced by standard models of geometrical, or spatial, growing networks. Like many other non-planned cities, London displays a very broad distribution of road length and capacities in addition to a broad distribution of connectivity of the dual graph, reminding of power-law behaviour. Previous studies described areas of cities as containers exchanging traffic (reservoir model), or topological statical objects. A small attempt at including a dynamical element in the generation of network traffic patterns was made by Tadic et al. using random walkers on the dual graph of the Chinese city of Nanjing. In this work, a flock of random walkers diffused on the dual graph of the real street network, where the choice of the dual graph was justified by the fact that moving on it better represents the navigation process across a network in comparison to moving on the primal graph. The timeseries of the queue lengths on the different nodes of the dual graph were then correlated in order to produce a correlation matrix between the nodes populations. The correlation matrix itself was then interpreted as a fully connected, weighted adjacency matrix and community detection methods were applied in order to identify the dynamical structure. The underlying assumption is that the topology of the network is the dominating element in the dynamics defined on the network, in this case (biased) random walkers. However, when other factors come into play, as they do in the case of travellers in an urban network, one cannot anymore be certain that the topological analysis alone will be sufficient to characterise the evolution of the dynamical process on the network. Even the simple fact that travellers have an origin and a destination can significantly change the evolution of the traffic network state (as we show later in section 4.4).
3.1. BACKGROUND AND AIM

The Freeflow data in principle allow us to study the same "congestion" communities and uncover the functional dependence of the various part of the street network. Surprisingly, we will see that the real picture is very different -and much richer- than what is usually found for dynamical processes on networks (see [168] for examples).

Most of the research based on Freeflow data so forth goes in the direction of pattern-matching techniques [169–172]. In these approaches, large amounts of historical traffic data are collected and used to build a database of network traffic states. Once this is done, these states are linked to the traffic control actions that were taken to relieve eventual problems together with a score describing how much a specific traffic control action improved the situation [173–175]. When a potentially problematic traffic state arises in the real traffic network, the state is matched to one (or more) classified historical ones that are close according to the previously specified metrics. Then, one chooses the traffic control strategy that scored the highest among the ones adopted as a reaction to the historical states. The rationale behind these methods is that the best control strategy can arise by successive trials and evaluations in the context of repeatable, or at least, predictable traffic states [176, 177]. These assumptions work very well when one is confronted with uniform road segments (such as in highways or arterial roads) or small network sections, however it becomes less reliable when the network size grows. The number of possible traffic configurations in fact becomes rapidly very large, hindering the capacity of pattern-matching techniques to find close historical states in the high-dimensional configuration space. More importantly, a very large amount of historical data becomes necessary to calibrate reliably the scoring method and this is not always possible or manageable.

Finally, pattern-matching methods assume that traffic states that are close according to some similarity measure will evolve along similar trajectories in configuration space. In other words, the system is assumed to be non-chaotic or display weak chaotic behaviour, so that the evolution on short-time scales trajectories of close states the traffic configuration space will not become very dissimilar. We will see that this hypothesis fails when the entire London urban network is considered, as so probably in other large cities too.

In this chapter, we describe the dataset and the preprocessing necessary
3.2. THE FREEFLOW PROJECT AND ITS DATASET

for the analysis (section 3.2). We then compare the empirical results for London with the current research on the existence and properties of the Macroscopic Fundamental Diagram and show that the spatial occupancy heterogeneity is a fundamental quantity to describe the macroscopic evolution of the network. In section 3.4 we investigate the correlation networks built from the traffic timeseries and frame the results as an example of self-organised criticality in a spatially extended, strongly interacting system (section 3.5).

3.2. The Freeflow Project and its Dataset

The Freeflow Project, a collaborative effort of 15 partners in academia, government and industry and led by Imperial College London, is one of the winners in the Future Intelligent Transport Systems (FITS) initiative, promoted by the UK Department of Transport. The aim of the proposals funded by FITS generally lies in the direction of developing novel solutions to transport problems, in particular by adopting multidisciplinary approaches that are able to stimulate and accelerate the industry investment toward innovation.

In this context, Freeflow’s goal is to change the way traffic data is used through the inclusion of new types of data (CCTV feeds, for example), the development of reliable network performance metrics and intelligent decision support.

At the moment, Freeflow is collecting data in London, York and Kent. In this thesis, we focus on the data coming from London for two reasons: they are at present the most complete, and they permit to investigate traffic dynamics in an extremely heterogeneous and dense urban environment. Across the Greater London Authority area, Freeflow is recording data from 3256 sensors, mostly distributed in the Central London area but reaching also Tottenham to the north, Brixton to the south, Stamford to the east and Chiswick to the west. The area covered is roughly a square with a side of 20 km, for a total of 400 km$^2$ (Figure 3.1).

Each sensor is an ILD able to measure the flow and occupancy, respectively the number of vehicles passing over the detector and the fraction of time that the sensor passes covered by a vehicle. The latter can be related to the traffic density [178–180] under relatively mild hypothesis of uniform
3.2. THE FREEFLOW PROJECT AND ITS DATASET

Figure 3.1.: The map of the part of Central London covered by the Freeflow sensors. There are in excess of 3000 ILDs, stretching from Wimbledon and Brixton in the south to Tottenham Hale and Walthamstow in the north, from Stratford in the east all the way to Kensington and Chiswick in the west, for a total catchment area of about 400 km$^2$.

Sensors aggregate data over 15 minutes, so the measures effectively are integrated flows and time-averaged occupancies over such temporal windows. This coarse-graining does not allow for a high-frequency analysis of the traffic features, but, as we will see in the following, is still sufficient to analyse the macroscopic statistical properties of the traffic network considered as a spatially extended, strongly interacting system.

The dataset analysed here covers the month of September 2009 for 3256 ILD sensors. Figure 3.1 shows the catchment area. The sensors, represented by the red pins, are clearly distributed unevenly on the street network. Indeed, the majority are located in the central region of the catchment area, clustering around Marble Arch and the surrounding areas of Camden, Hyde Park and Westminster. These regions are the most impor-
Figure 3.2.: Map of the time-averaged occupancy $\langle o_i \rangle$ for sensors (top). The sensors’ size and colour both represent the average value of the occupancy for the corresponding sensor. There is a large spatial disorder in the distribution of traffic with the most central sensors experiencing the largest densities. The values (bottom) are distributed around a well defined peak, which can be fit well by a normal distribution with average 8.6 and width 12.9, which accounts for 95% of the total sensors. The remaining 5% lies in the peak (close to occupancy 92%) for very high densities, corresponding to the red, large dots in the picture.
Figure 3.3.: The distribution of occupancies for the Freeflow sensors separated for periods of the day. The single sensor occupancy distributions show an initial plateau followed by a power-law decay.
tant ones in the daily London traffic, since most of the vehicles concentrate there during the day. In Figure 3.2, the distribution of average occupancy for sensors is shown. It is very easy to see that there is a large heterogeneity in the expected occupancies even of sensors that are very close in space. However, the largest densities are visibly concentrated in the central area. The distribution displays two separated peaks, a Gaussian one for low densities \((\mu = 8.5, \sigma = 12.8)\), and one for very large densities, centred at an occupancy of \(\sim 92\). The first one contains about 95\% of the sensors in the network, the remaining belonging to the high density peak. This means that there exist locations in the street network that are systematically undergoing large traffic densities. Interestingly, they are not tightly clustered, being on the contrary rather mixed up with other sensors. Moreover, there is no uniform distribution of average occupancies; sensors tend to be either very congested or very free.

This static picture is useful to develop an intuition regarding the traffic distribution on the street network. However it is unable to provide insights into the dynamical evolution of the network traffic states. More interesting is the distribution of sensor instantaneous occupancies (Figure 3.3). We find that the distribution is characterised by an initial quasi-plateau for low occupancies and a power law tail for larger occupancies. This implies that, in analogy to other systems displaying such distributions [120] [181] [182] and although low-density states are very likely, the probability of finding a sensor in a state of large density is significant due to the presence of the thick tail. The aggregated observations for the whole day mix together states of the network that are radically different, for example rush hours and the night. We see however that the distributions obtained separating different sections of the day are very similar, with a small bump around 20–40\% in the morning and a small increase for the probability of high events in the day and evenings. Therefore, from the point of view of the distribution of network densities, the heterogeneity is not accounted by the inter-day variability of the network traffic. Differences emerge if one studies the distribution of single sensor occupancy for classes of different average network occupancy \(O(t) = 1/N \sum_{i=1}^{N} o_i(t)\). In the dataset \(O(t)\) varies between 0 and 30. Define then the sets \(T_{\text{low}}, T_{\text{med}}, T_{\text{high}}\) to be:

\[
T_{\text{low}} = \{t \in (0, T)|O(t) < 10\} \quad (3.1)
\]
3.2. THE FREEFLOW PROJECT AND ITS DATASET

\[ T_{med} = \{ t \in (0, T) | 10 < O(t) < 20 \} \] (3.2)

\[ T_{high} = \{ t \in (0, T) | O(t) > 20 \} \] (3.3)

In Figure 3.4 we plot the distribution of occupancies obtained restricting to the three different sets of times. In this new cut of the data, the occupancy distributions are different. For low total occupancy, the distribution shows a nice power law behaviour, \( P(o_i) \sim o_i^{-\omega} \) with \( \omega = 2.4 \pm 0.1 \). When the density increase, as expected the distribution shifts to the right, enlarging the plateau, but the best fit to the tail has the same exponent as before (within errors), implying that, although somewhat reduced, the heterogeneity of traffic distribution carries through the whole total network occupancy range.

This implies also that the network traffic density field is very disordered. This observation is not unexpected: travellers tend to concentrate on few major roads, reducing the density of traffic on the smaller roads. As we argued in section 4.5, this disorder is likely to create instabilities, especially during periods of intense traffic, reducing the carrying capacity of the network and effectively decreasing its performances.

The aim of the next sections will be to quantify the effects of this disorder and to try to identify the peculiar dynamical structures that emerge in London’s traffic. In particular, we aim to highlight the recurring correlations in the traffic landscape (section 3.4).

To do this, we will be studying both the original timeseries and the ones where the dependence on the hour of the day is removed. The detrended ones will be very important when we discuss the behaviour of the system with respect to the propagation of traffic density fluctuations and the correlations between different locations on the network.

For sensor \( i \), denote by \( f_i(t) \) and \( o_i(t) \) its flow and occupancy timeseries and build its typical daily flow profile \( \Pi^f_i(\tau) \) and occupancy profile \( \Gamma^o_i(\tau) \):

\[ \Gamma^f_i(\tau) = \frac{1}{n_{days}} \sum_{m=0}^{n_{days}} f_i(\tau + mT_{day}) \] (3.4)

\[ \Gamma^o_i(\tau) = \frac{1}{n_{days}} \sum_{m=0}^{n_{days}} o_i(\tau + mT_{day}) \] (3.5)
3.2. THE FREEFLOW PROJECT AND ITS DATASET

Figure 3.4.: The distribution of single sensor occupancies for the Freeflow sensors clustered according to the total network instantaneous occupancy $O(t)$.

where $\tau$ is the time of the day ($\tau \in (0, T_{\text{day}})$), $n_{\text{days}}$ is the number of days the timeseries lasts and finally $T_{\text{day}}$ is the length in time steps of a day. The analysed timeseries span a month and have a temporal resolution of 15 minutes, so $n_{\text{days}} = 30$ and $T_{\text{days}} = 24 \cdot 60/15 = 96$. We then built for each sensor the detrended timeseries $\hat{f}_i(t)$ and $\hat{o}_i(t)$ by subtracting the profile function:

\[
\hat{f}_i(t) = f_i(t) - \Pi'_f(mod(t, T_{\text{days}})); \quad (3.6)
\]

\[
\hat{o}_i(t) = o_i(t) - \Pi'_o(mod(t, T_{\text{days}})); \quad (3.7)
\]

where $mod$ is the modulo function. Figure 3.5 reports an example of the resulting profiles (sensor n00/005g1).
3.2. THE FREEFLOW PROJECT AND ITS DATASET

Figure 3.5.: An example of the flow and occupancy profile (sensor no. n00/005g1). The errors are the standard deviation of the flow/occupancy values at given time.
3.3. London’s Macroscopic Fundamental Diagram

3.3.1. Current Understanding

The main tool in traffic studies to characterise the state of the flow is the Fundamental Diagram (FD). The FD for a single road segment is well defined and has been studied in depth over the years, since the seminal paper of Greenshields in 1935 [151]. Harder is the definition of a Macroscopic Fundamental Diagram to describe a traffic network or at least some part of it.

Very recently, due to availability of real and simulated data, various effort were made to build a convincing MFD [23, 160, 183]. In particular, Geroliminis et al. [23] showed that a MFD relating production (the product of average flow and network length) and accumulation (the product of average density and network length) can be obtained for neighbourhoods of cities in the order of 5-10 km$^2$. They also argued that, conditional on accumulation, large networks behave predictably and independently of their Origin-Destination (OD) tables and they suggested that the dynamics of the rush hour can be predicted quite accurately without the knowledge of disaggregated data (single-street capacity, junctions, signalling etc). Daganzo et al. [160] showed that the MFD relating average flow and average density exists on any street with blocks of diverse widths and lengths, but no turns, even if all or some of the intersections are controlled by arbitrarily timed traffic signals. For networks then, Daganzo et al. [160] also derived an upper bound for average flow, conditional on average density, that matches acceptably the experimental results of [23]. Ji et al. [184] suggest that these results regarding MFDs for neighbourhoods could be used to implement refined versions of perimeter control strategies, by limiting the number of vehicles in a certain area and thus keeping the neighbourhood near the MFD peak and proposed a clustering method to identify homogenous areas in an urban network.

Most of these treatments however disregard the additional degrees of freedom that are present when one considers a network instead of single

---

1Section 3.4 will analyse an issue very close to this problem and show how and why community detection methods fail to pick up such areas in London.
street. In fact, while the usual Fundamental Diagram relates well-defined densities and flows on a road, for a network any density defined as any type of averaged single link density will incur in a degeneracy of the states it describes. In other words, many different traffic states might have the same network density: for example, consider one state where all links have the same density (very uniform) and one where half of the links have zero density and half have double the average density. Therefore, in principle there is at least one extra degree of freedom needed –e.g. the standard deviation of the density (or flow) distributions– in order to have a meaningful description of the network traffic state. In the literature, a few studies addressed this issue. Mazloumian et al. [185] showed through simulations that the MFD of a grid network is well defined and single valued only when one cluster data points by the spatial density heterogeneity. The latter in fact changes the capacity of the network. So in a real network one would expect to see a superposition of curves in the MFD, corresponding to traffic states with different degrees of heterogeneity. De Martino et al. [44] in the context of adaptive learning drivers showed that the most efficient state was the one characterised by the highest uniformity across links in both time and space. The same conclusion can be seen as a natural requirement for a Wardrop equilibrium. Very recently, Geroliminis et al. [186] replaced the homogeneity constraint for the existence of a low scatter MFD with the following conjecture [186]:

\[
\text{If the spatial distribution of link occupancy is the same for two different time intervals with the same average network occupancy then these two time intervals should have the same average flows.}
\]

This conjecture is supported by the Yokohama arterial traffic data, but unsupported by the freeway network of the Twin Cities. This contradiction was explained by referring to the contrasting features of the two networks: the redundancy of the arterial network, the traffic control schemes and hysteresis phenomena in the freeway network.

The case of London is still different. The sensors are positioned both on arterial and minor urban roads, most are deeply entrenched in the urban environment and, while a large redundancy is topologically present, it is not always clear to what extent drivers do take advantage of it, for example because of the (high) risk of getting lost.
In this contest the main questions we ask are:

Does a well defined, uniquely-valued relation between network densities and flows hold for London (like in the standard MFD setup)?
If not, what are the other relevant observable quantities?
Does traffic pass from free flow to congested when the density (or occupancy) becomes larger than a critical value (usually around 30% of occupancy), or does flow break up happen through a different mechanism?

3.3.2. The Role of Spatial Heterogeneity

In order to approach these questions, we begin by adopting the prescription of Daganzo et al. [160] for an unweighted MFD and define the network occupancy \( O(t) \) and flow \( F(t) \) as the unweighted average values of the \( o_i, f_i \):

\[
O(t) = \frac{1}{N} \sum_{i=1}^{N} o_i(t) \quad (3.8)
\]

\[
F(t) = \frac{1}{N} \sum_{i=1}^{N} f_i(t) \quad (3.9)
\]

Figure 3.6 displays the MFD for London, obtained from the complete Freeflow dataset. The first observation is that, although a certain scatter is present, it appears that the MFD is defined for London, as there is no evident bifurcation. Second, although at every time \( t \) there always is certain number of sensors which are congested (see Fig. 3.4), the whole network as a whole does not seem to go into the congested region of the MFD (\( F \) is still monotonically growing with \( O \)). However, these observations do not take into account the previous remarks about the role of traffic heterogeneity.

Let us consider first the effects of the occupancy distribution. In the top panel of Fig. 3.6 we plot the MFD coloured according to the value of the rescaled occupancy heterogeneity, which we define as \( \Sigma_O(t) = \sigma_O(t)/O(t) \), where \( \sigma_O(t) \) is the standard deviation of the occupancy across the individual sensors at time \( t \). The definition of \( \Sigma_O \) is chosen as to provide a normalised measure of the broadness of the occupancy distribution. Figure 3.7 isolates the dependence of \( F \) on \( \Sigma_O \). We find that a very good fit is
3.3. LONDON’S MACROSCOPIC FUNDAMENTAL DIAGRAM

Figure 3.6.: The Macroscopic Fundamental Diagram for London ($F(t)$ vs $O(t)$) where data points are coloured according to the value of $\Sigma O(t)$.

given by a simple quadratic polynomial form ($a_2 x^2 + a_1 x + a_0$), whose coefficients are reported in table B.1.

In addition to the usual (concave) dependence of $F$ on $O$, this highlights a strong dependence on $\Sigma O$. London displays consistently smaller flows as $\Sigma O$ increases. This result suggests that, to obtain a complete description of the network dynamics, one cannot stop at a unique-valued MFD $F(t) = g(O(t))$, where $g(x)$ is a function of only $x$.

On the contrary, we need to generalise the expression to explicitly include the density heterogeneity, $F(t) = h(O(t), \Sigma O(t))$. The Geroliminis conjecture implies then that $F$ is a function of only $\Sigma O$ and $O$. That is, if $F = h(O, \Sigma O, X)$ where $X$ is some other quantity, then $X = v(O, \Sigma O)$. Conceptually, this means that the MFD is equivalent to a two-dimensional surface in $(O, \Sigma O, F)$ space, rather than to a curve. Based on the dependency found earlier, a naive attempt is to try and fit a polynomial quadratic in both traffic variables, that is we assume $F$ to be of the form $F = \sum_{i=0}^{2} a_{i,2-i} \Sigma O^{2-i} O^i$, where the $a_{xx}$ are fit parameters\(^2\). As shown in Figure 3.8, the fit we obtain

\(^2\)A quadratic form was chosen for the dependency on $O$ too out of simplicity, being the simplest (reasonable) concave form. Naturally, more efforts, both analytical and empirical, will be needed to establish which one is the exact functional form of $F$. 
Figure 3.7.: The relationship between $F$ and $\Sigma_O$ for the London data. The best fit is obtained with a quadratic polynomial (adj-$R^2 = 0.9862$, RMSE=12.85).
is very close to the observed data (adj-$R^2=0.99$, RMSE=10.1).
If we continue the fitted surface to include $O = 0$, a few inconsistencies appear though (Fig.3.9). For example, $\forall \Sigma_O$ when $O \to 0$, $F$ increases, instead of going to zero as one would expect for an empty network. These inconsistencies reflect the lack of data near the borders of the domain of definition of $O$ ($O \in (0,100)$).

If we had a theory for how the MFD dynamically emerges from a microscopic network model, we could impose boundary conditions on $O$ and $\Sigma_0$ and this would take care of the large flows near zero network occupancy. Unfortunately, we do not have such theory yet. Nonetheless, we can improve our fit by imposing some effective boundary conditions, in the form of additional data points to account for special cases. Great care however is needed, because it is hard to control the effects of perturbing the data in this way. Fortunately, there is one point we can be sure of, namely the origin, $O = 0, \Sigma_O = 0$: the situation of a zero network occupancy $O = 0$ and zero spatial heterogeneity $\Sigma_O = 0$ must correspond uniquely to the empty network. Hence, we can safely assert that $F = 0$. Figure 3.10 shows the fitting surface obtained with the addition of the origin to the Freeflow data (adj-$R^2=0.98$). Continuing the surface obtained over a larger interval of values of $O$ (Fig. 3.11), we find some very desirable properties. First, the maximal flow $F_{\max} = \max_{\Sigma_O, O} F$ is bounded and $F$ for vanishing $O$ is decreasing. Second, $\Sigma_O$ emerges as the other determining factor of the network capacity and thus the MFD. Finally, it provides a reasonable null model against which to compare new data, in order to test the robustness of the surface and whether it could potentially provide a tool for network managers.

3.3.3. Flow heterogeneity and Hysteresis

Until here, we focused on describing the network flow $F$ as a function of the occupancy and its spatial heterogeneity. We have not considered yet

---

3 A caveat is due here: $R^2$ can be badly behaved at the origin, so this value needs to be taken with care. The analysis of residuals however confirms the goodness of the fit.

4 Indeed, also in this case $F(O = 0, \Sigma_O = 0) \neq 0$. This is due to the fact that we did not require the fit to pass exactly through the origin, as it would be in the case of true boundary conditions. Nonetheless, the surface gradient near the origin goes toward smaller flows, in contrast to what happened to the surface obtained in Fig. 3.8.
Figure 3.8.: The Freeflow MFD and the surface obtained through the polynomial fit.

Figure 3.9.: The continuation of the surface obtained for wider intervals of $O$ and $\Sigma_O$. The MFD surface unrealistically grows for small occupancies. Negative flows have been set to 0 in this visualization.
3.3. LONDON’S MACROSCOPIC FUNDAMENTAL DIAGRAM

Figure 3.10.: The Freeflow MFD and the surface obtained through the polynomial fit with the addition of the origin as an additional data point.

Figure 3.11.: The continuation of the surface obtained from the Freeflow data with the addition of the origin. At variance with Fig [3.9] the new surface is concave and produces a reasonable behaviour for the MFDs. Negative flows have been set to 0 in this visualization.
the role or origin of the flow spatial heterogeneity.

In analogy, to what done before we define \( \Sigma_F(t) = \sigma_F(t)/F(t) \) as the rescaled flow heterogeneity at time \( t \), where \( \sigma_F(t) \) is the standard deviation of the instantaneous single sensor flow distribution. Colour coding data points according to \( \Sigma_F \) (Figure 3.12), the MFD shows that, similarly to what happens for occupancy, at a given \( O \), traffic configurations with larger \( \Sigma_F \) display smaller \( F \), confirming the diminishing effects of disorder on network performance.

It is not possible, however, to find a simple functional form for the dependence of \( \Sigma_F \) on \( O \) or \( \Sigma_O \). This is due to a large clockwise hysteresis loop\(^5\) (Figure 3.13) in \( \Sigma_F \). The continuous line links points that are successive in time. Hence, the loop accurately follows the temporal network dynamics. The periodicity, as could be expected, is daily. In the early morning, the traffic is characterised by low occupancies and high flow heterogeneity; during the day the heterogeneity decreases reaching a minimum in the middle of the day; during the late afternoon and night, the system decays to the initial state, along the less disordered branch (lower branch). In addition, \( \Sigma_F \) shows an anticlockwise hysteresis loop (inset) as a function of \( \Sigma_O \): during the load up of the network (early morning and first part of the day), \( \Sigma_F \) for a given \( \Sigma_O \) is larger than during the unloading phase (afternoon and night).

In the light of the previous results (see [187, 188] and Figure 3.10), one would expect to see hysteresis patterns in the MFD too. Although not obvious at first sight, the hysteresis loop becomes apparent when comparing the hysteresis in Figure 3.13 with the colour coding of Figure 3.12. For clarity, the top panel of Figure 3.14 shows the MFD for the first seven days of recorded data. The daily hysteresis pattern is clear, despite variations in the maximal occupancy and flow over different days. The hysteresis is present every day in the recorded data, so it cannot be imputed to exceptional events, as in the case of the truck drivers’ protest and following car diversions examined in [187]. Rather, it is a property of London’s Origin-Destination matrix coupled to the city’s topology and, possibly, traffic control system. The bottom panel of Figure 3.14 shows the typical hysteresis daily, obtained by averaging data points corresponding to the same times.

\(^5\)Attempts at producing a surface fit similar to Figure 3.10 yield very bad fits, due to the scatter of data and the hysteresis loop.
of the day. For network occupancies larger $O \geq 13$, the hysteresis loop is anticlockwise. This means that the loading of the network progresses along the lower branch (smaller flows), while the unloading along the upper branch (larger flows). By comparison with Figure 3.12 and 3.6, this also means that the spatial flow and occupancy distributions are more heterogeneous during the loading phase than during the unloading phase. Finally, the daily loop lies on the MFD surface obtained through the MFD surface fit.

Collecting all these observations, the typical scenario in London is the following: early in the morning, the traffic occupancy is low, but very spatially disordered, inducing suboptimal flows in the network (Figure 3.10); as the time passes and the occupancy increases, the network manages to uniform the traffic distribution, *climbing* along the MFD surface towards the optimal flow trajectory.

Since the hysteresis loop periodicity is daily, it is natural to think that a time dependent traffic demand is the reason for the observed evolution along the MFD surface. Indeed, in the morning and the evening the traffic is mostly inbound and outbound respectively, while during the day the traffic demand can be considered roughly isotropic across Central
Figure 3.13.: *Main plot*, the flow spatial heterogeneity $\Sigma_F(t)$ versus the network occupancy $O(t)$. The line represents the temporal succession of data points. A clear clockwise hysteresis loop is present: during the first part of the day the system evolves along the upper branch (larger disorder), until it reaches the minimum $\Sigma_F$. Successively, it unloads along the lower branch (smaller disorder). *Inset*, $\Sigma_F(t)$ versus $\Sigma_O(t)$ displaying an antclockwise hysteresis loop.
Figure 3.14: Top, anticlockwise hysteresis in the Macroscopic Fundamental Diagram for the first week of recording. Bottom, the typical hysteresis loop: each points has been obtained averaging over the data points corresponding to a certain time of the day (the resolution is 15 minutes, so daily there are 96 measure). The loop is anticlockwise: during the morning rush period, the network follows the lower branch in the plot, while the unloading in the afternoon happens along the upper branch.
London\cite{189}. At the time of writing, a (preprint) article by Zhang et al. \cite{188} appeared online, investigating the relationship between traffic spatial heterogeneity and the appearance of hysteresis patterns. Their results appear to support the role of demand in shaping the MFD: in their simulations, time-dependent MFDs on square lattices produce complicated hysteretic loops, which mix anticlockwise and clockwise patterns. Through an elegant analogy with the theory of charge holes in solids, they conclude that anticlockwise (clockwise) loops emerge in over-saturated (under-saturated) networks, due to the propagation of gaps (congestion seeds) in the traffic. However, in the Freeflow data we observe an anticlockwise loop already at very low occupancies, which is not consistent with the interpretation given in \cite{188}.

Another possible explanation is given by Gayah et al. \cite{183}, where the hysteresis is produced by the exchange of traffic between two traffic reservoirs, representing disjoint areas of an urban network. In this case, hysteresis loops of different orientation emerge from the ratio between the respective capacities of the reservoirs and the turning rates. In particular, Gayah et al. predict that anticlockwise loops would emerge when the capacities of these reservoirs are very unbalanced. It is unclear however what or how should be defined as a reservoir (a road, a neighbourhood, etc.). If one considers each road as reservoir, a massive multi-reservoir version of the model proposed by Gayah et al. might reasonably apply to London, due to the city’s complex and multi-centred street topology and its skewed fat-tailed capacity and length distributions \cite{161}.

The answer seems to lie somewhere in the middle. The complex structure of London makes the street network more likely (than regular networks) to develop congested areas even at low occupancies, in a similar way to the jamming transitions happening on scale-free networks \cite{190-192}. This provides seeds for congestion to emerge on the most central links on the network, which then reinforce and maintain the spatial heterogeneity, thus reducing the network flow. A change in the traffic demand might then allow some areas (the reservoirs) to alleviate their congestion, pushing the system up along the surface towards higher flows.

A natural question at this point is: do these reservoirs exist in London? How can they be defined?
In fact, if indeed one could identify a good set of reservoirs, perimeter control strategies could be used to effectively relieve congestion. In addition, strategies could be implemented to make the reservoirs more homogeneous internally, in order to obtain the best possible local flow.

In short, reservoirs represent the dynamical units composing the traffic network and -it could be expected- defining its performances.

There are two ways to go about their definition: one can define them by hand according to some external criterion, e.g. municipal boundaries; or one can let them emerge autonomously from traffic dynamics itself and then define the borders accordingly.

In the latter, finding reservoirs becomes equivalent to identifying units that evolve coherently under varying conditions, which is a common problem to a number of other applications. In financial networks for example it is important to understand which financial sectors evolve (and crash) together and how they rely on the others. For politics, identifying social units in the underlying tapestry of personal acquaintances and contacts can be used to predict political schisms and the appearance of new political parties. In neuroscience, disentangling the modular structure of the brain from its collective evolution is an open problem. Finally human interaction and mobility, can be characterised and predicted making use individuals’ contact patterns.

In the case of London’s traffic, the interesting object is the network of dynamical dependences between sensors.

It can be obtained easily by correlating the traffic variables measured at different sensors and thus defining a correlation matrix between them. Such matrix is then be interpreted as the adjacency matrix of a weighted fully connected network and the community detection performed on it. In the next section we follow this approach with some surprising results and discuss their implications.

3.4. Correlation Networks and Congestion

In this section we apply methods from community detection analysis to the correlation network built from the traffic sensor data. As mentioned in the introduction of this chapter, there have been some attempts to study
the topological and community structure of traffic networks. Most of these have been topological studies [161, 199, 202], with a few exceptions based on simulation data [165, 184].

Here we use the Freeflow data in order to uncover the underlying dynamical structure and, if possible, obtain the reservoirs that compose London. In order to do this, we want to build a network that is representative of the influence that one sensor (node) has on another and vice versa. One way of doing this is using a measure of similarity between quantities that belong to the sensors. Consider a pair of sensors \(i\) and \(j\) and denote the quantity under study by \(X\). The simplest similarity measure one can use is the Pearson correlation coefficient \(C_{ij}^X\), defined as:

\[
C_{ij}^X = \frac{E[(X_i - \mu_X)(X_j - \mu_X)]}{\sigma_X \sigma_X}
\]

(3.10)

where \(\mu\) and \(\sigma\) are the mean value and the standard deviation of the random variable \(X\). Considering all the pairs of sensors, the previous equation defines the elements of the correlation matrix \(C^X\) of the system. The following step is to interpret this matrix as the weighted adjacency matrix of the network of sensors.\(^6\)

The Freeflow data has two quantities so we can construct two correlation matrices, one for occupancy, \(C^o\) and the second for the flow, \(C^f\). However, some care is necessary. If we use the timeseries \(\{o_i(t)\}\) and \(\{f_i(t)\}\), we are including in our correlation measure also the effect of the daily traffic routine. This produces though unreasonably high correlations among sensors: any dynamical dependence or interaction between sensors would be washed away by the systemic effects of increased flow and occupancy during rush hours. It would seem that the whole system is extremely correlated, simply because it is forced it in every point with a similar driving. The resulting correlation network would have no interest, because all nodes would be more or less correlated to the others in the same way. Our aim instead is to identify functional, dynamical dependency among sen-

\(^6\)Note that this correlation matrix is different from the one defined in [60, 203]. Equation (3.10) is the correlation matrix obtained from the sensor data, that we interpret as the weighted adjacency matrix of a network, not dissimilarly from the interaction term of a spin-glass theory [57]. In [60, 203] (and reference therein) instead, the correlation matrix is built starting from a given adjacency matrix.
sors. Suppose that at a certain instant a given sensor is recording higher occupancies than its usual value at that time of the day. We want to predict which other nodes will more likely be influenced by the additional congestion that is developing on the first sensor. The object of interest are then the fluctuations over the expected values of flow and occupancy and how these fluctuations correlate over the street network. Therefore, to build the $C_{ij}^{f}$, we use the detrended time series $\hat{f}_i$ and $\hat{\delta}_i$ defined as in Eqs. (3.6-3.7) through the usual sample correlation coefficient:

$$C_{ij}^{f} = \frac{\sum_t (\hat{f}_i(t) - \mu_{\hat{f}_i})(\hat{f}_j(t) - \mu_{\hat{f}_j})}{\sqrt{\sum_t (\hat{f}_i(t) - \mu_{\hat{f}_i})^2} \sqrt{\sum_t (\hat{f}_j(t) - \mu_{\hat{f}_j})^2}}$$ (3.11)

$$C_{ij}^{o} = \frac{\sum_t (\hat{\delta}_i(t) - \mu_{\hat{\delta}_i})(\hat{\delta}_j(t) - \mu_{\hat{\delta}_j})}{\sqrt{\sum_t (\hat{\delta}_i(t) - \mu_{\hat{\delta}_i})^2} \sqrt{\sum_t (\hat{\delta}_j(t) - \mu_{\hat{\delta}_j})^2}}$$ (3.12)

Figure 3.15 shows a comparison of the distance matrix between pairs of nodes and the corresponding correlation coefficient. Already at a naive visual inspection, the distance matrix shows a higher degree of structure. For example, the sensors of the top left region of the distance matrix consistently have low values of distance, that is they are close to each other, while there is no such evident structure in the correlation matrix $C_{ij}^{f}$ shown. This suggests that the traffic recorded by the sensors does not show dynamical clustering on the network, painting a picture in which streets are correlated disregarding of their position, which is indeed an interesting hypothesis, although a rather surprising one. It would in fact imply that traffic is coordinated at the city level.

To verify this hypothesis and quantify the presence of community structure in the correlation matrices, we use three different methods of community detection analysis.

The first one, the Girvan-Newman modularity [51, 82, 204], is the most famous and well established method. It is based on the optimisation of the modularity function, which tries to maximise the number of links within communities and punishes links across communities. As detailed in section 2.1.2, Girvan-Newman modularity requires a null model, against which the data are contrasted in order to find the statistically significant group of nodes that are more linked than one would expect randomly. This can
be considered as the first level of refinement in finding communities on (weighted) networks.

The second method, the map equation [67, 72, 205], is very different in principle. It is based on the idea of a random walker moving on the given network. In particular, weighted networks can be seen as a Markov Chain describing a random walk on the network themselves. The map equation then tries to give the minimal length description (in the language of information theory described in section 2.1.2). In order to accomplish this, the network is not only partitioned into communities, but a hierarchy of partitions is created through the use of codebooks. This provides an additional dimension to the community detection analysis, since in this way it is possible to know also what communities are more related within a given partition.

Finally, the third method we use is a modification of the Girvan-Newman optimisation method [197]. Note that the two methods described above lack any consideration about the nature of the correlation matrix whose best partition they are trying to identify. In our case, however, we are interested in the correlation between vehicles moving on a network with the additional effect of a poor temporal resolution (15 minutes). One can easily imagine then that sensors that are close will be much more correlated than faraway ones not only because of their dynamical dependence, but also due to simple physical proximity. This third method, spatial modularity, is able to identify the former and eliminate the latter. In order to this, we need to account for the physical distance between sensors. In the context of modularity, this means using a null model that compares the observed correlations between sensors at a certain distance with the correlation expected at that distance. In this way, the method is able to discriminate between correlations due to proximity and to real dynamical dependences in traffic.

3.4.1. Results

We calculated the correlation matrices for both flows and occupancies and applied the community detection methods described in section 2.1.2. In our case, we assume the same mass ($N_i = 1\forall i$) for all sensors, because the relative strength of nodes is already taken into account by the Pearson co-
3.4. CORRELATION NETWORKS AND CONGESTION

Figure 3.15.: Top, visualization of the distance matrix between pairs of sensors. Bottom, the correlation matrix $C_f$. Already at a visual inspection it is possible to see, from the groups of sensors at small distances, that the network is spatially clustered as one would expect in a street network. On the contrary, the flow correlation matrix does not show any evident community structure.
3.4. CORRELATION NETWORKS AND CONGESTION  

efficient (see Eqs. (3.11)[3.12]). Following [197], for the distance function we use the expected correlation values $C_{i,j}^o$ at a certain distance $d$:

\[
 f^o(d) = \frac{\sum_{i,j | d_{ij} = d} C_{i,j}^o}{\sum_{i,j | d_{ij} = d} 1} = \frac{1}{n_{d_{ij}=d}} \sum_{i,j | d_{ij} = d} C_{i,j}^o = I_{ij}^{Spa,o} , \quad (3.13)
\]

\[
 f^f(d) = \frac{\sum_{i,j | d_{ij} = d} C_{i,j}^f}{\sum_{i,j | d_{ij} = d} 1} = \frac{1}{n_{d_{ij}=d}} \sum_{i,j | d_{ij} = d} C_{i,j}^f = I_{ij}^{Spa,f} . \quad (3.14)
\]

Figure [3.16] shows the results of the Girvan-Newman modularity maximisation on $C^f$. The top panel contains a map of the sensors coloured according to the community they belong to. The main feature are the two largest communities, which together account for $\sim 85\%$ of all the sensors. The red clearly identifies the central areas of London, while the black dots identify the peripheral sensors. The results are not unexpected, since the central area is where most of the general traffic is concentrated during the day, while the peripheral arteries sustain mostly the commute traffic in the morning and the evening. On the whole, this suggests that the whole traffic network is split in two large dynamical areas (we are correlating the flows here). At a closer look, communities show a large spatial overlap. In particular, such overlap appears not only near the spatial boundaries between the two, but also deep in the bulk. An overlap near the boundaries could be explained by small errors in identifying the maximum modularity partition (e.g. due to degeneracy near the maximum or resolution issues [206,207]). Here instead the overlap between the two largest communities is significant (see for example the right part of the top panel in Fig. 3.16) and the same applies to the other two communities detected (yellow and orange). This result cannot then be considered as an artefact of the method used and implies that the communities effectively overlap in space. However, it is hard to understand why a sensor deep in the (spatial) bulk of a community might belong to another spatially distant community while another sensor, close to the first, is not.

One way to think about this is hypothesise that a hierarchical structure of communities exists. The Girvan-Newman modularity would not be able to resolve since it looks only for the partition with the highest modularity. If

\[\text{See Appendix B.1.1 for a few technical details involved in this case.}\]
Figure 3.16.: Results for the Girvan-Newman community detection on the correlation network $C^f$. The Girvan-Newman modularity is optimised by a partition of the network containing 4 communities. The two largest ones (red and black) account for about 85% of the sensors and correspond respectively to the northern central area and the southern peripheral areas of London. The third (yellow) appears to be spatially scattered across the central area of London, but has no evident functional structure. The fourth is almost invisible in the network. All communities show a large degree of spatial overlap and it is not easy to identify units allowing to decompose the traffic in different dynamical units.
Figure 3.17.: Results for the map-equation community detection on the correlation network $C_f$. The optimal compression is obtained by a partition of the network containing 5 communities. The largest one (black) accounts for about 95% of the sensors and covers the whole city. The second largest (red) is itself scattered across the central area of London and has no evident functional structure.
some hierarchy is really present, then one should coarse-grain the network and obtain a highest modularity partition at each level of coarse-graining. This process is laborious and requires defining a process of box-counting, which on networks is an open problem in itself [208–210]. The multilevel map equation approach however is naturally tailored to identify the underlying hierarchical structure through the use of nested codebooks. In Figure 3.17 we show the results of the map equation on the network. Surprisingly, we find that, beyond the trivial cases - all nodes in one community or each node in a different community - there is only another level of the community hierarchy, which contains one very large community, spanning the whole system and containing almost all the sensors. The remaining four communities detected are extremely small and spatially scattered. This means that the trajectory of a random walker on $C^f$ is very hard to compress and does not show modules with high persistence times, i.e. dense network structures. In addition, there is no hierarchy in the network partitions.

A very similar result is obtained from the optimisation of spatial modularity (Eq. (2.22)). We find also in this case a very large community that spans the whole system (about 65% of the sensors). In contrast to the map equation results, there are 77 smaller partitions. Incidentally, if communities are ordered by descending size and their position in this ranking denoted by $r$ (Figure 3.18, bottom), one finds that the community size $s$ scales approximately like $s \propto r^{-\nu}$ with $\nu = 1.09 \pm 0.01$ (adj-$R^2 = 0.97$). The top panel of Figure 3.18 shows how the communities are distributed in space. The red dots covering the whole network correspond to the largest community, while the other colours correspond to the smaller communities. This visualisation is confusing however and does not allow to identify the details of the communities. While some of the smaller communities appear to identify recognisable elements of the street network, others instead are completely spatially spread out. For an example, the panels of Figure 3.19 show some of the largest communities (large dots) against the background of the largest one (small dots). The second largest does not show any recognisable feature and is spread all over the catchment area (top left). The third largest (top right) instead appears to identify two different branches of an important corridor in London, the north-south path that leads from Vaux-

\[8\]Some colours for the smaller communities had to be reused.
Figure 3.18.: Results for the spatial modularity community detection on the correlation network $C_f^S$. $Q^{Spa}$ is optimised by a partition containing 78 communities. The largest community covers the whole catchment area and is followed by a series of communities with power law distributed sizes that do not provide much functional information (see main text and Fig. 3.19).
3.5. Power-law correlation length and power spectra

All the community detection methods produced noisy, overlapping and ill-defined communities on London's street network. The Girvan-Newman modularity identified two large communities, roughly corresponding to the central and peripheral area of the city, while the map equation returned one spanning community, implying that there is no simple structure in the network defined by the correlation matrix. The spatial modularity produced a very similar result in addition to a large number of smaller communities (whose size scales as a power law). It also provides a key for the
Recall that spatial modularity is a way of quantifying the effect of spatial proximity and account for it in the community detection in order to cluster together nodes that are more similar or interacting that it would be expected at a certain distance. In the spatial null model, the fundamental element is the deterrence function $f$, defined in eq. (2.22). In general, $f$ measures the expected strength of a link between nodes at a given distance. In our case, the links are (same time) correlations and therefore the deterrence function is akin to the two-point correlation function at zero delay ($\tau = 0$). For the expected correlation $C(r, \tau)$ we find two regimes separated by a characteristic distance between sensors $r_0 \sim 200\text{m}$ (Figure 3.21). The case $r < r_0$ typically corresponds to sensors on the same link and the coarse-graining of our data does not allow us to identify the functional shape conclusively. Nonetheless there appears to be steep decay. For $r \geq r_0$ instead, $C(r, \tau = 0)$ displays a clear power-law decay $r^{-\xi}$ with $\xi = -0.26 \pm 0.01$ (adj-$R^2 = 0.98$) up to about 12 km. Note that in the case of delayed correlations (see Fig.3.21 for $C(r, \tau = 30\text{min})$), the functional form remains a power law for $r > r_0$ for delays up to a two hours, when the system starts decorrelating (not shown).

These results suggest the presence of two mechanisms behind these two regimes: one that is responsible for the internal unidimensional link dynamics, where correlations decrease rapidly (possibly exponentially) in space, and a second that emerges at the network level, as the outcome of the different interacting flows.

To test this hypothesis, consider the following spatial network model: place each sensor on the plane at its coordinates and consider it as a node; then link all the pairs of nodes $(i,j)$ such that their spatial distance $d_{ij}$ is smaller than a certain distance $d_0$:\footnote{The distances we are using here are geodesic calculated from the sensors’ longitudes and latitudes. In Appendix B.1.2 we justify this choice and show that statistically there is a linear proportionality between the travel time—or driving distance—between two sensors and their geodesic distance.}

\begin{equation}
\begin{aligned}
p(d_{ij}, d_0) &= \begin{cases} 
1 & \text{if } d < d_0, \\
0 & \text{if } d > d_0.
\end{cases}
\end{aligned}
\end{equation}
It is clear that varying $d_0$ between 0 and $\max(d_{ij})$, one passes from the empty network to the completely connected graph (Figure 3.20). In the middle, there must exist a value of $d_0$ for which most of the network becomes connected. In graph theory, this phenomenon is usually referred to as the birth of the largest connected component (LCC) [51, 212]. Strictly, the LCC is defined as the component whose size is a finite fraction of the total network in the limit of network size tending to infinite ($N \to \infty$). For example, in the Erdős-Rényi graph model, one can prove that the appearance of the LCC at a certain linking threshold probability $p_c = 1/N$ occurs as a second-order phase transition, akin to percolation [213]. However, we are interested in a finite, albeit large, system and, although these concepts are rigorously defined only in the thermodynamic limit, we can so still exploit the concept of LCC in a finite size network to test the hypothesis regarding the two regimes. If at $r_0$ the sensor network starts to develop a large component, it means that for $r < r_0$ we are probing only local dynamics, pertaining to a small disconnected components (e.g. a street and its related junctions), while for $r > r_0$ we are studying the integrated dynamics.
3.5. POWER-LAW CORRELATION LENGTH AND POWER SPECTRA

taking place at the network level. In Figure 3.22 we show that this is the case: the LCC undergoes a rapid growth at a value of the linking distance $d_0 \sim 100\text{m}$ very close to the value at which the correlation function becomes algebraic. Therefore, the power law correlations are a feature of the network traffic, that is lost when we consider single road segment or similarly sized areas of the city.

The effects are far reaching. Long correlations of these sorts usually emerge in systems with long-range interactions or systems with local interactions in the vicinity of a critical point [213], that is near a (second-order) phase transition, where the correlation length diverges. When this happens, the whole system is coupled as one dynamical module disregarding of the underlying topological structure or the details of the interaction. The large spanning community found in the community detection (sec. 3.4.1) is then just the representation of this phenomenon. What is less clear, however, is how the large correlation length comes to be.

Let us consider for a moment that London’s street network is, in some sense to be specified, close to a critical point. This concept of phase transition and criticality for traffic is not new. A large number of works in traffic flow theory investigated the properties of the transition from free flow to synchronised flow to congested flow on road segments [152, 214], with particular attention to highways. The focus on highways and main arterial roads was motivated by the fact that these models resemble closely hydrodynamical models based on fluid density. In order for the continuous approximation to stand, sizeable lengths of road were needed. The same holds for most of the phenomena predicted (and observed) in traffic flow, e.g. stop and go waves, wide moving jams, jam fronts propagation [157].

In a dense urban environment, on the contrary, the situation is more complicated: due to the restricted spaces and richer infrastructure, the endogenous flow dynamics does not dominate anymore, the concept of continuous flow density breaks down and most of the dynamics is in the queues, the network topology and the traffic control. Nagel [215, 216] proposed that the interaction between the traffic and some traffic control strategies might result into a self-organised critical state [217]. This state would maximise throughput, but also minimise predictability. Indeed it would be characterised by power-law spatial correlations, fluctuations on all scales and
Figure 3.21.: Plot of the spatial correlation function $C(r, \tau)$ as function of the spatial distance $r$ for delay $\tau = 0$ (blue) and $\tau = 30$ minutes (gray). The slope is $\beta = -0.26 \pm 0.01$ implying a very slow decay of correlations over a large interval of distances (<200 meters). The correlations for $\tau = 30$ minutes have been lowered for visibility, since the two curves showed remarkable overlap.
3.5. POWER-LAW CORRELATION LENGTH AND POWER SPECTRA 70

Figure 3.22.: Top, the size of the largest connected component of the spatial network defined by eq. (3.15) as a function of the linking distance $d_0$. Bottom, $C(r, \tau = 0)$ again as a function of the distance. Note that the size largest connected component starts to increase steeply at the same value of distance for which the correlation function start displaying a power law dependence.
usually long memory effects, encoded in the $1/f$ spectra of the observables in the system. Finally, it should be self-organised in the sense that no external tuning is tuning the control parameter to the critical point. In Nagel’s case, the control parameter was the network density and the order parameter the individual travel time \[ \text{[215]} \]. At the point of maximum flow in the Fundamental Diagram, the travel time started increasing and a divergence of the error on the estimated travel time for individual cars was observed. These divergences (technically of the derivative of the continuous order parameter at the critical point) are a general feature of critical systems and result in fluctuations of all amplitudes on all scales near and at the critical point. Due to the nature of our dataset, it is not possible to measure the individual travel times. Despite that, we can probe long-memory effects in the network looking at the spectra of the occupancy and flow timeseries. Figure 3.23 shows the power spectra for the detrended occupancy and flow data, averaged over the sensors. Both show a peak, corresponding to a daily
period ($10^{-5}$ Hz), and an underlying clear power law dependency with exponents between $-0.9$ and $-1$.

Scaling in power spectra of traffic flow was observed in a number of different traffic situations [156, 218, 219] and related to the existence of a self-organised maximum through-put state for unidimensional flows on long streets (with infinite length or periodic boundaries), where jams of all sizes appear producing a fractal landscape of densities. We see here that this feature survives also in urban flow, despite the finiteness of urban road segments and the presence of network infrastructure and traffic control.

More importantly, putting together the observed power law correlations with the sensors’ spectra, we have a very strong case to argue that London, as a large spatial interacting system, is a self-organised critical system.

### 3.5.1. Origin of Criticality

The traditional way to go about checking whether the system at hand is really critical would be to try to calculate the critical exponents, so as to verify to which –if any– universality class it belongs. However, due to the poor resolution of the data and the disordered distribution of the sensors, it is very hard to devise a reliable renormalisation scheme (see for example [208, 220]) to measure finite size scaling (FSS) on networks [221]. One way to approach this question is to avoid completely the (FSS) and look at the Fluctuation Scaling (FS), or Taylor’s law [125, 141], that was introduced in section 2.3. FS states that the fluctuations of a positive extensive quantity are well approximated by a power law of the average value the quantity itself.

Fluctuation scaling can be defined for *ensemble* fluctuations (EFS). Consider to have a system composed by many elements, indexed by $i$. Each element has some quantity associated to itself. For definiteness, in our case the elements will be the sensors and the quantity the corresponding flow ($i, f_i(t)$). Let us imagine further that each element can be grouped according to some other quantity, usually referred to as the *size* $S$ of the element due to the seminal paper by Taylor [125] which treated the crop yield of fields of different sizes. In our case, we can associate to each sensor $i$ its average occupancy $<o_i>$. We denote then the ensemble average over
sensors with similar occupancy \( o \) as \[141\]

\[
\bar{f}_o = \frac{1}{M_o} \sum_{\forall i: \langle o_i \rangle = o} f_i
\]  
(3.16)

where \( M_o \) is the number of sensors with average occupancy equal to \( o \). The variation inside each class is then given simply by:

\[
\bar{\sigma}_o^2 = [\bar{f}_o]^2 - […]^2
\]  
(3.17)

and finally FS is written as

\[
\bar{\sigma}_o \propto \bar{f}_o^{\alpha_E}
\]  
(3.18)

The ensemble fluctuations exponent \( \alpha_E \) is usually found in the interval \( 1/2 < \alpha_E < 1 \). For the Freeflow data, \( \alpha_E = 0.87 \pm 0.02 \) (adj-R\(^2\)=0.95) (see Figure 3.24).

A value of \( \alpha_E \) in the middle between the two extreme values 1/2 and 1, corresponding to fluctuations dominated by random internal dynamics and by external forcing, it implies that the network traffic dynamic is not dominated by either, but rather emerges from the inextricable interaction
between the two. More importantly, it provides us with new insights into
the nature of the criticality we observe in the dataset. Indeed, it is well
known [148, 149, 222] that for complex systems whose components dis-
play long-range correlations of the form $C(r) \sim r^{2H_V-2}$, one has that $H_V = \alpha$ [126]. For the correlations, we measured the exponent $\xi = 0.26 \pm 0.01$,
hence a predicted Hurst exponent $H_V = 0.87 \pm 0.03$, which is compatible
with the scaling exponent.

The relations just used were devised in the context of ecological mod-
els, in particular the Sataka-Iwase model of trees masting [222] after the
observation of large scale synchronisation of forests across the world (up
to thousands of kilometres of distance). In this framework $\alpha = 1$ would
correspond to perfect synchronisation and $\alpha = 1/2$ to random fluctuations.
For London the results indicate the presence of strong, yet partial spatial
(remember we used the ensemble fluctuations) synchronisation that spans
the whole system and is the driving mechanism behind the scale-free fluc-
tuations observed through the power spectra and the correlation function.
3.6. Summary of Chapter

This chapter investigated the features of urban traffic in the Freeflow London data and framed them in the context of the current discussion about the existence of a Macroscopic Fundamental Diagram. London displays a large heterogeneity in space and time as described in section 3.2. We complemented previous results on the effects of spatial distributions of traffic and showed that, under mild functional assumptions, an extension of the MFD to include the rescaled occupancy heterogeneity \( \Sigma O \) provides a good (and testable) description of the observations (section 3.3.2). The extended MFD defines the surface of allowed traffic configurations, but it does not give any prescription about the evolution of the network state on it.

With reference to the motivating questions of this chapter, our results imply that the heterogeneity of the traffic distribution plays a fundamental role in determining the Macroscopic Fundamental Diagram of an urban network. The same network density can give rise to very different network flows (and thus trip completion rates) if traffic is spread uniformly or not. Therefore, the MFD has to be a function of the first two moments of the density/occupancy distribution.

This also implies also that network flow break down can happen in two ways. The first, similar to the classical single-link FD, takes place when the network occupancy exceeds a certain threshold and the flow starts decreasing. The second can happen at any network density and is driven by the heterogeneity. In other words, even for low network densities (e.g. 10% occupancy), a large heterogeneity of traffic distribution can lead significant flow break down.

London also displays a daily anticlockwise hysteresis loop, characterised by a low-flow, high-disorder branch during the loading period (morning rush hour) and a high-flow, low-disorder branch during the unloading period (late afternoon and night) (section 3.3.3). In the light of recent work on urban hysteresis, we argued that London’s complex structure makes it likely to develop congested areas at lower occupancies than otherwise expectable and that effective reservoir definitions would be needed to counter
the insurgence of hysteresis.

Our attempt at finding a robust data-driven definition for such reservoirs through the study of the correlations in the traffic was frustrated by the surprising scale-free character of the traffic flow fluctuations, which was reflected by the community detection results (section 3.4). In particular, we found that the two meaningful level of descriptions for London are the single road and the whole urban area. This is a consequence of the algebraic decay of the correlation between flows at different points. That is, London behaves like a single unity, an entangled complex network.

At a closer look, the spectral and correlation properties revealed the typical signs of self-organised criticality, which together with the analysis of the fluctuation scaling, offers a potential explanation for the evolution on the MFD surface. During the early morning the traffic is mainly directed from the outskirts toward the centre of London and is therefore characterised by low measured occupancies and large heterogeneities. In turn, this forces the network state to move along the low flow branch of the observed hysteresis loop. As the densities increase during the day, the flows on different links begin to interact through spill-over effects, reactive control and queueing. These interactions are the most likely cause behind the appearance of the observed partial spatial synchronisation, which in turn smoothes the traffic density landscape and results in the sharp increase of the flow at the top of the hysteresis loop. During the late afternoon, most traffic is directed towards the outskirts. This should reproduce a similar scenario as the early morning. Instead, the spatial synchronisation developed previously, together with the long memory effects showed by the $1/f$ spectra, allows the system to maintain a smaller heterogeneity and so to unload along the higher flow branch.

In brief, as the occupancy increases during the day, partial spatial synchronisation develops, producing the observed long-range correlations, smoothing the traffic disorder and therefore improving the network flow. Once the network starts unloading, synchronisation is maintained while the system unloads along the higher flow branch of the hysteresis loop on the MFD surface.
In the light of the considerations above, it is of great importance to understand and devise control and routing schemes to smooth the density landscape. However, due to the coarse graining of the dataset (aggregated over 15 minutes), it is not clear in our analysis whether and how much traffic control actually affects this dynamics. It would indeed be interesting to test this with data recorded under different traffic regimes. A second way of testing this hypothesis would be to look for leftover pockets of synchronised traffic at night, when the density is lower, which might act as seeds for the network-wide one during the following day. Unfortunately, the current dataset is too noisy and short to obtain any conclusive evidence at this stage. However, datasets of 1-2 years of length should make these effects visible and such datasets are within Freeflow’s reach (even though not this thesis’).
4. Models of Information Propagation

Traffic signals in New York are just rough guidelines.

David Letterman

4.1. Background and Aim

In the last few years, improvements in information and communication technology have dramatically increased our capability to gather and disseminate data of all kinds. Access to this new deluge of information raises the question of whether, how and by whom this information should be used. These questions are becoming increasingly important for traffic networks where the availability of real-time information is crucial for traffic planning, management, and the control of congestion. For example, traffic prediction models are used every day to devise traffic light schedules (e.g. the SCOOT or GLIDE systems [223]). These optimisation procedures not only depend on large amounts of computational power but also crucially depend on the availability of reliable real time data. To make the problem harder, in the next few years new sources of data will need to be integrated in the traffic control scheme, e.g. CCTV feeds, pollution data, crowd and cycles positions. The above-mentioned, classical methods for managing traffic information are highly centralised: the information gathered from the network is used to obtain an optimal set of instructions. It is evident that such systems cannot be robust, as any computational failure at the central level or data acquisition error at the peripheral level could ultimately influence the set of instructions sent to the traffic controllers. More importantly, as the amount of data increases, centralised methods
might soon become unsustainable.

In addition to the technical challenge, another factor is coming into play. Recent advances in small, mobile technology mean that now soon enough an army of sensing devices will be on the roads. Already now for example Tom Tom collects traffic information through the mobile phone data of its users. This can be a boon for the data-mining scientist, but one can easily see how this technology might wrestle the control of information from the hands of the network managers. Imagine, for example, populations of travellers receiving route recommendations from independent providers (non network managers) obtained via crowd-sourcing, or vehicles carrying and exchanging information. Some of these applications already exist and promise to be advantageous, like in the case of AAC on highways \[224\], but it is not clear how similar technologies would influence urban traffic patterns, especially once out of the hands of the traffic managers. It is therefore crucial to develop an understanding of what are the effects of information per se, not just as a tool.

These themes are not new to the traffic literature. Starting with Arnott et al. \[48\], the literature on Advanced Traffic Information System (ATIS) and dynamic traffic assignment \[31\]-\[36\] \[225\]-\[227\] is extensive, ranging from assignment with recurrent congestion \[228\] to modelling of route choice behaviour \[229\] \[230\]. In particular, the availability of real-time information prompted ATIS to develop traffic assignments schemes which included account also en-route travel information \[31\]-\[37\]. These schemes in turn relied on the capacity to estimate the effects of real and perceived information uncertainty \[231\] \[232\] its cognitive cost \[233\], inequality among drivers \[234\], user response to the travel information \[235\]. Although their accuracy and effects on route behaviour have been questioned \[38\]-\[41\], ATIS are generally considered to improve network performances \[42\] and the dominant paradigm assumes that (a proper use of) complete information will reduce congestion and lower travel times. This idea, however, has not been proven to be general, but rather disproven in some cases, for example in the case of over-reaction leading to herding effects that can cause congestion to be simply shifted to a different area instead of relieved \[43\]-\[45\], reminding of Braess’ paradox \[48\] \[230\] \[236\] \[237\].
More recently, attention shifted to the information feedback effects due to ATIS [238–241] and the most promising (non equilibrium assignment) routing methods based on decentralised collection and projection of real-time information [242–244], usually inspired by swarm or ant colony methods [245]. To illustrate this, consider the recent paper by Claes et al. [242]: they showed that an elegant system based on the cooperation between road infrastructure and vehicles through "digital ants" is able to outperform routing strategies based only on historical data (basically weighted shortest path) or regular, yet infrequent radio updates (TMC [246]).

All of these methods focus on devising the best strategy to collect traffic information and forecast traffic conditions in order to minimise travel times. The underlying assumption is that complete information will guarantee the best network performance.

The perspective on information and network dynamics radically shifts when one looks at the physics network literature. There, most approaches to network traffic are motivated by navigation or congestion problems on large scale-free networks, the prominent example being the Internet. Other common problems relate to epidemics and gossiping in social networks, search in large databases, resilience problems of infrastructure networks. In most –if not all– these models, one tries to optimise network performance based solely on local information, that is information that is available to a packet sitting on one router or that you could ask your friends (your social network neighbours).

The reason is twofold: ignorance and resources. Ignorance reflects the fact that in most cases the networks over which the diffusion takes place are not known or completely mapped. There exists no map of the Internet or of the WWW. The same holds for social networks. Yet, the Internet works extremely well and Milgram’s experiment showed that social networks can be efficiently navigated on the basis of local information alone [247, 248]. This conclusion might seem trivial, since it is well known that most real-world networks are small world networks: that is, the maximum distance between any two nodes scales as $\log N$ where $N$ is the number of nodes in the network. However, the fact that short paths exist does not mean that they are easy to find.
Moreover, local information is economical in terms of memory, storage and computation. If packets on the Internet were requested to bring with themselves a record of their path—say, for navigation purposes—they would grow quickly very large due to the extra bits needed for coding that information with the result that, for example, online streaming would become unfeasable.

The literature is extensive. A good review on simple particle-hopping models is given in Evans et al. [249], dynamical processes on network are fully reviewed in Barrat’s book [168]. Among the many studies, Kleinberg [250] elucidated the principles of navigation on small-worlds, Tadic et al. [251, 252], Fronczak et al. [253, 254] studied routing based on local information. They considered routing rules with different degrees of congestion awareness, ranging from random diffusion to rigid congestion-gradient driven flow. They found that the strictly congestion-gradient driven routing easily leads to jamming, while introducing elements of randomness usually guaranteed a degree of flexibility resulting in alleviation of congestion. Carmi et al. [255] presented a physical solution of the problem of effective routing with minimal memory resources, while Toroczkai and Bassler [256] investigated the influence of network architectures on the congestion, uncovering that homogeneous networks are more resilient to congestion, while scale-free networks appear to be optimised for transmission. Sinatra et al. [257] concluded a dispute about the possibility of defining maximally entropy-producing walker based on local information only. Rosvall and Sneppen [166] discussed how to use limited information to find the optimal routes in a network, with regard to search problems on networks [258] and to the informational description of cities [259, 260].

Another important difference with respect to DTA models is that most of these systems are cast as dynamical non-equilibrium problems. Particles come and go, there is no fixed demand, but rather a driving on the system. The performances of a routing or dissemination strategy are therefore not obtained through an equilibrium assignment (although some simple zero-range models admit stationary distributions [190]) and therefore are usually measured directly in simulations through network wide metrics. Congestion for example is usually identified with a positive first derivative of the network density (averaged over small time intervals) and its emergence
associated with a phase transition, from the stationary (free) phase to the congested one. In this perspective, the features of such transition are crucial for the understanding of network dynamics, where by features usually it is meant how smooth is the transition to a congested state, or in the language of phase transitions, first-order or second-order.

The limitations of these studies is that they did not address the issues of explicit navigation AND congestion at the same time, due to mathematical difficulties in some occasions, and to lack of interest in others.

A few works \cite{238,239} introduced density feedback mechanisms to reduce oscillations in a two-route scenario, but had no explicit discussion of the role of information.

Of great interest for this discussion is the paper by Scellato et al. \cite{47}. They showed that a simple model of routing based on local dynamical information is able to outperform routing strategies based on complete congestion information. In particular, they simulated various urban networks, where each link was equipped with a Nagel-Schreckenberg automata. A penalty function that depended on the distance from the destination and the congestion state was the basis for the navigation. They compared the cases in which congestion information only regarding the next link was available and the case in which travellers had complete, global information. They found that the difference in performances were not appreciable for low traffic loads and that local information was preferable for high traffic loads. Although their model had limitations (for example, the total network load was stationary, i.e. the system was open), their result raise the question of whether local information constitutes an optimum or whether an optimal information length scale between the local/global extremes exists.

Similar results were also obtained for routing on complex networks, usually with reference to packet transmission on scale-free networks or transport in supply networks\cite{263,266}.

Here we address this conundrum, asking how much information is advantageous and under what traffic conditions.

---

1 The different perspective in the two literatures is due to the fact that the physics literature mostly focused on large communication and social networks \cite{166,261,262}, where it usually is hard to obtain complete information about the network topology and therefore one must resort to local information or strategies that do not need the infrastructure's help.
For this reason, we will analyse a few models of traffic networks where the traffic dynamics is coupled to the traffic information via local and non-local dissemination strategies, often recurring to self-organised processes. A little care is needed here: for local here we mean information propagation is limited to short distances, typically one link in a network, the road segment until the following decision point in a road network. In this sense, digital ants crawling ahead of a vehicle and then reporting back are non-local. This caveat is necessary to avoid confusion, since all decentralised methods are by definition local, in the sense that the interaction usually happens only with the surrounding environment.

The chapter begins by studying an extremely simple model of flow on artificial networks (section 4.2), where we focus on the comparison between local and global control strategies. This model already gives insights into the differences between the effects of the two regimes and shows that a local information regime generates smaller fluctuations in the number of congested nodes. Building on these results, in section 4.3 we relax the local-only or global-only conditions on the information regimes and ask whether it is possible to let the system itself devise an optimal length-scale for the information dissemination. In section 4.4 we implement a mechanism of self-organisation on realistic urban maps and discover that usually it does not perform better than a strictly local information regime. The reason is that such a mechanism, under low loads, produces only local propagation, while, under large loads, it produces system-wide informational cascades, that block the whole system, thus decreasing its performances. A natural step forward is to allow the propagation of information in space, but only in a limited way. This makes the traffic network more resilient to a larger inflow of travellers. Finally, we shift the focus from the infrastructure alone, and consider a model of information exchange among agents on the urban network, reminding the models of [242–244] where the traffic infrastructure assumes the role of the information broker rather than of the supplier. We find a striking similarity with the previous cases regarding the relation between the emergence of congestion and the information regimes and highlight the importance of information ageing in preventing the development of congestion. Finally, making use of the results of the previous chapter, we explain the results linking the information dissemi-
4.2. MINIMAL FLOW MODEL

4.2.1. Description

As a first investigation in this context, we study a minimal network flow model where the nodes represent junctions on a traffic network or stations in a rail network. Nodes react to their environment under two different regimes of information. Under the first regime, local information dissemination, each node is only aware of the state (e.g. congestion, flow, occupancy) of its immediate neighbours. Under the second, global dissemination, nodes are aware of the situation of the whole network and react accordingly [267].

The information that is available in this model is the number of congested nodes. In the local dissemination case, this means that a node is aware of how many of its immediate neighbours are congested and uses this information to modulate its own outgoing flow. In the global case, instead nodes have access at each time to the total number of congested nodes present in the whole system. Let us call \( J_i(t) \in [0,1] \) the time-dependent fraction of congested nodes of which node \( i \) has knowledge. In the local case, this is given by the number of congested neighbours of \( i \), \( n_{i\text{ cong}}(t) \), divided by the number of \( i \)'s neighbours, that is \( i \)'s degree \( k_i \), so \( J_i(t) = \frac{n_{i\text{ cong}}(t)}{k_i} \). In the global case, every node has full knowledge, \( J_i(t) = \frac{1}{N} \sum_i n_{i\text{ cong}}(t) = J(t) \), and so \( J \) is independent of the node.

In contrast with communication networks, where the maximum number of packets on a server (the queue length) can be very large (effectively infinite), we assume nodes to be characterised by a finite capacity threshold, \( T_i \) for node \( i \), representing the value of the load \( L_i \) (people, trains, cars...) at
which the node is considered congested. We denote by \( S_i \in \{0, 1\} \) the state of a node, respectively when the node is free or congested. Congested nodes (that is \( L_i > T_i \)) do not accept incoming flow from neighbouring nodes until their load decreases under their capacity threshold. This is done in order to mimic the behaviour of real systems, where stations can block due to too many incoming passengers, effectively cutting off their neighbours. The state of node \( i \) at time \( t + 1 \) depends on its load at the preceding time-step, so one has \( S_i(t + 1) = \Theta(T_i - L_i(t)) \), where \( \Theta \) is the Heaviside step function. In terms of nodes state, for the local dissemination we can rewrite \( J \) as:

\[
J = J_i^{loc}(t) = \frac{1}{k_i} \sum_{j \in \Gamma_i} \left(1 - S_j(t)\right),
\]

where the sum is restricted to the neighbourhood \( \Gamma_i \) of node \( i \), while in the global dissemination each node has information about the whole network, therefore \( J \) takes the form:

\[
J_i^{glob}(t) = \frac{1}{N} \sum_{m=0}^{N} \left(1 - S_m(t)\right).
\]

To calculate the flow between two neighbouring nodes, we assume that at each time nodes divide equally and send out their load to the neighbours. Hence, the outgoing flow \( F_{ij}^{out} \) from node \( i \) to node \( j \) is given by

\[
F_{ij}^{glob/loc}(t) = \min\left(\frac{L_i(t)}{k_i}, T_i\right) S_j(t) \left(1 - J_i^{glob/loc}(t)\right)
\]

where the \( \min \) function limits the maximal total outgoing flow from a node at given time to \( L_i \). Note that, for growing \( J \), the outgoing flow from nodes decreases in general. Also, the term \( S_j \) on the r.h.s. of (4.3) accounts for the impossibility of sending flow to an already congested node (\( F_{ij} = 0 \) if \( S_j = 0 \)). The idea here is that a node, if informed that many others are jammed, will gradually reduce its outgoing flow to avoid congesting the system any further. At the same time, no flow is sent to congested nodes in order to relieve them.
4.2. MINIMAL FLOW MODEL

4.2.2. Model setup

Initial Conditions The system is driven with two different mechanisms. In the first one, we introduce in the network a total load, \( L_{\text{tot}} = \beta \sum_{i=0}^{N} T_i = \beta T_{\text{tot}} \), controlled by the parameter \( \beta \), that represents the filled fraction of the total network capacity. This load is distributed in the network randomly, the only condition being avoiding to jam nodes at the beginning of the simulation. Operatively, one can accomplish this by assigning to node \( i \) a load equal to a random fraction of the node’s threshold and then iterating over the nodes until the load introduced in the network reaches the desired value. The initial conditions are thus:

\[
S_i(0) = 1 \quad \forall i \\
L_i(0) = X_i, \quad 0 < X_i < T_i \quad \forall i
\]  

(4.4)

where \( X_i \) a random variable \( \in [0, T_i] \), so that \( \sum_i L_i = \beta T_{\text{tot}} \). Because of the shape of Eq. (4.3), the state without congested nodes is an absorbing state. Therefore, to initialise the dynamics, a single congested node is needed. To obtain it, we select pairs of nodes and transfer a random fraction of the load of the first on the second and repeat this process until one congested node appears. At that moment, the redistribution of load is stopped and the system evolves autonomously until there are no more congested nodes. Then, we restart the redistribution until a new period of activity is triggered.

In the second case, the entire load is placed on a single node, denoted as the seed, from which the load diffuses into the network, without the need of any further driving. So the initial conditions are:

\[
S_i(0) = 1, \quad L_i(0) = 0 \quad \forall i \neq \text{seed} \\
S_{\text{seed}} = 0, \quad L_{\text{seed}} = L_{\text{tot}}
\]  

(4.5)

The rationale behind the first method is testing the response of the system to a small perturbation in a typical load configuration (load distributed around the network). The second instead tests the response to a configuration in which most of the load is concentrated in one location and relaxes through the system.
Dynamics  The parameter $J$ in (4.3) represents the fraction of jammed nodes that a station sees in the system. In the local dynamics, each station has information only about its nearest neighbours, thus $J$ is node-dependent, $J = J_{i}^{\text{loc}}(t)$ restricted to the neighbourhood of node $i$, while in the global dynamics each station has information about the whole network, therefore producing an unique value for $J$.

We note that the completely jammed and unjammed states, $N_{J} = N$ and $N_{J} = 0$, are absorbing states. Indeed, if the system reaches the state $N_{J} = N$ we have $J_{i}^{\text{loc}} = J_{i}^{\text{glob}} = 1 \forall i$, implying $F_{ij} = 0 \forall i,j$. Similarly, if $N_{J} = 0$, consider the load $L_{i}(t_{0})$ on node $i$ at time $t_{0}$. Since all the neighbours of $i$ are unjammed, the total outgoing flow from $i$ is $\sum_{j \in \Gamma_{i}} F_{ij} = L_{i}(t_{0})$ while the maximum incoming flow is $\sum_{j \in \Gamma_{i}} F_{ji} = k_{i} \alpha = T_{i}$. Thus,

$$L_{i}(t_{0} + 1) = L(t_{0}) - \sum_{j \in \Gamma_{i}} F_{ij} + \sum_{j \in \Gamma_{i}} F_{ji} \leq k_{i} \Omega = T_{i} \tag{4.6}$$

and node $i$ does not jam, since the condition for jamming is $L_{i} > T_{i}$.

We have seen that starting from $N_{J} = 0$ the system cannot in itself trigger jamming perturbations. However, this state is very unstable because, in presence of jammed neighbours, nodes keep some fraction of their previous load on themselves. This will move them closer to their jamming threshold, making them more susceptible to jamming. So if we produce in some way even a single jammed node, we expect the jamming perturbation to expand through a sort of chain reaction up to a stationary state that depends on the amount of load introduced. This is a first indication that inhomogeneity in the system is the driving force behind jam propagation.

Simulations  We performed extensive simulations on Erdős-Rényi random graphs (RG) and scale-free networks (SF) with the number of nodes $N$ varying between $10^{2}$ and $5 \times 10^{4}$ under both driving mechanisms. Realisations initialised as in (4.4) were driven by redistributing a fraction of a randomly chosen node’s load to another randomly chosen node ($L_{i} \rightarrow L_{i} + cL_{j}, L_{j} \rightarrow L_{j}(1-c), c \in (0,1)$). When a jam appeared, the driving was suspended for the duration of the active phase ($N_{J} \neq 0$). Realisations initialised as in Eq. (4.5) were allowed to relax to their stationary state without active driving.
4.2. MINIMAL FLOW MODEL

4.2.3. Results

Figure 4.1: $N_j$ vs $\beta$ plot for global and local models on SF (full line) and RG (dashed line) networks ($N = 1000$) with seeded and uniform (inset) initial conditions. Error bars are standard deviations of fluctuations of $N_j$ around its mean value, averaged over the realizations at a fixed $\beta$. Data shown are obtained over 100 realizations of the RG and SF networks.

Figure 4.1 shows the characteristic behaviours observed for the asymptotic jammed population $N_j(t \to \infty)$. The main features are:

1. there is no asymptotic jammed populations for a large range of $\beta < \beta_c$, where $\beta_c$ depends on the driving and topology;

2. at $\beta_c$ the system undergoes a first-order transition to a state with a stationary jammed population;

3. the system then has a plateau, showing little or no response to increases in $\beta$ until a second transition to complete saturation;

4. the fluctuations around the mean stationary value are much larger and broadly distributed under the global dynamics then the local one (see Fig. 4.4).
Let us now go through these results in detail. For $\beta < 0.75$, there are no congested nodes in the stationary state, meaning that the load in the system is efficiently redistributed among the nodes. For higher values of $\beta$, $N_J$ starts growing. The exact values of $\beta$ at which $N_J$ becomes different from zero depend on the topology and the information regime. Once a stationary jammed population is present, one would naïvely imagine $N_J \not\to N$ as $\beta \to 1$. Instead, we observe that $N_J$ remains in the vicinity of $N/2$ for a broad range of $\beta$ values. Indeed, the applied load is sufficient to jam the entire network. However, most of the load is trapped in the congested nodes and is redirected from there. This prevents a uniform redistribution of the load across the network keeping $N_J$ far below its maximal possible value.

As can be seen from Fig. 4.1, the local dissemination regime develops a non-vanishing population of congested nodes for $\beta \geq 0.76$ in scale-free networks and $\beta \geq 0.83$ in random graphs. For the global dissemination we find larger threshold values, $\beta \geq 0.8$ on scale-free networks and $\beta \geq 0.86$ on random graphs. It appears then that the global dissemination mechanisms allow the network to sustain a larger load before developing congestion in both topologies. Hence, from this point of view, global information is preferable.

The fact that scale-free networks develop congestions for lower loads as compared to random graphs is expected (see [46] and citing articles): due to the power law degree distributions, hubs are present, i.e. nodes with very large connectivity; their large degree increases the likelihood that they will receive large amount of load or have a congested neighbour and will therefore start to accumulate load on themselves, eventually propagating the jam. This must be compared with the homogeneity of random graphs, where there all nodes are have approximately the same degree and therefore any diffusing quantity tends to be uniformly distributed.

In addition, we observe that $N_J(t)$ under the seeded driving (Fig. 4.2) increases through a sequence of quasi-plateaus and jumps. These are the signature of an oscillating core mechanism (Fig. 4.3): the initial congested seed sends flow toward its neighbours, while not accepting any from them; this brings the neighbours closer to their capacity threshold, making them more vulnerable to jamming. If there is enough load on the seed node, it will eventually jam its own neighbourhood (Fig. 4.3a). When the neigh-
4.2. MINIMAL FLOW MODEL

Figure 4.2.: $N_J$ as a function of time for global (green) and local (black) models in a RG network with $\beta = 0.9$, averaged over realisations.

Figure 4.3.: Jammed component expansion mechanism.

bours relax onto their non-congested neighbours (the seed’s second neighbours), this makes the second neighbours vulnerable or even jam (Fig. 4.3b); in this fashion, the congestion can propagate as long as the flow is not distributed homogeneously. The growth of the core ends then in two scenarios:

a) the seed node’s neighbours manage to distribute the load without making their own neighbours congested and so the load is distributed freely outward (Fig. 4.3d), or

b) the second neighbours jam and while they relax, they jam the seed’s first neighbours again thus creating a bigger, stable jammed core (Fig. 4.3c). The quasi-plateaus in $N_J(t)$ correspond to periods of growth of
the vulnerable population, and are followed by sharp jumps, where
the jammed core grows by rapidly invading the vulnerable nodes.

Another interesting feature is that the local model exhibits much narrower
fluctuations ($\sigma_l \approx 30$ for $N = 10^3$) than the global model ($\sigma_g \approx 150$). These
fluctuations are another signature of the core expansion mechanism: when
the systems reaches its stationary $N_f$ value, it keeps fluctuating around
the value due to the repeated overshooting between congested and uncon-
gested nodes. The more violent fluctuations observed under the global dy-
namics are indeed consistent with the core mechanism. Indeed the global
dynamics, through its global flow suppression, produces smaller "packets"
of flow across the network that give rise to a slower and more uniform dis-
tribution of the load. This, on the one hand, allows the system to avoid
the emergence of congested nodes for larger $\beta$s as compared to the local
dissemination regime. On the other hand, it also makes nodes able to stay
closer to their thresholds than under the local dynamics, because of the
smaller flow packets, therefore generating a larger population of vulnera-
ble nodes, which are the ones responsible for the large fluctuations in $N_f$.

4.2.4. Rate Model

Although the model is deterministic, the fluctuations due to spatial in-
homogeneity of the load distribution suggest that it might be possible to
describe the observed behaviour by making use of a set of transition pro-
cesses. We separate the population of nodes in to three types: unjammed
(U), vulnerable (V) and jammed (J), with rate equations,

\begin{align}
U_n(t) + J_m(t) \xrightarrow{\alpha} V_n(t) + J_m(t) \quad (4.7) \\
V_n(t) + J_m(t) \xrightarrow{\nu} J_n(t) + J_m(t) \quad (4.8) \\
J_n(t) + U_m(t) \xrightarrow{\gamma} U_n(t) + U_m(t) \quad (4.9)
\end{align}

Here $U_s(t)$ means that node $s$ is unjammed at time $t$ and analogously for
the other two possible states of a node. Equations (4.7), (4.8) and (4.9)
translate the interactions described in the previous section. For example,
equation (4.7) describes how a free node $n$ can be made vulnerable by a
neighbouring congested node $m$, because $m$ can send load to $n$ but not in-
verse is forbidden by eq. (4.3); eq. (4.8) describes instead the transition of
a node from vulnerable to jammed. Finally, eq. (4.9) accounts for the possibility that a congested node relaxes on a free node without making the latter vulnerable.

Following standard methods for mean-field approaches in networks [51, 268] instead of considering single nodes, we group them by degree classes and introduce $U_k, V_k, J_k$ as the probabilities of choosing randomly a member of a $k$-degree node belonging to one of the three types. The rate terms are very intuitive and can be written down directly from the rate equations, considering the contributions from the different degree classes. For example, the contribution of (4.8) to $J_k$ is of the form

$$\Gamma^J_{\nu}(k,t) = \nu \sum_{k'=1}^{k_{\text{max}}} P(V_k(t), J_{k'}(t)).$$

The equations for the probabilities are:

$$U_k(t+1) - U_k(t) = \Gamma_\gamma(k,t) - \Gamma_\alpha(k,t) \quad (4.10)$$

$$V_k(t+1) - V_k(t) = \Gamma_\alpha(k,t) - \Gamma_\nu(k,t) \quad (4.11)$$

$$J_k(t+1) - J_k(t) = \Gamma_\nu(k,t) - \Gamma_\gamma(k,t) \quad (4.12)$$

Later on, we will set $\nu = 1$ as definition of vulnerable population.

For the moment we neglect degree-degree correlations and write the probabilities $P(a,b)$ as the probability of randomly choosing a node belonging to $a$ and a second node belonging to $b$ times the probability of having a link between the two. For example, $P(V_k, J_{k'}) = \frac{V_k \cdot J_{k'}}{N^2 M (N-1)}$, in other words the process rate times the number of relevant couples $VJ$. If we now substitute the $\Gamma$ terms inside the rate equations for the populations we obtain,
for each degree class $k$, the set of equations:

\[ J_k(t+1) = J_k(t) + \frac{k}{N(N-1)2M} \times \sum_{k'=1}^{k_{\text{max}}} k' \left[ \nu V_k(t) J_{k'}(t) - \gamma J_k(t) U_{k'}(t) \right] \]  

\[ U_k(t+1) = U_k(t) + \frac{k}{N(N-1)2M} \times \sum_{k'=1}^{k_{\text{max}}} k' \left[ \gamma J_k(t) U_{k'}(t) - \alpha U_k(t) J_{k'}(t) \right] \]  

\[ V_k(t+1) = V_k(t) + \frac{k}{N(N-1)2M} \times \sum_{k'=1}^{k_{\text{max}}} k' \left[ \alpha U_k(t) J_{k'}(t) - \nu V_k(t) J_{k'}(t) \right] \]  

with $J_k + U_k + V_k = p_k$ and $\sum_k p_k = 1$. Performing the sum over $k'$, we obtain the average degree for the respective populations. The eqs. (4.13)-(4.14) cannot be solved easily. We restrict to the case where the average degree is the same in the three populations, $\langle k_J \rangle = \langle k_U \rangle = \langle k_V \rangle$, a strong assumption that is however confirmed by the simulations. So for the stationary state we find:

\[ J_k = \frac{\nu}{\gamma} V_k \quad U_k = \frac{\nu}{\alpha} V_k \quad V_k = \frac{p_k}{1 + \frac{\nu}{\alpha} + \frac{\nu}{\gamma}} \]  

Assuming a simple proportionality between $\beta$ and the "jamming rate" $\alpha \propto (\beta - \beta_0)$, we can compare the prediction from the rate model

\[ J(\beta) \propto \frac{(\beta - \beta_0)}{\gamma + (\gamma + 1)(\beta - \beta_0)} \]  

with the simulated data. Figure 4.5 shows this comparison and we can easily see that the analytical solution reproduces well the behaviour of the simulated data up to $\beta \approx 1$. In particular, the fit correctly identifies the value of $\beta_0$ and reproduces the steep transition from vanishing stationary $N_J$ to a significant non-zero $N_J$. As expected, eq. (4.17) does not predict the second jump in $N_J = N/2 \to N$, because there the approximation of uncorrelated degrees and constant rates breaks down. Indeed, also eqs. (4.7)-(4.9) are not a good description anymore, since in this case the system is better
described by only two competing population V and J. Moreover, our mean-field-like rate model does not account for random fluctuations that could bring the system in the absorbing $N_f = N$ state.

Also, the more violent fluctuations observed under the global dynamics are indeed consistent with the core mechanism: the global dynamics, through its global flow suppression, produces smaller "packets" of flow, that allow nodes to stay nearer to their thresholds than under the local dynamics, and therefore produces a larger population of nodes susceptible of jamming, thereby generating bigger fluctuations. These results are reminiscent of the ones obtained in a recent paper by De Martino et al. [190], under very different assumptions. The authors developed a particle hopping model where each particle was random-walking on the network, with transition probabilities proportional to the population on the departure node. In their model, target nodes have a rejection probability of the form $\eta(n) = \eta^* \Theta(n - \bar{n})$, where $n$ and $\bar{n}$ are the population and the threshold for rejection of the target node. Also, at each hop a particle has a probability $\chi$ of being removed from the system, while particles are produced with a rate $p$ at each node. The order parameter in [190] is the total number of particles in the system at a certain time and it was shown to undergo a phase transition, the order of which depended on the value of the rejection prob-

![Fluctuations PDF](image.png)

Figure 4.4.: An example of fluctuations PDF for RG for $\beta \approx 0.91 (N = 1000)$. 
ability $\eta$. For $\eta$ smaller than a critical threshold $\eta^*$, the transition from free flow to congestion is smooth and continuous. For $\eta > \eta^*$ instead, the transition changes from continuous to discontinuous, that is congestion emerges suddenly at a critical value of the rejection parameter without warning. This is consistent with the abrupt, discontinuous transition we found in the rate model, which can be thought as the continuum, $\eta^* = 1$ limit of [190]. In [190] it is claimed that traffic control in the system, in the form of the rejection term, improves system’s global performances by enlarging the free-flow region in their parameter space. However, as shown in Fig. 4.4, a global information regime (which is in this case plays a similar role to traffic control/routing) is very likely to produce broad fluctuations in the number of congested nodes, than a purely local mechanisms. Indeed, for real systems, this model suggests that under heavy loads it might be better to let the system evolve under local information rather than attempting global management, since jams sweep a much larger part of the system in the latter case.

Also, a recent simulation study by Scellato et al. [47] appears to support this idea: Scellato’s model included a congestion-aware vehicle routing mechanisms, in which a traveller at node $i$ on route to node $t$ was provided...
with a penalty function of the form

$$P_n = (d_{in} + d_{nt})(1 + c_{int})^\alpha$$  \hspace{1cm} (4.18)

for each outgoing link \(i \rightarrow n\), depending on the link’s congestion. In this way, vehicles are encouraged to travel topologically longer but less congested paths. The authors found that such a mechanism produces a better spreading of traffic, increasing of a little amount the flow on many (previously) scarcely used streets, but effectively reducing the effects of congestion. Also, when they considered a global control (or rather global information scheme), in the form of an path-averaged \(c_{int}\) coefficient, they found that the global mechanism is outperformed by the purely local one. While they calculated the fundamental diagram for the whole network under different loads, it is not clear how the system evolves toward the completely congested state.

### 4.2.5. Conclusions

The model studied the two limits in information dissemination, local versus global. It is important to investigate at first minimal models of traffic coupled to information, in order to gain insights into the robust features of this interplay. We found that global information tends to delay the onset of congestion, because it produces a better diffusion of the flow on the network. At the same time, it is more likely to create strong fluctuations in the number of congested nodes as compared to local information only. However, the model studied in this section has a number of limitations that hinder its capacity to provide with detailed insights in the dynamic of transportation networks. For example, it is usual to consider costs or capacity associated with links and nodes, e.g. travel time along a link, whereas in this model there is only a congestion threshold on the nodes and no notion of length of travel time beyond that of topological distance between nodes. The model does not include any model of route choice for the travellers or an Origin-Destination (OD) demand matrix, relying only on the topological degree centrality of nodes and the status of nodes. Ideally, one would want to explicitly represent two types of behaviour — both of which will be influenced by some dynamic information field—
4.3. INFORMATION PROPAGATION WITH SELF-ORGANIZED PROPAGATION

the route choice behaviour of travellers and the network management behaviour of system managers (e.g. changes to capacity or link availability). As per the model studied above, these two elements are mixed, making harder to interpret the results. Indeed, in contrast to what is usually assumed for the packet-transmission models, one should assume two different —yet coupled— dynamical responses, one for the drivers and one for the network itself, be it considered as self-organised or the effect of network managers’ decisions. Finally, the way in which the information dissemination is presently embedded in the model is not very different from a particular routing mechanism (as shown by the comparison with the ZRP model in [190] with $\eta^* = 1$). Sections 4.4 and 4.5 will approach some of the issues described above. Before taking that step forward in modelling complexity, we take a conceptual step in a different direction. This section focused on the two extremes, local and global information. In the next section, we ask whether an optimal mesoscale exists for information diffusion. Also the next model will make drastic simplifications. The reason to make them is to study what are the fundamental, robust (to changes in the details of the models) features that differentiate the informational regimes.

4.3. Information propagation with self-organized propagation

In this section we study how a simple model of decentralised information dissemination on traffic networks affects the emergence and response to congestion.

In particular, we are interested in having some form of traffic information propagate through the network in a self-organised way, effectively producing a mesoscopic information dissemination length scale able to bridge the two extremes studied in the previous section, by allowing the information strategy to change. To do this we need a physical traffic network to act as a substrate for a superimposed information network to work. Also, we need the two to be interacting, that is the traffic state at a certain location must be able to influence the information network and vice versa. There are many ways of doing this and in general the two networks need not have the same topology: given a common set of nodes $V$ for the two networks,
4.3. INFORMATION PROPAGATION WITH SELF-ORGANIZED PROPAGATION

the set of edges for the physical network, $E$, and information network, $E'$, need not be the same, $E \neq E'$ in general.

It is important to specify what we mean by locality in this context. At the network level, a local interaction is defined as an interaction taking place over an edge. Let us assume now that the edge $(v, w)$ connecting node $v$ to node $w$ belongs to the information network, $(v, w) \in E'$, but not to the traffic network $(v, w) \notin E$. In this case, anything propagating on the information network from $v$ to $w$ would appear to be transmitted non-locally in the traffic network. Hence, when we refer to non-local propagation we intend this with respect to the (spatially embedded) traffic network. If instead we assume $E = E'$, the two networks are topologically the same and what becomes relevant is the different set of rules for the traffic flow and the information propagation. We focus on this second case, because we are interested in mechanisms of local propagation of information that create emerging non-local effects, without resorting to long-range effects. In particular, we consider cascading models similar in spirit to the classical sand-pile model of Bak-Tang-Wiesenfeld [84, 269, 270]. In these models, the system is composed by a network of sites that can hold a number of sand grains (see section 2.2); grains are then randomly added to the system until one site reaches its threshold; at the point, the driving of the system (adding the grains) is stopped and the over-threshold site is allowed to relax by toppling some of its grains on its neighbouring sites; this relaxation process is repeated until all sites are below their threshold and then the driving is resumed [217]. The main ingredient here is the separation of timescales, the slow driving scale and the fast relaxation one. Along these lines, we can think the information as propagating very fast along the network, modifying the state of the nodes, and the traffic slowly diffusing, according to the state of the nodes. This is reasonable, because the shorter time-scale represents the faster elaboration and transmission of information. For example, one might think of junctions, communicating with each other and receiving inputs only from their own road section and their neighbours. In such a case it would be technologically feasible to have information dissemination on a much shorter timescale than the physical (cars’) one.
4.3. INFORMATION PROPAGATION WITH SELF-ORGANIZED PROPAGATION

4.3.1. Model setup

We start then with a minimal theoretical model composed of a physical network layer (PL) coupled to an informational network layer (IL). On the PL we use a particle-hopping model ([190] [216] [249]) with first-in-first-out (FIFO) queues. For the information dissemination on the IL, we hypothesise sandpile-like dynamics ([217] [271]). For the physical network, we consider the dual graphs of traffic networks [165] [272] [273]. In the dual representation, a node represents a street and the crossings between streets become the links between nodes. In this way, central urban roads are generally represented by high betweenness nodes in the resulting dual network, which allows us to study the general mobility on the network without resorting only to the nodes’ betweenness centrality (we will make use of this in Eq. (4.21)). In this first model we focus on the queue dynamics, neglecting the internal link dynamics (the relation between flow and density). In this perspective, the maximal number of cars that a road can store, its storage capacity, can be at first order approximated by its length, which in the dual representation is related to the node’s degree: major urban roads are more likely to have more crossings, thus a higher degree.

We want each node on the PL to communicate directly only with its neighbours (decentralised system), but at the same time we want it also to be able to obtain information from a larger distance, when the network state requires it. To achieve this, we use the cascading process on the IL. The physical layer obeys a modified Zero-Range Process (ZRP) with a physical timescale $t_{\text{phys}}$ [249], while the information layer evolves on a faster timescale $t_{\text{inf}}$, such that $\Delta t_{\text{step}}^{\text{phys}} \gg \Delta t_{\text{step}}^{\text{inf}}$. For our purposes, one can imagine the information propagation to be instantaneous.

Each node is characterised by its population (or queue) $n_i$ and its informational state $S_i \in \{0, 1\}$. A particle hops from node $i$ with rate $\phi_i$ and chooses a target neighbour $j$ with probability $p_{ij} \propto b_j (1 - c_j S_j)$, where $b_j$ represents the betweenness centrality of node $j$ and $c_j$ tunes the importance given to the state of node $j$. We put $c_j = 1 \forall i$, because we are interested in the effects of strong correlations. So, whenever a node is congested ($S_j = 1$) the

---

2 The choice of a particle hopping model as opposed to a flow model is due to the fact that in the section 4.4, we will want to assign a destination to each particle (vehicle) and this will change radically the results.

3 The effects of introducing explicit navigation will be investigated in section 4.4.
probability for a particle to hop in its direction vanishes, similarly to what happened in the model described in section 4.2. The state of a node \( S_j \) is determined by:

\[
S_j(t_{\text{phys}} + 1) = \theta \left( k_j \frac{n_j(t_{\text{phys}})}{\bar{n}_j} + \sum_j a_{jl} S_l(t_{\text{phys}}) - k_j \beta \right),
\]  

(4.19)

where \( k_j \) is the node’s degree, \( \bar{n}_j \) a suitable capacity threshold, \( a_{ij} \) the adjacency matrix and \( \theta \) the Heaviside step function. The state \( S_j \) of node \( j \), parametrised by 0 (free) and 1 (congested/signaling), is therefore the result of its own population and of the state of its neighbours. As can be seen from the argument of the \( \theta \) function in eq. (4.19), \( \beta \) sets the threshold at which \( j \) changes its status to 1: the first term inside the \( \theta \) function is proportional to the fraction of the node’s filled capacity (\( n_j/\bar{n}_j \)); the second accounts for the information arriving from the node’s neighbours. For example, consider the case in which no neighbour is signalling (\( \sum_j a_{jl} S_l = 0 \)), then the node will change state \( (0 \rightarrow 1) \) only when the population is large enough \( (n_j/\bar{n}_j \geq \beta) \). At the other extreme, if the population vanishes, the node will change state only when a large enough number of neighbours signals that they are getting close to their own thresholds \( (\sum_j a_{jl} S_l(t_{\text{phys}}) \geq \beta) \). Naturally, the combinations between these two extremes are the most interesting ones, as they will be the ones to contribute the most to the information diffusion.

In line with the arguments above and previous studies\([163, 165]\), we choose to let the capacity be proportional to the node’s degree:

\[
\bar{n}_j = \alpha k_j.
\]  

(4.20)

The parameter \( \alpha \) is free and in the simulations it was set to \( \alpha = 100 \). We performed simulations on different topologies, square lattices (SL), Erdos-Renyi random graphs (ER) and scale-free (SF) networks, both Barabasi-Albert and random with degree distribution \( \propto k^{-\tau}, \tau \approx 2 \). It might seem strange to analyse ER and SF networks in the context of traffic networks, because we are used to think of urban areas as (almost) planar graphs that live embedded in 2-dimensional space. If we however go to the dual representation of the traffic network, we discover that the dual graph’s topology
4.3. INFORMATION PROPAGATION WITH SELF-ORGANIZED PROPAGATION

Figure 4.6.: Example of timeseries of the total number of particles on a Barabasi-Albert scale-free network for different values of $\beta$. From the bottom up, respectively $\beta = 0.60, 0.90, 0.59$. It is evident the sharp transition to a congested state for $\beta = 0.59$, where the nodes are effectively frozen as shown by the linear population increase. For $\beta_c = 0.6$ instead the system manages to deliver many more particles, entering a quasi-stationary state. Interestingly, increasing $\beta > \beta_c$ does decrease performances significantly.

is often close to random graphs or scale-free networks, since we find a large heterogeneity in the dual graphs of real urban networks [163, 164]. As an example, compare the situation of a high street together with all its side streets, and a grid topology like Manhattan. The former would present a hub (the high street) and many lowly connected nodes (the side streets), while the latter would have two classes of nodes with approximately the same degree (the avenues and the streets). For all topologies we simulated networks containing $L \approx 10^3$ nodes. The system is updated as follow:

1. Each node $i$ produces a particle with probability $p_i$.

2. Each node $i$ with a non-zero population tries to send out a particle. The target node $j$ in the neighbourhood of $i$ for the particle is selected
with probability $p_{ij}$:

$$p_{ij}(t_{phys}) = \frac{b_j(1 - S_j(t_{phys} - 1))}{\sum_l b_l(1 - S_l(t_{phys} - 1))},$$  \hspace{1cm} (4.21)

where $b_j$ is the betweenness centrality of node $j$ and obviously $\sum_j p_{ij} = 1$. The particle attempts to move to the target node and at each jump it is removed (absorbed) from the system with a probability $\mu$, representing the particle reaching its destination. If it is not absorbed and the target node is over capacity, the particle is not accepted with probability $\eta$ and remains node $i$. In our simulations, we chose $\mu = 0.2$ and $\eta \approx 0.3$.

3. When all nodes have been updated, the informational layer comes into play. At the beginning of each time step, all the nodes’ states are set to uncongested ($S_i = 0 \ \forall i$).

4. The system is scanned for nodes that satisfy

$$k_j \frac{n_j(t_{phys})}{\bar{n}_j} - k_j \beta > 0$$ \hspace{1cm} (4.22)

and their states are set to 1. We refer to these nodes as primary critical nodes, because they are the nodes who are congested due to their population only and are responsible for starting the informational cascade.

5. Each neighbour of a primary critical nodes is checked (eq. 4.19) and its state is set accordingly. We refer to the nodes that change their state (0 $\rightarrow$ 1) at this stage as to secondary critical nodes. The procedure is iterated, until the avalanche stops.

6. When the spreading on the IL stops, the information landscape, i.e. the set $\{S_i\}$, is fed back to the physical layer, namely in the $p_{ij}(t_{phys} + 1)$. This is the mechanism that provides the informational loop between the movements on the PL and the information spreading on the IL.

The choice of using the betweenness centrality as the bias on the routing mechanism certainly overlooks many details of the more realistic routing
protocols, e.g. origin-destination tables; however, at this stage, we are interested in accounting for the higher likelihood of a street to be used to move between two points and how this affects its congested behaviour and this is well approximated by the betweenness centrality in the case of urban networks [163]. Moreover, in general it is possible to tailor the production and absorption rates $p_i$ and $\mu_i$ in order to approach realistic origin-destination tables.

We record the system performances as a function of the parameter $\beta$ over a time window $T = 10^4$ physical timesteps, using high production probability ($p_i \geq 0.05$, see step 1 of the update algorithm) to stress the system and study the appearance of congestion, because we want to test the different behaviours at the emergence of congestion.

### 4.3.2. Discussion

The parameter $\beta$ is allowed to vary on $(0, \infty)$. The limit $\beta \gg 1$ represents the free system, where the particles are just routed following the unmodified $p_{ij}$, while $\beta = 0$ corresponds to the nodes signalling to each other as soon as they receive a particle. Because of eq. (4.19), the most interesting range for $\beta$ is $(0, 1)$. This interval expresses active control, with nodes signalling to their neighbours before they reach their capacity. Since we chose $p_i$ large (see for comparison [165]), we find that the system is not able to deliver all the produced particles and thus the total population in the system, $N(t) = \sum_i n_i(t)$, is usually increasing in time. However, varying $\beta$ one finds very different patterns. Figure 4.6 shows the results of simulations on a scale-free network for three values of $\beta$ and the same production rates. It appears that, for very low value of $\beta$, the system becomes very quickly unable to move particle around, effectively freezing. The population then increases linearly due to the nodes’ production rates. Increasing the value of $\beta$, the timeseries suddenly stop diverging and instead grows sublinearly in time. Increasing further the value of $\beta$ does not change qualitatively the evolution but the network’s performances worsen. In this regime, the secondary critical nodes populations are stationary for a wide interval of $\beta$’s (see Fig. 4.7) and this allows to extract the power spectrum of the timeseries. For scale-free and random graphs, the spectra scale as a power-law
4.3. INFORMATION PROPAGATION WITH SELF-ORGANIZED PROPAGATION

Figure 4.7: Stationary timeseries of secondary critical nodes for $\beta \approx \beta_c$ on a scale-free network for $\beta$ near the transition.

$f^{-\nu}$ of the frequency with a scaling exponent close to $\nu \sim 1$, typically found in self-organised critical systems, while the lattice produces instead a much flatter spectrum.

In figure 4.9 the final particle density per node, $n_f = N(t_f)/L$, is plotted versus $\beta$. Starting with $\beta$ very large and progressively reducing its value, we observe a decrease in the final density, meaning that the system is able to deliver more particles and have a smaller stationary population, thanks to effects of the information spreading. We find that $n_f$ is minimal for a certain critical value $\beta \approx \beta^+_c$, which represents a global optimum for the network dynamics. The critical value depends on the topology and, in principle, on the $\{p_{ij}\}$. Figure 4.9 shows that the value is larger for heterogenous networks. Approaching the critical value from above ($\beta \rightarrow \beta^+_c$), the final density has a jump reminiscent of a phase transition. This jump appears to be discontinuous for the square lattice and the scale-free network, while it is smoother in the case of random graphs. This can be understood thinking that, as $\beta$ is progressively reduced, the effective correlation length, represented by the dimensions of the connected components of critical nodes in the systems, increases both due to the lower threshold for the primary
critical nodes and to the facilitated spreading of secondary critical nodes. Nonetheless, the transition to complete congestion for $\beta < \beta_c$ limits the possibility of exploiting the network optimal point. In figure 4.10 we plot the distribution of sizes of the connected components of secondary critical nodes in the IL. The distributions shown are the ones obtained near the respective $\beta_c$’s of the topologies studied. Random graphs show a steep power-law dependence on the cascade size (exponent $\approx -3$), which is consistent with the $1/f$ spectrum and the smoother transition. Note also that the maximum cascade size in RG is significantly smaller than $N$, meaning that the avalanches on the RG topology are small. The homogeneity of the lattice produces instead a broad distribution of size, with an exponential cutoff around the system’s size ($L \approx 10^3$ in the simulations). The distribution for scale-free networks instead has two components, a steep exponential for small sizes and a broad peak roughly corresponding to the maximal degree in the network. Indeed, this is a sign of growing congestion on the hubs and their extended neighbourhoods. This behaviour is consistent with previous studies on condensation and failure spreading [122, 274], since it is known that in scale free networks large degree

---

4The size of a cascade is the total number of nodes that change state ($0 \rightarrow 1$ until it stops. If at some point in time two cascades originated from different nodes merge, they are considered as one with size equal to the total number of involved nodes.
Figure 4.9.: Final particle density $n_f$ vs $\beta$ for Erdös-Renyi graphs (red), Barabasi-Albert networks (green) and square lattice with periodic boundary conditions (blue). A transition from a completely congested state to a highly optimised state clearly appears in all three topologies.

nodes are the main (and most vulnerable) actors in diffusion phenomena. As a final remark, we note that in our study the total population $N(t)$ are not stationary themselves and our results are relative to observations over a time window $T$. Therefore, asymptotically we expect the population to diverge also for $\beta \geq \beta_c$. However, the development of congestion is delayed significantly: near the optimal $\beta_c$, $n_f$ is about 2 orders of magnitude smaller than the congested value and one order of magnitude smaller than the density obtained for $\beta \gg \beta_c$. In other words, defining the mean effective production rate $p_{eff} = n_f/T$, which corresponds to the constant production rate needed to obtain final population $n_f$ in case of no absorption, the system at $\beta_c$ needs to wait until $t_{phys} \approx 10T$ to reach the $n_f$ values obtained with $\beta \gg \beta_c$. 
4.3. INFORMATION PROPAGATION WITH SELF-ORGANIZED PROPAGATION

4.3.3. Conclusions

The model studied in this section is a model of routing, based on local interaction that allows for information to be propagated non-locally, thanks to a separation of timescales. We showed that, by tuning a congestion-awareness parameter $\beta$, it is possible to increase the network performances, namely increasing the delivery rate and significantly delaying the onset of congestion. However, this can be done at the cost of increased correlations in the system that invariably lead to a violent transition to system-wide congestion. Different topologies show peculiar patterns near the transition. For example, random graphs and scale-free network present power-law distributed temporal autocorrelation functions, but the way the two network topologies approach the transition is different, as shown by the shapes of the critical nodes’ size distribution. Scale-free networks accumulate their avalanches around a characteristic scale (related to the neighbourhood of the maximum degree hub in the network); avalanches on homogeneous graphs (lattices and random graphs) instead distribute broadly but with different functional shapes.

This feature could be exploited to further control the congestion spreading on the system (e.g. in exceptional circumstances, large flow interruptions, street works etc.). Indeed, the $\{p_{ij}\}$ enforce a dynamical topology on top of...
the physical one. In general, it would be possible to redesign this dynamical topology to comply with local and global parameters, without breaking the self-organisation. For example, one could define a new routing strategy $\{B_{ij}\} = B\left(p_{ij}, x_i^{local}\right)$, where the first term represents the information spreading, the second the node's internal structure.

Unfortunately, tuning $\beta$ to approach the theoretical $\beta_c$ can be dangerous. As mentioned before, $\beta_c$ in general depends on the topology and the set of $\{p_{ij}\}$. Hence, incomplete knowledge on network structure and inflow might yield a wrong value for $\beta_c$ and therefore one might drive the system to congestion instead of exploiting the optimality near it. Fortunately, there is a catch. Since for $\beta \geq \beta_c$ the performances of the network decrease slowly, as measured by the dependence of the final density of $\beta$, $n_f(\beta) \approx \beta^{-0.5}$ (see Fig. 4.9), one could choose values not too close to the transition, losing little on the performances but greatly reducing the risk of accidentally triggering the emergence of congestion.

### 4.4. Self-organised propagation on realistic networks

In the previous section, we showed that a self-organised mesoscopic information dissemination length can help reduce congestion and increase the capacity of a network. The topologies and navigation rules we studied however were not very realistic. In this section we study a similar model of traffic dynamics with a non-local model for information dissemination, using however a more detailed description. First, we describe explicitly OD couples and navigation for the travellers (section 4.4.1); then, in section 4.4.2 we test this model on networks based on actual urban topologies (Fig. 4.11) under different information schemes and discuss the results using two modifications of the model (sections 4.4.2 and 4.4.3). Finally, we discuss the results in section 4.4.3.

#### 4.4.1. Model Setup

**Network structure**

The traffic network is composed of streets (links) and junctions (nodes) equipped with traffic lights. Junctions are indexed by latin letters ($a, b, c, ...$),
Figure 4.11.: NYC Manhattan. The red highlighted section is composed 426 junctions and 1426 one-way links. Two one-way links connecting the same pair of junctions represent a two-way street.
while streets are identified by an ordered pairs of letters, i.e. the street going from junction $z$ to junction $t$ is labeled $zt$. Each link $ij$ can hold a maximum number of cars, $\bar{n}_{ij}$, and, when it is served by the traffic lights at junction $j$, has a maximum outflow $O_{ij}$. If a car tries to turn into an already full link, it is refused entrance and is stuck at the head of its current link. This design represents spill-over effects and is the main cause for congestion spreading through the network. Junctions are set up as fixed-cycle lights, serving one incoming link at a time, and junctions are processed in random order. Each time step in the simulations corresponds to one green phase. Turning probabilities at each junctions are calculated individually for each traveller.

Figure 4.12.: The primal graph built from the red section of figure 4.11, the picture shows how the graphs were obtained from map images of the cities. Junctions were detected and the distances were reconstructed walking along the links.
Navigation

The two key elements are the travellers’ navigation rule and the information dissemination rule. We assume travellers have a complete knowledge of the network’s topology, the map. However, they only have information regarding the congestion on the nodes neighbouring the one they currently reside. Thus, a traveller $A$ on route to $d$ and currently on node $i$ will choose one of the links outgoing from $i$ with a probability:

$$
p_{ij}^d(A) = \frac{\beta_j (1 - c_A S_j(t)) / l_{ij}^d}{\sum_{m \in \Gamma_i} \beta_m (1 - c_A S_m(t)) / l_{im}^d},
$$

where $\beta_j$ is the betweenness centrality of node $j$, $c_A$ the compliance of traveller $A$ to the received information and $S_j(t)$ the state at time $t$ of node $j$. The quantity $l_{ij}^d$ is the shortest path length from $i$ to $d$ going through $j$ and the exponent $\alpha$ is calibrated on simulation runs with vanishing population in order to provide meaningful travel times for the travellers. The denominator just normalises $p_{ij}^d$ to obtain a well-defined probability. The betweenness values $\{\beta_j\}$ were included to mimic the preference of travellers to choose more central routes. The state $S_j$ is the quantity that relates the information spreading with the traffic dynamics. At time $t + 1$ it is given by

$$
S_j(t + 1) = \Theta \left( n_j(t) + \sum_{m \in \Gamma_j} \bar{n}_{jl}(t) - \beta \sum_{m \in \Gamma_j} \bar{n}_{jl} \right),
$$

where $\beta$ is a parameter and the Heaviside theta function $\Theta$ takes values $\Theta(x) = 1$ $\forall x > 0$ and 0 elsewhere. In eq. (4.24) similarly to eq. (4.19) in the previous section, the second term in parenthesis encodes the influence of a node’s neighbours: if any of the neighbours of $j$ has state $S$ this increases the chances that $j$ itself will have $S = 1$ at time $t + 1$. Like in section 4.3 the states of nodes are updated in two steps:

- first, we look for the nodes that satisfy $n_j(t) - \beta \sum_{m \in \Gamma_j} \bar{n}_{jl} > 0$, which we label primary critical nodes, as they are the one that change their state independently of their neighbours’ behaviours;

---

5Specifically, in an empty network, one traveller was introduced and the value of $\alpha$ increased until the travel time as a function of distance stabilised.
• we then check whether any of the neighbours of primary critical nodes satisfy the condition \( n_j(t) + \sum_{m \in \Gamma_j} \bar{n}_{jm} S_m(t) - \beta \sum_{m \in \Gamma_j} \bar{n}_{jm} > 0 \). If we find any, we set its state to 1 and label it a secondary critical node. Every time we find a new critical node, we repeat the procedure. In this way the same primary critical node may trigger an avalanche of state changes in the system, depending on the distribution of populations in the system and the topology of the network[122, 217]. This is again what allows nonlocal propagation of the information.

The above rules introduce two timescales to the dynamics, the shortest relative to the state switching and cascade forming, the longest relative to the travellers’ movement, indexed by the time stamp \( t \). In a realistic setting, the shortest scale would represent electronic communications between neighbouring junctions, therefore justifying the separation of timescales between the information and physical dynamics.

If the second step is skipped, i.e. one does not look for secondary critical nodes, the state of a node does not propagate across the system and has thus purely local effects, that is local propagation.

### 4.4.2. Simulation Results and Discussion

We simulated different types of networks, both realistic and artificial. It is well known that the topology of a network influences the evolution of the superposed dynamical process, e.g. the vanishing of the infection threshold for epidemic spreading on scale-free networks [275] or viral propagation on the Internet [271]. We are interested in urban traffic networks, which in their primal representation do not share the heterogeneity in degree distribution and small-world characteristics of other real-world networks[163, 273]. This is mainly due to the spatial and planarity constraints enforced by the underlying geometry and the costs involved in building long links [27, 80, 276]. However, our results turn out to be consistent over the different topologies we studied. From Fig. 4.11 we extracted the part of street network highlighted and built its primal graph (Fig. 4.12). The reconstructed network has an average degree \( z = \langle k \rangle = 3.3 \pm 0.8 \), meaning that the degree distribution is very narrow and most nodes have either 3 or 4 streets connecting to them, as expected from Fig. 4.12. When the most
time is spent passing from one queue to the next, as opposed to the actual flowing, a good proxy for the capacity \( n_{ij} \) of link \( ij \) is its length, \( l_{ij} \).

We therefore extracted the node to node distances from Fig. 4.11 and used them to assign capacities to the different links. Figure 4.13 shows the distribution of capacities obtained.

In addition to the Manhattan section shown in Fig. 4.11, simulations were performed on Erdős-Rényi graphs, square lattices, a part of the urban network of San Francisco and another section of the Manhattan network. For the Erdős-Rényi the average degree was set at \( z \approx 4 \) for comparison with the other networks studied. For artificial networks, links were given a capacity \( C = 30 \) and outflow \( S = 15 \), which are compatible with realistic networks, as shown in Fig. 4.13a. Both Erdős-Rényi graphs and lattices showed much larger \( N_{c}^{cong} \approx 65 - 70 \) due to the greater homogeneity. This was expected since they are more homogeneous than realistic networks both in degree distribution and capacity distributions and it is know that these properties allow to support more traffic (at the cost of speed of delivery however)[261, 274]. The (storage) capacities for the links in real networks were instead set proportionally to the link’s length. The simulations results did not show significant differences from the ones obtained for the Manhattan network. As such, we examine here in detail only the case of Manhattan (shown in Fig. 4.11).

In order to establish the comparison baseline for the local and non-local information dissemination models, we first examine shortest-path routing without information. The rate \( N_c \) at which cars are introduced in the system is our primary indicator. Indeed, we are interested in finding the lowest value of \( N_c = N_c^{cong} \) for which the shortest-path routing develops congestion in the system \(^6\). The red dots in figure 4.13b show the total number of cars that reached their destination, \( D_{tot} \), as a function of the rate, \( N_c \), at which cars are introduced in the system. Already for very low values of \( N_c \), the shortest path routing alone starts to perform poorly (\( N_c^{cong} \approx 5 \)). The explanation is that travellers are completely unaware of the congestion state of their surroundings and naturally navigate straight to their destination. Central nodes then are forced to withstand higher traffic flows and

\(^6\)The shortest-path routing regime can be easily reached by setting \( c_A = 0 \) for all travellers. In this way the information is completely discarded and thus equivalent to the case of no information.
eventually break down, leading to congestion formation. The introduction of information should counter this dynamics, allowing stressed nodes to warn their neighbours before they reach saturation and thus become seeds for congestion. The threshold $\beta$ is therefore crucial as it plays a role both in the primary signalling process and in the dissemination of it. In figure 4.14 we plot the total number of delivered cars for the Manhattan network for different values of $N_c$ and $\beta$. The red colour identifies the regions of parameter space that maximise $D_{tot}$. As expected for $N_c < 5$, both mechanisms reproduce the results of the shortest-path routing and there is no dependence on $\beta$ as the spreading of information does not activate. For $N_c > 5$, $\beta$ becomes important. In the following, we refer to $\beta$ for the local and nonlocal cases respectively as $\beta_l$ and $\beta_{nl}$. The nonlocal mechanism is most efficient for $0.5 < \beta_{nl} < 1$, while the local mechanism for $0.1 < \beta_l < 0.6$. Also, both the information spreading strategies extend the range of $N_c \leq N_c^{cong} \approx 15$ for which the traffic network stays fluid, allowing travellers to reach their destinations.

It is interesting to note that when $N_c$ increases the interval of effective values of $\beta$ shrinks. This can be seen from the reduction of the red-yellow region in Fig. 4.14. Also, for all $N_c$ we find similar performances for values of the threshold that satisfy $\beta_{nl} \geq \beta_l$. This can be understood by considering the dynamics of how information cascades. Indeed, in the nonlocal spreading mechanism $\beta$ is the also threshold for the secondary critical nodes (see eq. (4.24)). Therefore, a low value of $\beta_{nl}$ does not only increase the probability for a node to change status due to its own population, but also the probability that a single node can cause large cascades. In the limit $\beta_{nl} \rightarrow 0$, one car on one node would cause the whole system to switch to $S = 1$. For this reason, we find better performances for higher values of $\beta_{nl}$, which allow for both small and large cascades. Conversely, in the local information mechanisms, nodes have no way to steer away incoming cars as long as they are not in their direct neighbourhood. Thus, in order to avoid overloading, they need to keep a larger fraction of their capacity free, in a sense pre-emptively "protecting" themselves more from incoming traffic.

To compare the performances of the two spreading mechanisms with the shortest-path baseline, we select $\forall N_c$ the maximum number of vehicles that reached their destination during a simulation, $D_{tot}$, and plot them in Fig. 4.13b. Local (blue) and non-local (green) both outperform by far the
shortest-path strategy in terms of the number of delivered cars, but also as already seen in Fig. 4.14 shift $N_{c}^{cong}$ to higher values. Surprisingly and in contrast to the results of section 4.3, the most efficient strategy turns out to be the local one. The plot shown refers to the Manhattan section, but we consistently observed the same behaviour in the other topologies studied (see Appendix). It is important to notice that the value of $N_{c}^{cong}$ for both the information strategies is very similar, and so is $D_{tot}(N_{c})$ for $N_{c} < N_{c}^{cong}$. In section 4.3, we found that a value for the signalling threshold existed such that the network would deliver the maximum number of vehicles. In the model analysed in this section, the presence of an explicit navigation makes the local regime more able to deliver more vehicles to their destination. The main difference is the shape of the $D_{tot}$ curves for $N_{c} \geq N_{c}^{cong}$: the local strategy, while reducing its performances, systematically delivers more cars than the non-local strategy and the shortest-path routing. How is this possible? Our results paradoxically seem to suggest that increased access to free\textsuperscript{7} information is detrimental to the system, at both the user and global level.

A possible explanation comes from considering the following picture (see Figure 4.15). Imagine a network where a congested area is forming. A traveller possessing only local information would proceed straight toward its destination until she is informed of the presence of congestion on the next link; at that point, with a large probability, she would continue its trip around the border of the congested area and thus travelling along a trajectory which produces the smallest increase of travel time, compatible with the presence of congestion.

Under nonlocal information, travellers have a chance to reroute along slightly longer routes, which allow them to reach their destination with small delays or extra distances. However, at the network level, these small delays sum up and globally reduce the maximal traffic demand that the network is able to sustain. In other words, nonlocal information seems to promote the use of a larger set of routes, thus promoting fluidity in the traffic, but in this context is not at the same time able to guarantee enough travel completion. Our results suggest that the best tradeoff is given by local only congestion, which allows travellers to avoid getting trapped in long queues, but also produce a very small disruption of the shortest path routes. It is

\textsuperscript{7}In the model, travellers do not have a cost to access the information.
4.4. SELF-ORGANISED PROPAGATION ON REALISTIC NETWORKS

Figure 4.13.: a) Distribution of capacities \( \bar{n} \) for single links. The capacity \( \bar{n}_{ij} \) of link \( ij \) corresponds to the maximum number of cars it can hold. \( \bar{n}_{ij} \)s were approximated by the ratio \( l_{ij}/h \) between the link length \( l_{ij} \) and a standard length \( h \approx 5 \text{mt} \), representing a standard measure for car length plus headway. b) Number of delivered cars \( D_{tot} \) as function of the forcing rate \( N_c \) (akin to traffic demand): the shortest-path routing (red) systematically delivers less cars than the routing with nonlocal (green) and local (blue) information dissemination. For local and nonlocal dissemination, the plotted points correspond to the maximum value of \( D_{tot}(\tau) \) at a given \( N_c \) (see Fig. 4.14 for details).
possible to investigate to which extent this picture holds, by studying the Fundamental Diagram of the network, which represents the flow through the network as function of the network load density. In Fig. 4.16 we plot the ratio $F/F_{\max}$ versus the density $\rho = \frac{M}{\sum_{i} n_{ij}}$, where the flow $F$ is defined as the number of cars that move during one timestep, $F_{\max}$ is the maximum possible outflow, $F_{\max} = \sum_{i} O_{ij}$, and $M$ is the total number of cars in the network. At any given $\rho$, the local and nonlocal strategies (blue and green points respectively) allow the system to exploit much higher flows than the shortest-path routing, effectively keeping the traffic fluid and letting it flow around potentially disruptive nodes. However, after the maximum at $\rho \approx 0.25$, the slope of the curves is steeper than for the shortest-path rout-
4.4. SELF-ORGANISED PROPAGATION ON REALISTIC NETWORKS

Figure 4.15.: Under local information dissemination (top), a traveller continues until she encounters congestion and then continues following the shortest deviation around the congested area. Under nonlocal information (bottom), a traveller might instead reroute across a trajectory that is further away from the congested area and thus involves areas which were not affected by the emerging congestion. The extra length (or travel time) in the second case can eventually stack up at the network level, making the network more unstable and prone to develop network-wide congestion.

ing, even though they remain above it. Note that the slope of the curve corresponding to the local strategy is slightly smaller than the one corresponding to the nonlocal information strategy. This result supports the picture described above: although the information dissemination strategies studied both allow to increase the maximum flow, they induce correlations that can also hinder the flow, as for example in the case of cars being redirected around a congested area. The performances of the system depends on the balance between these two competing factors. The nonlocal information dissemination strategy creates correlations on a longer range than the local one and therefore tends to break down earlier as observed. In other words, what happens is that, where the non local information mechanism works, it is because it produces only small cascades, comparable to the local mechanism. When instead the congestion state worsens, the size of the cascades grows quickly engulfing the whole net-
Limited propagation

One possible way to further enhance the performances of the nonlocal mechanisms might be considering two different thresholds, one for the primary signalling and one for the propagation of the signal itself. Although such a setup would delay the onset of system-wide cascades, there would still exist a density at which the emergent length scale becomes comparable with the size of the system and the argument regarding local and nonlocal correlations would stand.

A way around would be to set a threshold for propagation, as in Eq. (4.24), but limit the maximum number of secondary critical nodes that can be present at each time. This method does not incur in the same problem of a system-wide propagation by construction, but requires a global control on
the number of critical nodes to stop the information propagation at some level.
This can be done by considering, in addition to the signalling threshold $\beta$, a parameter $\epsilon$ representing the maximum fraction of the system that can be critical at any one time, $\epsilon = n^{\text{cong}} / N$. We allow any number of nodes to be primary critical nodes (the one who satisfy $n_j(t) - \beta \sum_{m \in \Gamma_j} \bar{n}_{jl} > 0$), but we only allow a total number of secondary critical nodes smaller than $\epsilon N$. This constraint can be enforced by stopping the propagation of the information cascade once the maximum number $\epsilon N$ is reached. If the number of primary critical nodes is already larger than $\epsilon N$, we simply do not propagate at all. We simulated the system as before and studied the performances as a function of both $\beta$ and $\epsilon$ for different values of the forcing $N_c$. Fig. 4.17 shows the results for a simple lattice grid. The forcing $N_c$ increases from left to right and top to bottom. The quantity shown on the z-axis is the total number of delivered cars. We can see that for low forcing there are two regions of the parameter space ($\beta, \epsilon$) that allow for high delivery values. For $0.4 \leq \beta \leq 0.7$, the performances of the system do not depend on the value of $\epsilon$. However, as the forcing is increased ($N_c \sim 76$), the number of delivered cars decreases and eventually vanishes. Interestingly, for the same forcing, in the region ($0.1 < \beta < 0.3$, $\epsilon < 0.2$) of the parameter space, the system manages to remain fluid and deliver a large number of cars. This part of parameter space corresponds to junctions being very sensible to congestion, due to the low value of $\beta$, but also to a small maximal number of critical nodes (less than 20% of the network), implying that the purely self-organised information length model is outperformed by a model characterised by frequent, yet small localised cascades. Notice however that the peak delivery values are always larger for $\epsilon = 0$, which corresponds to local information alone (only primary critical nodes). Therefore, also in this case, we see that the local congestion awareness is the preferable strategy.
4.4. SELF-ORGANISED PROPAGATION ON REALISTIC NETWORKS

Figure 4.17.: Results of limited propagation dynamics for a 10x10 lattice grid. The plots show represent the number of delivered cars as a function of the signalling threshold $\beta$ and the maximal critical fraction $\epsilon$. The forcing $N_c$ increase from left to right, top to bottom.
Betweenness analysis

The results obtained in the previous subsection are supported also by a different analysis. In equation 4.23 we biased the navigation rule with the betweenness $\beta_j$, in order to represent the fact that travellers on the way to their destination tend to favour central rather than peripheral nodes. Imagine now to give only to nodes with $\beta_j$ larger than a certain $\beta_0$ the capacity to signal. For ease of visualisation, let us consider a 10x10 square lattice network with $N_c < N_{cong}$ and $\beta$ in the range for stability. Fig. 4.18 shows how the probability of developing congestion in a lattice network changes as function of the fraction of signalling nodes. We see that trimming the low betweenness nodes does not increase significantly the probability to develop congestion. This might not seem very surprising, since it is well known in network theory that low betweenness nodes tend to have little influence on dynamical processes on graphs\cite{277,281}. However, even when only 16 out 100 nodes can signal, the probability of congestion is still very small (red nodes in the bottom right square of Fig. 4.18). Only when the inner nodes able to signal begin to be removed (centre square), the probability changes, growing abruptly and quickly reaching one (top left square). Indeed, it appears that, to avoid the emergence of congestion, it is more important to protect a few very central nodes (in our case, high betweenness nodes), rather than to disseminate the information to all the nodes. In particular, the model performs well as long as the connected core of signalling nodes is able to communicate and steer away incoming traffic, avoiding the emergence of a first seed of congestion, that can eventually grow into a network-wide jam (akin to the congestion core of section 4.2).

4.4.3. Conclusions

In this section we studied a model of decentralised nonlocal information dissemination, with an emergent length scale coupled to the traffic state, and compared it with models with local or no information. In line with previous results for global and local routing \cite{47}, we show that under the proposed mechanism non-local information performs significantly better than shortest-path routing without information, allowing the system to sustain a larger flow for higher densities and increasing the total through-
Figure 4.18.: Probability of congestion development as function of the cutoff of nodes able to signal. The nodes highlighted in red are those that are able to signal, selected on the basis of their betweenness centrality.
The long range effects in the traffic flow, introduced by the emergent dissemination length scale, allow for effective congestion avoidance by automatically rerouting travellers around congested areas. However, this turns out to be detrimental to performances when the traffic density increases: when information cascades become too widespread and probable, the traffic is strongly constrained and the performances decrease. This solves the apparent paradox of congestion favoured by distributed information: information cascades allow the system to self-organise in order to react effectively for moderate densities, but when the density increases cascades easily percolate through the system and become useless, because all nodes are critical.

Constraining the maximal number of critical nodes improved the maximum forcing $N_{\text{cong}}$, but did not outperform the simple local information routing. Indeed, for the class of propagation models similar to the proposed one, our results suggest that the local routing strategy is the optimal one at high traffic densities, as it is the one that better reconciles avoidance of congested nodes with the shortest correlation lengths in the traffic. This result is in line with recent (and independent) work by Meloni and Gomez-Gardenés [282] for congestion in communication networks. In their article, minimisation of congestion was obtained through local congestion-aware routing schemes, which are known to cause the phase transition from free flow to congestion abrupt (first-order); however, through the use of local empathic routing, a scheme in which a node considered the state of its first neighbours (like in Eq.4.24), they were able to bring back the transition to second-order, making it smoother and hence preferable from the point of view of applications.

In this chapter we have, until now, focused on the property and effects of information dissemination on traffic, finally showing that limited (local) dissemination is in most cases preferable to large-scale dissemination. However, these models all treated information as part of the traffic infrastructure, rather than an asset that could be exchanged among travellers. Intuitively, from the point of view of a single traveller, access to personalised information appears to be an obvious advantage. In the next section, we study this case: exchange and dissemination of information by agents, rather than the infrastructure.
4.5. Model of Aging and Propagating Information

This section will focus on agent-mediated information models reminding of [243, 244]. We will compare three paradigms. The first two, shortest path routing and routing with local congestion awareness, do not need further introduction. The third model attempts to explore the gap between purely local and purely global routing mechanisms by putting information in the hands of travellers: we take inspiration by navigation methods of ants, in the sense that we exploit the cooperation between vehicles and traffic infrastructure. Vehicles are at the same time users, producers and carriers of traffic information spreading. Users, because they can request information to choose their route. Producers and carriers, because they themselves collect and carry information around during their trip through the network. An example of how a similar cooperative multi-agent mechanism could be implemented in practice was proposed by Adler [283]. Here, however, we focus on the effects, rather than on the details of the realisation. In particular, our approach allows us to study a range of different length scales for information dissemination, through the tuning of a "forgetting", or information removal, parameter (akin to the evaporation of ants’ trail [284]). For fast removal of information, the model tends to the model based only on local information; while, for slow removal, such information comes into play effectively modifying the environment and therefore the behaviour of the agents.

This is different from the models of global information considered in the previous sections. There, the information was propagated on a shorter timescale than the evolution timescale of the traffic itself. Moreover, the information was managed by the infrastructure of the network, making the travellers simple passive users. Here instead, we investigate whether strategies similar to (and inspired by) information processing in ants’ societies but with limited information can be effective for congestion relief.

This section is organised as follows: in subsection 4.5.1 we describe the implementation of the models on a network traffic model, the navigation rules used for the simulations and how the travel time information is embedded in them; we then present and discuss the results in subsections 4.5.2 and 4.5.3 concluding in subsection 4.5.4.
4.5.1. Description of the model and setup

Implementation

Simulations for the three models have been performed on square lattices with \(N = 100,400\) and \(900\) nodes. Each node is a junction and has 4 outgoing and 4 incoming links, with the exception of the nodes situated on the border of the lattice. Each link is thought of as a road section with a buffer \(B\) of 30 cars, so the setup represents a regular grid network with two-way, double-carriageway roads. Junctions have fixed cycle traffic lights.

At each discrete time step of the simulation, junctions serve one of the incoming links (referred to as active link in the following). For each junction, the first link to be served is randomly selected in the beginning of the simulation, then the links are served sequentially. During a green light, up to \(S = 15\) cars in the queue of the active link can attempt to cross the junction and continue their trip. Starting with the first car of the active link’s queue, we calculate the turning probabilities (that depend on the car’s destination and the travel times stored on the neighbouring sites), we roll the dice and let the car attempt the turn. If the selected outgoing link for the turn is full, the car is stuck at the head of the active link’s queue and so are all the cars behind it for that cycle. This is implemented in order to simulate the effect of congestion spill-over and is the main ingredient of congestion spreading. Cars are introduced with rate \(N_c\). We performed simulations both with constant \(N_c\) and with a daily variation profile with average rate \(\approx 0.95N_c\). They navigate through the network according to the rules described below (see section 4.5.1).

When a car moves to a new link, it is placed at the back of the queue of the new link. Their origin and destination nodes are chosen randomly, then a random outgoing link of the origin node is selected and the car is placed at the back of its queue. A car is considered to have reached its destination when it reaches an ingoing link of the destination junction and it is then removed from the system. We monitor the populations on links, the travel times, the frequency of information updates and the number of cars that reach their destination. All runs last up to \(T_{\text{max}} = 24 \times 10^3\) time steps, unless network-wide congestion emerges. In such case, we interrupt the simulation run. Each time step represents a green light cycle. The very long time window is chosen in order to study the asymptotic stationarity.
of the system under different choices of parameters.

**Car Navigation**

Suppose a car on the route to $d$ reaches junction $l$. It asks from junction $l$ the travel time information along each of the outgoing links ($l_k, l_m, \ldots$) and selects them with probability [285]:

$$p^d_{lk} = \frac{\exp(-\beta(T_{lk} + \hat{T}_{kd})/T_0)}{\sum_{m \in \Gamma_l} \exp(-\beta T_{lm} + \hat{T}_{md})/T_0)}$$  \hspace{1cm} (4.25)

where $\Gamma_l$ is the neighbourhood of $l$, $T_{lk}$ the travel time on the link $lk$ (local information) and $\hat{T}_{kd}$ the travel time information from the next junction ($k$) to the destination $d$. The $\{T_0^i\}$ and $\beta$ represent respectively the reference travel times and how likely travellers are to follow the shortest route: for large $\beta$, a car almost surely chooses the link with the shortest reported travel time. For a network with vanishing car density we find that the reference times are well approximated by $T_0^i \sim 2d_{ij}$, where $d_{ij}$ is the topological distance between junctions $i$ and $j$ (see Appendix).

Note that the form of the turning rates in Eq. (4.25) is general and encompasses the three models we will compare. Indeed, the differences among the navigation models are determined by how the travel times involved in Eq. (4.25) are obtained. In the shortest-path routing, agents have complete knowledge of the map, but are completely unaware of the traffic conditions. Therefore, $T_{lk}^{loc}$ and $T_{kd}^{inf}$ coincide with the reference values $T_0^{lk}$ and $T_0^{kd}$ respectively. The turning probabilities take then the form:

$$p^d_{lk} = \frac{\exp(-\beta l_{lk}^{st}/l_{ld}^0)}{\sum_{z \in \Gamma_l} \exp(-\beta l_{lz}^{st}/l_{ld}^0)}$$  \hspace{1cm} (4.26)

where again $l_{kd}^0$ is the shortest-path distance between $i$ and $d$ passing through node $k$ that belongs to the neighbourhood of $i$ ($k \in \Gamma_i$), and $l_{ld}^0$ is the actual shortest-distance ($l_{ld}^0 = \min_k (l_{kd}^k)$). In the case of local congestion awareness, the updated travel time on the next link, $T_{ij}^{loc}$, is supposed known at every time and estimated as $T_{ij}^{loc} = 2.2Q_{ij}(t)/S$, where $Q_{ij}(t)$ is the length of the queue on the link $ij$ at time $t$ and $S$ the maximum outflow from the link during a time step. The turning rate for the local congestion-aware
model are therefore:

\[
p_{lk}^d = \frac{\exp(-\beta(T_{lk}^{\text{loc}} + T_0^k)/\hat{T}_{ld}^0)}{\sum_{m \in \Gamma_l} \exp(-\beta T_{lm}^{\text{loc}} + \hat{T}_0^m/T_{ld}^0)}
\]  

(4.27)

For the routing with dynamical information, Eq. (4.25) cannot be further simplified and thus we need a prescription to calculate \(\hat{T}_{kd}^{\text{inf}}\). Before giving the functional form for \(\hat{T}_{kd}^{\text{inf}}\), it is important to explain how the routing with decentralised information works. The first step in this direction is promoting cars and junction to active interacting agents. Up to now, in fact, cars and junctions did not have an active role in the traffic dynamics: in the shortest path mechanism only topological distances matter; in the local congestion-aware mechanism junction simply communicate to cars the updated travel times on the outgoing links. The third mechanism instead requires travel times to be sourced by the cars along their routes and exchanged with the junctions. More specifically, we characterise each car by:

1. Its origin junction \((o)\) and origin time \(t_o\), which will be used to assess the relevance of the information carried by the car;
2. its position, as defined by the link on which it is introduced;
3. its destination junction \(d\);
4. the cumulative travel time toward \(o\) collected as it moves towards \(d\), \(T_{io}^{\text{cum}}\).

Junctions, on their part, collect both the travel time information carried by cars that pass through them, and provide them with travel times toward their destinations, taking into account the age of such information. In particular, for each destination \(d\) a junction \(i\) stores the following set of numbers:

1. a reference travel time \(T_{id}^0\) to \(d\) (free flow or expected travel time for example);
2. the most recent travel time information received from cars, \(T_{id}^{\text{inf}}(t_o)\), where the \(t_o\) is the origin time of the car that carried the last obtained information and it is used to decide whether to update \(T_{id}\) or not when new information is received;
3. the time \( t_{up} \) at which \( T^{\text{inf}}_{id} \) was last updated.

To explain how information is collected and disseminated, let us consider an example. Suppose a car is introduced on an outgoing link of its origin node \( o \) and that we denote by \( i \) the junction at the other end of said link. When the car reaches \( i \), it collects the travel time information \( T^{\text{loc}}_{io} \) from the current junction and adds it to its own \( T^\text{cum}_o \). In this case, the information stored in \( i \) about \( o \) is not updated because \( i \) and \( o \) are neighbours and we assume junctions to be aware of their own links’ state. Suppose further that, on its route to \( d \), the car chooses to follow the link joining \( i \) and \( j \). When it arrives in \( j \), it updates \( T^\text{cum}_o \rightarrow T^\text{cum}_o + T^{\text{loc}}_{ji} \). We stress that each car carries just the total travel time from its current position to its origin \( o \), not the travel times on the various links. The junction then compares the car’s \( t_o \) with the stored origin time \( t'_o \). If the new car is more recent (\( t_o > t'_o \)), then the junction updates \( T^\text{inf}_{io} \) with \( T^\text{cum}_o \), otherwise it neglects it. If the junction does not have any stored information about \( o \), it stores the one carried by the car.

Through this mechanism, each junction collects information about the time needed to reach different destinations in the network, building a travel time landscape from itself to all the other junctions. In case it has no dynamical information (travel time carried by a car) regarding a specific destination, we assume that it uses the standard reference times set, \( \{T^0_{ij}\} \), which is specified a priori. It is clear at this point that an additional element is needed: limited memory or, more precisely, information removal. Indeed, suppose that congestion develops in a certain area of the network. Junctions receiving this information will redirect traffic on other routes as travel times through congestion is increased. Suppose furthermore that, after a while, congestion is relieved and traffic returns to normality, but some junctions do not receive the new information, since no cars carrying it pass by them. Without a mechanism to remove the outdated (and by then probably wrong) information, cars will be routed along longer paths in order to avoid a bottleneck which does not exist anymore. Such vehicles would prolong without reason their own travel times and potentially hinder other vehicles. The prescription for \( T^\text{inf}_{kd} \) needs to embody the elements mentioned above. We choose the form:

\[
T^\text{inf}_{kd} = (1 - e^{\lambda (t - t_{up})}) T^0_{kd} + e^{-\lambda (t - t_{up})} T^\text{inf}_{kd} (t_o)
\]  

(4.28)
where the dynamical time $\hat{T}_{k,d}^{inf}$ emerges as an exponential averaging between the reference time and the travel time information received by the passing cars. The exponent of the exponential weights is a composition of the time passed since the last update, $t - t^{up}$, with the forgetting parameter $\lambda$. Small values of $\lambda$ correspond to retaining for long times the travel time information carried by the cars. Note that in the limit $\lambda \to \infty$ we obtain the local congestion-aware model, since $\hat{T}_{k,d}^{inf} = T_{k,d}^{0}$ for all $t$ and $k,d$. Through this setup it becomes possible to study the behaviour of the system under a continuous tuning of the memory parameter. Finally, note that, in the navigation models analysed here, travellers choose at each junction (decision point) which direction to follow, based on the topology of the network and the available information. At variance with the more classical approach of calculating and storing routes, the navigation rules do not lay out routes, even dynamical, that travellers can follow. Each decision is in principle independent from the previous ones, but in practice is not. Indeed, the navigation rules tend to push travellers along minimal travel time routes, similarly to particles following an attractive potential to their own desti-
4.5. MODEL OF AGING AND PROPAGATING INFORMATION

Figure 4.20.: The normalized time to congestion $T_{\text{cong}}/T_{\text{max}}$ as a function of $\lambda$ for different values of the forcing rate $N_c$. The full lines indicate the expectation values, the dashed the standard deviation values.

4.5.2. Results

In Figures 4.19 and 4.20 we plot the time to emergence of congestion as a function of the arrival rate $N_c$, in the cases of constant forcing and of a daily forcing profile. We define the network as congested when the average queue length is constantly growing. When the travellers follow shortest path routing, the network does not develop congestion for driving rates $N_c < N_c^{\text{sp}} = 100$. Beyond that point, the time it takes for the system to become completely congested decreases steeply, implying that the system has become unstable and more prone to develop network-wide congestion. It can be seen that the threshold value for the dynamical infor-
4.5. MODEL OF AGING AND PROPAGATING INFORMATION

The information model is $\sim 105$ in the case of constant forcing and slightly higher ($\sim 110$) in the case of variable forcing. The little discrepancy is due to the small difference in the average value of the two forcing profiles (see section 4.5.1). Despite it however, in both cases we find the same qualitative result for the behaviour of $T_{\text{cong}}$ as a function of $N_c$. When the local congestion-aware model is considered instead, the maximum driving rate grows up to $\hat{N}_{\text{loc}} = 135$ ($\sim 140$ for the variable forcing), larger than the one for dynamical information. This is surprising, since it suggests that a traffic network under the limited information regime implied by the local congestion-aware routing performs better, with regard to congestion resilience, as compared to one where more information is available. Naturally, in the decentralised dynamical information routing model, the time to congestion depends also on the value of in a non trivial way. The examples of Figures 4.19 and 4.20 show only the case of $\lambda = 0.1$ and we already know from Eq. (4.28) that in the limit of $\lambda \gg 1$ the decentralised dynamical information routing converges to the local congestion-aware routing. It is interesting then to understand how the time to congestion varies for different values of the memory parameter. We plot the normalized time to congestion, $\frac{T_{\text{cong}}}{T_{\text{max}}}$, for the case of variable $N_c$ (modelled after a typical daily profile). There are a few interesting points to be made. The first one is that there exist a critical value $\lambda_c$ of the memory parameter, such that, for $\lambda \geq \lambda_c$ the system is stable ($T_{\text{cong}} \to T_{\text{max}}$, i.e. no congestion emerges). The second is that $\lambda_c$ appears to systematically shift toward larger values for larger forcing rates. This indeed was expected: in a system were new cars are introduced at a quicker pace, the network state is likely to change more rapidly and therefore old information needs to be removed accordingly. The information that is effectively used depends on the product $\lambda(t - t_{\text{up}})$, but the average inter-update $t - t_{\text{up}}$ time can only decrease for increasing forcing rate, as more cars with fresher information are introduced. Therefore, in order to retain, flexibility in the informational landscape, $\lambda$ needs to increase faster than the decrease of the update time. However, the fact that the system becomes stable for $\lambda \geq \lambda_c$ implies that the information spreading in itself reduces the performances of the dynamical routing in comparison to the local congestion-aware one. From this perspective, $\lambda_c$ plays the role of the control parameter in a phase transition between two qualitatively different states, stability and instability. Further evidence in
this direction comes from the behaviour of fluctuations of $T_{cong}$ for $\lambda \rightarrow \lambda_c$:

- for $N_c \geq 120$ the fluctuations of the time to congestion, which are vanishing in general, grow very steeply, becoming almost comparable with $T_{cong}$ itself. Borrowing from the language of statistical physics, a continuous order parameter with diverging (or large in finite systems) fluctuations is typically associated with a second-order phase transition.

- Interestingly, for lower forcing rates we do not observe large fluctuations; for example for $N_c = 110$, fluctuations vanish and at the same time $T_{cong}$ discontinuously shifts from 0 to $T_{max}$, closely resembling a first-order phase transition. Until now, we showed that the presence of travel time information can significantly alter the probability of emergence of congestion in a network. However, it is not clear yet which are the dynamical effects, induced by such information, that create instabilities in the system.

Previous works (see for example [23, 44, 185]) linked traffic heterogeneity with inefficient use of network capacity, resulting in increased instability and a higher probability for the emergence of congested states. Intuitively, the most heterogeneous states at a given traffic density are the ones where most of the traffic is concentrated on few links, while the others are scarcely used. The full links impair the outflow from links upstream and through spill-over congestion spreads upstream. On the contrary, a more uniform density landscape is less likely to produce the fluctuations needed to break down the flow. Hence, if for some values of $\lambda$ the system is indeed destabilised, there should be traces in the statistical properties of the queue distributions.

To test this hypothesis we consider the following quantities [44]:

\[ H = \langle \rho(t) \rangle^2 - \langle \rho(t) \rangle^2 \]  
\[ \sigma^2 = \langle \rho(t)^2 \rangle - \langle \rho(t) \rangle^2 \]

where $\langle \ldots \rangle$ stands for time average and $\cdots \cdots$ stands for spatial average, that is over all the links. $H$ accounts for the spatial fluctuations (from link to link) of the average population: large values characterise configurations where
few links have a heterogeneous load distributions. $\sigma^2$ instead measures both the spatial and temporal fluctuations: for example, a network with high traffic flow volatility will have a significantly larger $\sigma^2$ than one with a very constant flow.

For comparison, in figure 4.21 we plotted $H$ and $\sigma^2$ for the three cases $N_c = 110, 120, 130$. We observe that there is a direct connection between the stability/instability $T_{\text{cong}}$ threshold and the behaviour of $H$ and $\sigma^2$. Notice that for $\lambda \geq \lambda_c$, both quantities decrease rapidly. This means that the average density values on the different links become progressively more and more uniform, producing a flatter landscape. More interestingly, we also observe the reduction of the time fluctuations ($\sigma^2 \to 0$), which implies that the volatility of the single link densities are reduced: the system tends to a uniform state both in space and time. In our simulations, the travel time across one link is proportional to the population on the link. Configurations with large $H$ therefore correspond to situations where links on similar routes can statistically exhibit very different travel times and so, in the set of routes of a given length between two points, one encounters very
different expected travel times for different routes. Such state however is far from user-equilibrium (UE) and system optimum (SO). In the case of UE, this is because each traveller can single-handedly change his/her preferred route and decreasing his/her own expected travel time. In the case of SO, this can be seen from the comparison of the travel time distributions obtained for different values of $\lambda$. Figure 4.22 shows such an example. Here the forcing has been chosen small ($N_c = 105$), so that the system is stable for a large interval of $\lambda$. As the memory parameter is decreased from $\lambda = 0.2$ down to $0.06$, two things happen: first, the expectation value of the distributions shifts slowly toward larger values; second, and more important, the distributions become broader and the tails fatter. In other words, there are more travellers that, under the same forcing, take longer to complete their trips, thus steering the system away from UE. The presence of these tails is also the reason for the appearance of congestion at lower $N_c$ as compared to the local congestion-aware routing.
4.5.3. Discussion

The results above show a counter-intuitive, if not paradoxical, feature of decentralised information diffusion. In fact, we find again that access to more travel time information is likely to reduce the performances of a traffic network, on both the user and system level. A simple local congestion-aware routing (obtained as the limit \( \lambda \to \infty \)) outperforms the dynamical travel information routing for all values of \( \lambda \). The result is surprising as one would naively expect that conveying information about the congestion state upstream would improve the global network flow. It seems odd indeed that the only outcome of more information would be more congestion. The quantities we studied until now give a dynamical picture of the evolution of network traffic. However, they do not describe appropriately how the flow-density relation is modified by the different information mechanisms. In Figure 4.23 we plot the macroscopic fundamental diagram (MFD), the plot of network flow versus the network density, for the dynamical travel time information routing with different values of the parameter \( \lambda \), for shortest-path navigation and local information navigation. The plots shown refer to a driving rate \( N_c = 150 \), for which none of the navigation strategies studied can avoid the emergence of network-wide congestion. This is intended in order to probe the effects of the various routing mechanisms after the appearance of congestion. The density and the flow are normalised by the network capacity and maximal flow (per time step). Interestingly, the MFD curve for the local information shows a peak at a traffic density \( \rho_{peak}^{loc} \simeq 0.2 \), smaller than the one of the dynamical information navigation, \( \rho_{peak}^{dyn} \simeq 0.4 \) for \( \lambda \simeq 0.07 - 0.1 \). Also, in the same range of \( \lambda \), the maximum flow under the dynamical information mechanisms is the largest. As expected, for any further increase in \( \lambda \) the system approaches the behaviour of the local information navigation (see for the curve corresponding to \( \lambda = 0.5 \) in Fig. 4.23). We see then that the dynamical information mechanism is less able to avoid the emergence of congestion, but, once congestion has emerged, it expands the free flow region of the MFD. This happens because, through the long-range information dissemination, incoming travellers can steer away, at least for a time, from

\[ ^{8} \text{That is, when queues in a region of or all the network are growing. In other words, when the first derivative of the population, averaged over the network, is positive.} \]
the growing congestion front. This mechanism does not manage to relieve congestion completely, but can delay the network-wide onset for a time by automatically rerouting travellers away from the congestion front.

As a final remark, notice that we did not make any initial assumption regarding traffic equilibrium or fixed traffic densities. The simulated network was driven with various inflow rates in order to study the characteristics of the considered information regimes. With regard to congestion emergence, the transition from instability to stability is controlled by the memory parameter. This transition corresponds to a transition for the density field from a very heterogeneous configuration to a uniform one, both in time and space, that approaches the Wardrop equilibrium for the system. We obtain therefore a notion of stability of the system that depends on the spatial and temporal patterns of the density field. In the next chapter we will discuss recent work [23, 160, 162, 187] linking the network capacity
to the spatial variability of the traffic flow. Our results are consistent with these findings: if the memory parameter can be thought of as a feature of the network itself, rather than of the travellers, it is obvious that it should contribute to the determination of the network capacity. Hence, the transition point $\lambda_c$ can be seen as the point at which the capacity of the network exceeds the demand and so it does not develop congestion.

4.5.4. Conclusions

The aim of this section was to identify the effects of different agent-based information dissemination regimes on the performances of a dense urban traffic network. We compared the shortest path routing model with fixed route cost (no information), a local congestion-awareness model and a model based on dynamical spreading of information. As expected, shortest-path fixed cost (travel time) routing was outperformed by navigation mechanisms based on local or dynamical travel time information. However, we found that local congestion awareness improved the resilience to the development of congestion more than the mechanism in which information is dynamically sourced and disseminated. The former may be thought as causing the smallest deviation from the shortest path, while at the same time allowing travellers to avoid localised congestion when it is directly encountered. For $N_c > N_{c,loc}$ or equivalently large densities, when congestion is unavoidable, the dynamical information mechanisms allows to keep the traffic flow more fluid, as shown by the analysis of the fundamental diagrams. With regard to congestion emergence, we found a transition from instability to stability, controlled by the memory parameter. Interestingly, this transition corresponds to a transition for the density field from a very heterogeneous configuration to a temporally and spatially uniform one, that approximates well the Wardrop equilibrium for the system. Different information regimes generate different spatial patterns for the flow, which in turn determine the network capacity. Hence, the type of information dissemination can be thought as a property of the network itself, rather than of the travellers.

For real-world applications, our results suggest a twofold policy regarding information dissemination: for low densities and no congestion, local
information should be used (e.g. travel time along the next link after a decision point); when heavy congestion appears, dynamical information should be used in order to avoid (or at least delay as much as possible) the gridlock.
The chapter addressed the role of decentralised information dissemination in traffic networks [47, 286] by comparing four models of increasing complexity:

- a simple model of flow subject to local or global information constraints, without explicit navigation;
- a particle hopping model without explicit navigation, subject to a variable information horizon determined by a cascading process, which in turn was due to the network population;
- a model on realistic urban network, with explicit navigation (OD pairs and routing) subject to the same variable information horizon as in the previous point;
- a model of information propagation mediated by the travellers themselves, with explicit navigation and ageing of information.

The rationale for studying these models is the attempt to understand how information dissemination affects the performances (traffic demand, congestion emergence, etc.) of a traffic network, with particular attention to the case of dense urban network, i.e. networks with a large number of paths between pair of nodes. In particular, which are the factors that appear to be robust under changes in the details of the dissemination or navigation and thus are more likely to hold also in real-world applications.

A list of the main results for the different models follows:

1. In the flow model of section 4.2, both local and global information are able to delay the onset of congestion; however, global information induces much larger fluctuations in the populations of jammed sites than the local one, making it easier of the system to hit the (absorbing) state of full congestion.

2. The model of self-organised information propagation (section 4.3) showed that, in the case of particles hopping on a network, with an absorption rate for each jump, it is possible to find a signalling threshold (see Eq. 4.19) such that the system is able to maintain a low
population and effectively avoid congestion even under large rates of particle production; this mesoscopic information range outperforms local routing.

3. In section 4.4 the model studied was very similar to the one in section 4.3 but travellers navigated according to realistic rules from their origin to their destination on realistic networks; the optimum found in the previous section disappeared and non-local dissemination was more likely to produce congestion than the local information only.

4. This meant that the self-organised dissemination is able to effectively promote fluidity in the flow, by redirecting excess traffic to lowly congested areas, but not to guarantee an efficient tradeoff between such fluidity and the need to deliver travellers to their destinations.

5. When the information is taken away from the infrastructure and put in the hands of the travellers, to mediate and to transport, the general picture does not change: local information dissemination allows the network to sustain larger inflows without jamming. This is because it provides the minimum route elongation compatible with low congestion states.

6. When traffic demand exceeds the network capacity (which is increased by local information), non-local dissemination, mediated by the travellers and with limited memory, is able to effectively spread travellers across more routes and delay, even though not avoid, gridlock.

7. It is possible to tune the memory parameter $\lambda$ so that it guarantees effectively local information dissemination (best congestion resilience), but also allows for increased mobility when demand exceeds capacity and network-wide congestion is emerging (Fig. 4.23).

With reference to our initial question about the existence of an optimal information horizon (section 1), our results (in particular points 5 and 6) imply that for decentralised non-predictive systems local information dissemination is optimal to improve network capacity, while a moderate dissemination has mitigating effects when congestion is emerging.
In relation to previous work on dynamic traffic assignment, it would be interesting to compare the ageing information model with the anticipatory vehicle routing models near the stability-instability congestion of the former. Indeed, anticipatory mechanisms might provide a better sampling of the routes that cross or avoid the congested areas. However, fluctuations in the traffic field would affect the digital ants in the same way as they affect the car measurements in section 4.5. This would likely cause a thickening of the tails of the travel times distribution due to vehicle taking slightly longer routes due to noise in the information field. It is unclear then whether the gain from the improved sampling is able to compensate for this effect.

Moreover, work on network routing in the physics literature often focused on models where no explicit navigation is implemented (for example [190, 251]. Rather, particles disappeared from the network with a certain probability when they jumped from one node to the following one. This probability was used to mimic the chance that the particle would hit its "target" and be removed from the system. However, such a setting only measures how fluid the network traffic is, while it neglects whether particles actually reach a certain particular node or not. The comparison of the models in section 4.3 and 4.4 shows that a routing mechanism producing an optimum for the network for the first case (no explicit navigation and destination) does not necessarily produce similar results when particles need to reach a given destination. In other words, mobility is necessary but not sufficient to guarantee good network performances in realistic setups.

Finally, the results of section 4.5 link back to the analysis of London’s traffic. We showed that, for the values of λ which made the system unstable, the corresponding traffic configurations were very disordered (in space) and volatile (in time). Thus, the information field reflects strongly on the traffic configuration and, consistently with the results of chapter 3, an information field characterised by long memory produces very heterogenous traffic configurations, that eventually result in network-wide congestion.
5. Discussion and Conclusions

It is not incumbent upon you to finish the work, yet neither are you free to desist from it. 

---

Rabbi Tarfon

The main results of this thesis regard, on one side, the effects on network performance of various type of information dissemination, where the attention was mostly on the importance of the dissemination range, and, on the other side, the role of traffic heterogeneity and the potential presence of a mechanism for large scale coordination.

5.1. Macroscopic Fundamental Diagram and Urban Dynamics

5.1.1. Findings

The analysis of London’s traffic showed, by aggregating data over a month, that the rescaled traffic heterogeneity constrains the network to move on a well defined surface in the tridimensional space ($O, \Sigma_O, F$). In particular, we showed that the spatial heterogeneity plays a role as important as the network density itself in determining network performances.

We found also that London exhibits a daily anticlockwise hysteresis loop: a disordered, low-flow branch in the morning, followed by an ordered, high-flow unloading branch.

Although the MFD surface constrains the value of $F$, it does not define the evolution of the network state on it, which therefore depend on the details of origin-destination assignments.

Finally, while looking for dynamical traffic communities in the correlations network of sensors, we found that London, viewed as a spatially extended,
interacting system, displays the typical signs of a self-organised critical system. In the light of the relations between the scaling exponents of the correlation function and of the Taylor’s fluctuation scaling law, there features were interpreted as rooted in an underlying spatial synchronisation arising from the interaction between flows. [222].

5.1.2. Conclusions

Recent work in this topic focused on the conditions of existence for low scatter MFDs, which amounted to considering small neighbourhoods (a few square miles [23]) as reservoirs, or using data points characterised by the same spatial occupancy distribution. Against these findings, our results show that the scatter observed in urban MFD is the artefact of the two-dimensional projection on the \((O, F)\) plane of the full MFD surface. The request of uniform traffic distribution for a low scatter MFD (in the \((O, F)\) plane) [186] is equivalent to intersecting the MFD surface with a plane at constant \(\Sigma O\), which produces a well defined curve. In addition, note that our analysis did not require the full information about the occupancy distribution, but only on its first two moments. This is consistent with the presence of critical, and therefore long-range, correlations in the traffic fluctuations. Large scale coordination results in the system behaving as a single unit and, in our interpretation, provides the mechanism for the smoothing of the traffic heterogeneity, thus pulling the network up, along the MFD surface, toward the maximal flow states. At the same time, the induced network-wide correlations can become very dangerous when significant congestion emerges, e.g. as result of accidents or failures near important or central bottlenecks, because of the long correlation length.

Therefore, understanding the origin of such synchronisation is necessary in order to be able to destroy them when they become dangerous. Without data collected in absence of traffic control we cannot proceed further. However, there are a some considerations that can be made on theoretical bases regarding the role of the traffic control. Behaviours similar to the ones described above have been observed previously in simulations of lattice gas models [287], where there was no control and the dynamical properties emerged from the repulsive interactions between particles.
on adjacent sites. Hence, the fluctuations emerge from the queueing and spill-over between links across the city, independently of the traffic control.

On the other hand, London’s Urban Traffic Control currently divides the city in four regions for which the SCOOT system optimises traffic signal cycles. We found no trace of such regions in the flow correlation matrix, which would go against an active role of traffic control in the large scale coordination. SCOOT however belongs to the class of adaptive traffic control systems, which means it takes into account also traffic measurements coming from ILD loops. If the effect of these local measurements is the largest in the calculation of the traffic cycle times at junctions, then, one might expect the emergence of large-scale coordination behaviours, if not outright coordination (see for example [288] and references therein).

5.1.3. Implications and Future Work

Recent work on a utilisation-based derivation of the MFD [159] showed that it is possible to predict a theoretical MFD for regular networks using as independent variables the network occupancy and the number of delayed vehicles, which are seeds for congestions.

From this point of view, if it were possible to estimate this latter number from a measure of spatial heterogeneity, this would constitute a first theoretical model for the MFD surface.

Another important issue is the understanding of how the structure of the underlying physical street network affects the shape of the MFD surface, e.g. what makes it more robust against fluctuations in the spatial heterogeneity (i.e. a “flatter” surface along the $\Sigma_O$ direction).

Further data analysis is needed to confirm the presence of spatial synchronisation, especially whether it is a persistent phenomena over the whole year or when and how it disappears. If confirmed, however, it is extremely important to understand where the origin of synchronisation lies and, in particular, whether it can be controlled through a tuning of the traffic control systems. If this is the case, one could exploit this phenomenon in order to allow the traffic landscape to climb the MFD surface toward less hetero-
5.2. INFORMATION PROPAGATION AND ITS EFFECTS

geneous, higher flow states, and deactivate or at the very least reduce the effects when significant congestion emerges in order to hinder the spreading.

In the light of these considerations, traffic management policies should focus on strategies to reduce as much as possible the spatial heterogeneity, for example by encouraging discovery of new routes or even recurring to the closure of links if that improves the traffic homogeneity (see [237] for an example based on the price of anarchy). This consideration leads directly to the second main subject of the thesis.

5.2. Information Propagation and its Effects

5.2.1. Findings

The main results of the modelling effort of chapter 4 are listed below.

1. In a simple flow model (section 4.2) we found that both local and global information are able to delay the onset of congestion; global information regime produces larger fluctuations of jammed sites than the local regime;

2. For self-organised information propagation without explicit navigation, there exists an optimal value of the signalling threshold, such that the system is able to maintain a low population and avoid the emergence of congestion; this optimum disappears in the case of realistic networks with explicit navigation and local dissemination is preferable;

3. In the case of information dissemination mediated by the agents, the results do not change significantly; local information dissemination allows the network to sustain larger inflows without jamming. When congestion appears however, non-local dissemination (i.e. a smaller $\lambda$) with limited memory is able to promote mobility and delay, even though not avoid, the onset of gridlock.
5.2.2. Conclusions

In the light of these results, we can conclude that, for decentralised non-predictive information dissemination strategies, local dissemination regimes are more resilient than mid- and long-range dissemination strategies to the emergence of congestion.

In other words, under local information a network is able to sustain a larger inflow (demand) than under the other regimes. This is due to the fact that it causes the minimum stretch of the shortest-path route, compatible with the current congestion state of the network (Fig. 4.15).

At the same time, when significant congestion is present, long range information is able to delay gridlock: the information trickles down to travellers that are far from the congested area and allows them to choose a route which -at least partially- avoids the jammed area.

These considerations suggest a two-fold approach to dissemination in real networks: local under normal conditions and until the emergence of significant congestion, in order to exploit the higher demand sustainable by local routing; non-local, if traffic demand the local information limit and thus congestion is not avoidable, in order to exploit the delay in the onset of gridlock.

Contrasting the results of the two chapters, the results obtained for the various information regimes can be cast in terms of traffic disorder. We analysed different types of information regimes and showed that, in most realistic cases, non-local information propagation can deteriorate performances of urban networks in comparison to strictly local dissemination. This is due to the fact that long-range information dissemination can enhance the heterogeneity of traffic and, in so doing, lower the overall network performances (as seen in chapter 3). Under a constant inflow, this results in an increase of network density which leads to congestion in the network. This conclusion is supported by the results of our study of local and non-local models on abstract and realistic traffic networks (similarly to [47]) under different dissemination means (infrastructure or agents) and therefore appears to be robust to changes in the details (functional form, propagation method, etc.) of the models.

This robustness can be understood if we relate information to a perturba-
tion of the traffic landscape, and therefore within frame of the discussion on spatial disorder.

Consider a moderately congested traffic state (say 10AM in London), characterised by some heterogeneity. From the network point of view, the optimal situation would be to obtain a uniform density everywhere. Now, imagine that information is given to travellers about the low traffic areas. It is easy to see how this can result in herding effects, which would not improve the situation but only move the low traffic areas.

In this picture, local information produces the smallest ripple in the density field, while still allowing rerouting around congested links. With a physical analogy, it is promotes diffusive behaviour, as opposed to turbulent mixing potentially induced by long-range information (Fig. 4.21 and 4.23).

5.2.3. Implications and Future Work

From the point of view of traffic network management, our conclusions imply that methods such as TRC or similar long-range broadcast should be dedicated to major network congestion events, substituted by many information exchanges with short-lived memory on a short scale.

Further work is required to compare these results with the most recent anticipatory routing mechanisms in situations of strong congestion. Moreover, as mentioned in the motivation of this modelling effort, it will be important to study what happens when more populations of travellers with independent information sources compete for the same network. It was shown in fact that competition and cooperation in a spatial system with limited information horizons can produce non trivial organisation of the system [289]. In the case of traffic networks, however, there is the additional constraint of capacity. An interesting case in this direction, then would be to investigate the optimal information range for flocks of communicating traffic-aware vehicles, which share information within the flock, but not with members of different flocks.
Bibliography


Duncan J. Watts. *Proceedings of the National Academy of Sciences*.


A. Notes on self-organized propagation model

A.0.4. Topologies

In addition to the Manhattan section in Fig. 4.11, simulations were performed on Erdős-Rényi graphs, square lattices and a subsection of the San Francisco network presented (shown in Fig.A.1). In the case of San Francisco, capacities were allocated proportionally to the road length, as described in the main text. For the Erdős-Rényi the average degree was set at $\langle k \rangle \approx 4$ for comparison with the other networks studied. Each link was given a capacity $C = 30$ and outflow $S = 15$, which are compatible with realistic networks, as shown in Fig. 4.13a. Both Erdős-Rényi graphs and lattices showed much larger $N_c^{cong} \approx 60$. This was expected since they are more homogeneous than realistic networks both in degree distribution and capacity distributions and it is know that these properties allow to support more traffic (at the cost of speed of delivery)[261, 274]. The simulations results did not show significant differences from the ones presented for the Manhattan network.

A.0.5. OD matrices

Throughout the simulations, we used random Origin-Destination matrices for traffic assignment. However, we did simulate different regimes. As an example, in the case of Erdős-Rényi graphs, we performed simulations where cars received origin and destinations nodes with frequency proportional to the respective nodes. In this way a highly connected node would generate and be selected as destination by more cars than a low degree node. Results did not change qualitatively, although as expected the additional traffic on high degree nodes (which usually occupy central positions in networks) decreased the rate $N_c^{cong}$ at which congestion appeared.
A.0.6. Compliance

In equation (4.23) we introduced the parameter $c_A$, relative to the compliance of travellers to the information received. We performed the simulations under different regimes ($c_A = 1, 0.75, 0.5, 0.2$). As can be expected, when the compliance of travellers decreases, the threshold $\beta$ has to decrease too since the nodes need to be able to signal more to steer off the same number of travellers. Conversely when compliance increases, $\beta$ should too. As an example, consider figure A.2 obtained on a subsection of the Manhattan with $c_A = 1$ and compare it with Fig. 4.14 obtained with $c_A = 0.75$. The effective signaling areas is shifted to the right in Fig. A.2.
Figure A.1.: The total number of delivered cars $D_{tot}$ as function of the forcing rate $N_c$ and the signalling threshold $\beta$ for the San Francisco network section. The top plot refers to the non-local information spreading mechanism, the bottom plot to the local one.
Figure A.2: $N_c$-\(\beta\) plot for nonlocal (top) and local (bottom) information dissemination strategies for $c_A = 1$. 
B. Freeflow data

B.1. Correlation matrices

In the main text of section 3.4 for brevity we showed results only for the flow correlation networks. The occupancy matrices however reproduces the same results, but are much noisier, due to the smaller average correlation values and worse resolution on the occupancy measures as compared to flow. Figure B.1 shows the comparison of a detail of the correlation matrices $C^o$ and $C^f$ (53 sensors). The correlation values between corresponding elements in the two matrices is

![Correlation Matrices](image.png)

Figure B.1.: Details of $C^o$ (left) and $C^f$ (right) for comparison.

B.1.1. Modularity trick

Correlation matrices can contain negative elements, that is anticorrelated elements.

The modularity defined in section Eq. 2.9 does not explicitly take into account this and uncontrolled effects might result. Dealing with negative links has been a longstanding problem in the network community. Indeed, intuition says that one should maximise the number of positive links within communities and scatter the negative among as many communities as pos-
B.1. CORRELATION MATRICES

This certainly holds for social networks, where negative links usually represent enmities of various kinds. In the case of correlation matrices interpreted as network adjacency matrices, the question is more subtle. This is because anticorrelation is in itself a strong correlation between two signals. A trivial example would be two identical oscillators locked by some reciprocal interaction in opposed phases: from the point of view of networks, it would be very much the case to put them in the same community although they would be in perfect anticorrelation. To account for this, one could consider the absolute values of the $a_{ij}$ and study the modularity as usual. Another possibility would be to ignore the nodes characterised by a negative total strength or simply ignore all the negative links.

Finally, there exists a very recent method that accounts explicitly for the existence of negative links, by separately maximising modularity for the positive links and minimising for negative ones developed by Reichardt and Bornholdt \[57\] and generalised by Traag and Bruggeman \[290\]. Formally, the method is equivalent to finding the ground state of a spinglass system: the modularity matrix is mapped to the interaction term of the spinglass Hamiltonian and the optimisation is done through standard MonteCarlo methods.

The detection was repeated for all the cases and the various partitions compared using the normalised variation of information. The variation of information (VI) is an informational-theoretic measure representing the informational distance between two partitions. If larger VI, the more dissimilar the two partitions will be. The VI of a partition is normalised by its maximum value $\log_2 n$ which gives the normalised version ($\text{NVI} \in (0,1)$):

$$\text{NVI}(C,C') = \frac{\text{VI}(C,C')}{\log_2 n}$$

where

$$\text{VI}(C,C') = \mathcal{H}(C) + \mathcal{H}(C') - 2\mathcal{I}(C,C')$$

The terms in equation (B.2) are the entropy $\mathcal{H}(C)$ of partition $C$ and the mutual information between two partitions $C$ and $C'$:

$$\mathcal{H}(C) = -\sum_{i=1}^{k} P(i) \log_2 P(i) \quad \mathcal{I}(C,C') = \sum_{i=1}^{k} \sum_{j=1}^{l} P(i,j) \log_2 \frac{P(i,j)}{P(i)P(j)}$$
where \( P(i) = \frac{|C_i|}{n} \) is the probability that an element of \( T \) chosen at random belongs to community \( C_i \in \mathcal{C} \), and \( P(i,j) = \frac{|C_i \cap C_j|}{n} \) the probability that an element belongs to \( C_i \in \mathcal{C} \) and to \( C_j \in \mathcal{C}' \).

We found that all the partitions had \( NVI < 3\% \), which means that they are virtually identical and implies that the community structure results obtained are reliable.

To ensure robustness of the results, the community detection was also repeated with different implementation of the detection methods (Radatools [291], Louvain [290], iGraph library spinglass tool [292], Matlab software developed by Expert et al. [197]).

B.1.2. Geodesic and Euclidean Distances

In section 3.4 we used the geographical distance for the calculation of the correlation function, instead of a measure of driving distance between sensors. The result however is not changed by these metric because travel times and driving distances are proportional to their geographical distance, as shown in Figure B.2.

![Figure B.2.](image)

**Figure B.2.:** Travel time between couple of sensors (from the Google Maps Api) versus geographic distance, calculated from their coordinates. The line is a linear fit with slope \( \sim 410^{-3} \text{km/s} \) (adj-R\(^2\)=0.98).
B.2. COEFFICIENT OF MFD FITS

We report here in tables B.1 and B.2 the coefficient for the fits performed in section 3.3.2.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimated Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_2$</td>
<td>29.57 (27.28, 31.86)</td>
</tr>
<tr>
<td>$a_1$</td>
<td>-323.6 (-331.8, -315.4)</td>
</tr>
<tr>
<td>$a_0$</td>
<td>754.7 (748, 761.4)</td>
</tr>
</tbody>
</table>

Table B.1.: Coefficient of the quadratic polynomial fit of $F$ versus $\Sigma_O$.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{00}$</td>
<td>25.95 (5.532, 46.36)</td>
</tr>
<tr>
<td>$a_{10}$</td>
<td>175.7 (158.2, 193.2)</td>
</tr>
<tr>
<td>$a_{01}$</td>
<td>59.6 (58.12, 61.09)</td>
</tr>
<tr>
<td>$a_{20}$</td>
<td>-37.6 (41.36, -33.86)</td>
</tr>
<tr>
<td>$a_{11}$</td>
<td>-24.1 (-24.7, -23.51)</td>
</tr>
<tr>
<td>$a_{02}$</td>
<td>-0.9337 (-0.9665, -0.901)</td>
</tr>
</tbody>
</table>

Table B.2.: Coefficients of the quadratic polynomial in $O$ and $\Sigma_O$ fitting the Freeflow data.