Additional File 1
Supplementary Information

Complex delay dynamics on railway networks from universal laws to realistic modelling

Bernardo Monechi$^{1,2}$*, Pietro Gravino$^{2,3}$, Riccardo Di Clemente$^{4,5}$, and Vito D.P. Servedio$^6$

$^1$Institute for Scientific Interchange Foundation, Turin, Italy
$^2$SONY Computer Science Lab, Paris, France
$^3$“Sapienza” University of Rome, Roma, Italy
$^4$The Bartlett Centre for Advanced Spatial Analysis, University College London, London, United Kingdom
$^5$Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, USA
$^6$Complexity Science Hub Vienna, Vienna, Austria

*Corresponding author. E-mail address: mone.berna@gmail.com
1 Datasets Information

Novel information technologies enabled real-time monitoring and sharing of any kind of traffic data. Impressive instances are the visualised datasets about marine traffic that can be easily found on the Internet\(^1\). Also, several websites display live air traffic, by gathering and visualising official data from various sources\(^2\). These sources of information were found to be crucial in order to improve the understanding of the related transportation systems \([1, 2, 3]\). However, it is still a hard task to aggregate and analyse global or continental datasets about railway systems. In fact, due to historical reasons and to the typical usage scale, each nation has a network with few international connections and the available datasets are not homogeneous in coverage and format. Thus, tailoring the analysis on national systems seems a natural choice, whereas the most interesting characteristic behaviours appear in all systems, suggesting some kind of universality in the dynamics. We focused our analysis on the European continent both for the historical importance of railways and for the recent institutional efforts to raise their adoption. The actual railway system is composed by three distinct layers: high-speed passenger trains, normal-speed trains (mostly of regional type) and freight/military trains. While the freight trains use different stations and traffic handling rules (e.g., they operate mainly during night time) and can be discarded from our analysis, the other two layers can possibly interact each other. In the high-speed layer, correlations are identical to those in the regular layer, while no appreciable correlations can be spotted across the two layers so that considering them as independent is a fairly good approximation. Since no additional information can be gained by studying the whole system, we focus on the regular-speed layer only, both for Italy and Germany.

1.1 The Italian Dataset

The dataset regarding the Italian Railways has been collected by means of the “ViaggiaTreno” website\(^3\). The purpose of this website is to provide real-time information to travellers regarding the position of a certain train on the network, its delay and possible adverse occurrences like cancellations or strikes. Despite the fact that the information is in real-time, i.e., the instantaneous delay of a train can be checked at any time during the day, whenever the train arrives at its final destination its record is not deleted from the site. Instead, it is possible to check its route and its delay at each intermediate stop from the departing station to the arrival one until the end of the day, at 23:59. Hence, we downloaded all the information displayed on the website each day at 23:30 in order to be sure that each train would have arrived at destination. Starting from the 1st of January 2015 and for the whole 2015, we collected 12 months of historical data about the dynamics of regular and high-speed trains in Italy. For each train we get an identifier, the ordered list of stations the train has to cross, the scheduled arrival time at each station and its delay. The resulting dataset comprehends the traffic running on 2253 stations, with a daily average schedule pertaining 8112 trains on 7062 links. Note that “ViaggiaTreno” does not collect information about the geographical position of the stations. Such information has been integrated by means of Wikipedia and Google Maps, allowing us to represent the geo-localized network of Italian railways. Each dot corresponds to a station and each link corresponds to a route between two stations. In other words, a line between Rome and Naples means that there is a direct train route linking them without intermediate stops. Lacking real point-wise tracks data, the route has been simply represented with a straight line.

\(^1\)E.g.: \url{https://www.shipmap.org}, \url{https://www.marinetrack.info}, \url{https://www.vesselfinder.com}  
\(^2\)E.g.: \url{https://www.flightradar24.com}, \url{https://flightaware.com}, \url{https://planefinder.net}  
\(^3\)\url{http://www.viaggiaTreno.it}
1.2 The German Dataset

The data about German Railways have been collected through the OpenDataCity\footnote{https://www.opendatacity.de} website. This site gathers different datasets collected by a variety of on-going or terminated projects dealing with open data. In particular, the data we analysed come from the “Zugmonitor” project, which aimed at providing a web-app and an API to German travellers in order to have real time information on the position of the trains on the German Railway Network and their delay. The project is no-longer running and the API is not accessible anymore. However, some dumps of historical data collected during the project are still available. In particular, we downloaded all the data regarding year 2015, covering the same period of the Italian dataset. This dumps collected not only the delay at each station like in the Italian case, but also the delay at intermediate points between two stations. All the points are also geo-localized so that it is possible to reconstruct a quite accurate trajectory of the trains. In order to be consistent with the Italian dataset we used the geo-localization only to identify the position of the stations in the map. Scheduled arrival times at each station were also stored in the dump, so that in the end we managed to reconstruct a dataset with a structure identical to the Italian one. The resulting dataset includes data for 5979 stations with a daily average schedule containing 11,975 trains on 16,277.

2 High-speed layer

The structure Italian and the German Railway Networks is the overlap of two distinct layers, the normal-speed and the high-speed one. These two layers are different from the structural point of view. The high-speed layer in fact has to allow for fast travelling trains and have a different kind of rails connecting stations and, in general, it is reasonable to assume that high-speed trains and normal-speed ones do not interact when travelling from a station to another. However, the nodes of the network (i.e. stations) are shared between the layers, making the network a “multiplex”\cite{4,5} and allowing for interactions of the two different kind of trains.

Our datasets contain information about high-speed and normal-speed trains, for both the Italian and German case and in principle it could be possible to study the dynamics of the high-speed layer and its interaction with the normal one. In the main text we decided though to focus on the normal layer cutting-out the high-speed part. This choice was made for sake of simplicity, since the rules of interaction between the layers might have been hard to understand or derive with data analysis. Here, we will show that this approximation is reasonable due to the smaller numbers of the high-speed trains travelling and their poor effects on the dynamics of the normal-speed one.

In our datasets it is possible to identify high-speed trains thanks a specific identifier (“ES*” for the Italian Network and “EC”, “IC”, “ICE” for the German Network) and use them to build the High-speed layer of the Railway Network in a similar way that has been done for the normal layer in the main text. The number of travelling high-speed trains per day is considerably smaller with respect to the normal-speed trains, being of $\sim 210$ and $\sim 1055$ for the Italian and German case respectively. As a consequence also the two networks are smaller compared to the normal-speed ones as shown in table\footnote{Table1}. The smaller number of nodes indicates the fact that high-speed trains usually connects fewer, more important and distant stations, since it is used mainly for mid-long range movements. This is also reflect in the distribution of the length of the links in the network (Fig.\footnote{Fig2}), showing a tail which is considerably longer with respect to the normal-speed layer. Other topological properties are similar in the two case, like the degree distribution (Fig.\footnote{Fig1}) and the associativity coefficient (table\footnote{Table1}). As an example of the dynamics taking place over the high-speed layer, we show in Fig.\footnote{Fig3} the distribution of the final (positive) delays of high-speed and normal-speed trains. We can see that both for the Italian and German case, the distributions are fat tailed, so that also trains on the high-speed layer can experience large delays and major disruptions. As a final remark, we validated the approximation of neglecting
Figure S1: Degree distributions for the Italian (left) and German (right) Railway Networks. Continuous lines correspond to the normal-speed layers, while dotted lines correspond to the high-speed layers.

Figure S2: Links Length distributions for the Italian (left) and German (right) Railway Networks. Continuous lines correspond to the normal-speed layers, while dotted lines correspond to the high-speed layers.
Table 1: Network Metrics of the High-Speed Layers of the Italian and German Railway Networks.

<table>
<thead>
<tr>
<th></th>
<th>Italian HS Layer</th>
<th>German HS Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Nodes</td>
<td>162</td>
<td>712</td>
</tr>
<tr>
<td>Number of Links</td>
<td>509</td>
<td>2281</td>
</tr>
<tr>
<td>Avg. Degree</td>
<td>6.28</td>
<td>6.41</td>
</tr>
<tr>
<td>Degree Assortativity</td>
<td>0.06</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Figure S3: Final delay distributions for the Italian (left) and German (right) Railway Networks. Continuous lines correspond to the normal-speed layers, while dotted lines correspond to the high-speed layers.

this layer by looking at the cross-correlations between the time-series of average delays on the links in the networks, similarly to what we have done for the normal-speed layer in the main text. In this case though we checked for correlations not just between the links of the same layer, but also between couples of links from different layers in order to see whether we can spot a signal of a possible inter-layer interaction. Fig. shows the cross-correlations in the “Forward”, “Backward” configurations, between the couples of links of the high-speed layer and the couples of links made by a link in the high-speed layer and one in the normal-speed layer. As for the normal-speed layer, decaying correlations exist for non inter-layer couples of links in the Backward configuration, while in all the other cases the signal of correlation is very close to 0. Hence, it is possible to considered the high-speed and normal-speed layer as independent and non-interacting. It is worth noticing that this measure of correlation might hide possible local interaction effects due to the fact that it is an aggregation of all the couples of links in the network. Such approximation will be then valid when considering global or aggregated metrics (e.g. the delay distributions), but it is not unlikely that more “fine-grained” observations (e.g. the distribution of delays on a single link or station) might be influenced by our choice.
Figure S4: Cross-Correlations between the average delay time series of pairs of nearby edges in the high-speed layers of the railway networks and between the time series of pairs of edges coming from different layers. Decaying correlations are observed only in the “Backward” case for pairs of edges both in the high-speed layer. No signal of inter-layer correlations can be observed.

3 Exogenous Delay Distributions

The most trivial way to group the links of the railway networks is according to the geodesic distance between the stations they connect, behind this a rough estimate of the length of the railway between them. Fig. 5 show the distribution of these distances $d(e)$ for all the edges $e$ in the Italian and German Railway Networks. From these distributions we can see that the distances are distributed around a typical value of $\sim 5$ km, but then span with a long tail until $\sim 100$ km. In order to characterize correctly the exogenous delay on the links, we measured the positive and negative exogenous delay distributions aggregating the links according to $d(e)$ as can be seen from Figures 6, 7, 8, and 9. In all these cases, we modelled the distribution using a $q$-exponential functional form $[6, 7]$: 

$$e_{q,b}(\delta t) \propto (1 + b(q - 1)\delta t)^{1/(1-q)}, \quad q \in [1, 2], \quad b > 0,$$

(S1)

so that in these cases the parameter $q$ and $b$ are depending on $d(e)$. The behavior of the parameters with respect to $d(e)$ are shown in the main text. We find that in general:

$$q(d) = \text{const}, \quad b(d) = Ad^{-a}.$$

(S2)
Figure S6: Distributions of the positive exogenous delays according to the length $d(e)$ of the links in the Italian Railway Network. Dotted lines represents the $q$-exponential fit of the distribution. The parameters obtained with the fits are shown in the legend.

Figure S7: Distributions of the negative exogenous delays according to the length $d(e)$ of the links in the Italian Railway Network. Dotted lines represents the $q$-exponential fit of the distribution. The parameters obtained with the fits are shown in the legend.

Figure S8: Distributions of the positive exogenous delays according to the length $d(e)$ of the links in the German Railway Network. Dotted lines represents the $q$-exponential fit of the distribution. The parameters obtained with the fits are shown in the legend.
Figure S9: Distributions of the negative exogenous delays according to the length \(d(e)\) of the links in the German Railway Network. Dotted lines represents the \(q\)-exponential fit of the distribution. The parameters obtained with the fits are shown in the legend. The parameters for equation (2) can be found in table 2. Since these distributions are all conditioned on the fact that the acquired exogenous delay is either positive or negative, we can check whether the probability of these conditions are influenced or not by the length of the link the train is travelling on. Fig. 10 shows these dependencies for both the considered Railway Networks. Despite the fact that a small dependence can be observed in the probability of having positive delays (i.e. it is slightly increasing with \(d\)), assuming that such probabilities are constant is a good zero-order approximation that we have used in the main text.

<table>
<thead>
<tr>
<th></th>
<th>(q)</th>
<th>(A)</th>
<th>(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITA positive</td>
<td>1.15</td>
<td>0.66</td>
<td>5.25</td>
</tr>
<tr>
<td>ITA negative</td>
<td>1.03</td>
<td>0.22</td>
<td>1.33</td>
</tr>
<tr>
<td>GER positive</td>
<td>1.28</td>
<td>0.99</td>
<td>37.5</td>
</tr>
<tr>
<td>GER negative</td>
<td>1.15</td>
<td>0.57</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 2: Parameters for the equation (2), governing the behavior of the parameters \(q\) and \(b\) of the \(q\)-exponential distribution as the links length \(d(e)\) varies.

Figure S10: Probability of having a positive (dashed line) and negative (continuous line) exogenous delay as function of the length of the links in the Italian (left) and German (right) Railway Networks.
4 Departure Delay Distributions

An approach similar to the one adopted for the exogenous delays on the links of the networks can also be adopted for the departure delays at the stations. In this case we categorized the departing stations (i.e. a subset of the nodes in the network) according to their out-degree $k_{out}$ whose distributions are shown in Fig. [1]. Having divided the nodes of the network according to $k_{out}$, we can fit the aggregated departure delay distributions as $k_{out}$ varies as shown in Fig. [13, 12 and 14]. Note that negative departure delays are not present in the German dataset. These distributions have been fitted using a $q$-exponential functional form as in equation [1]. The behavior of the $q$ and $b$ parameters for these distributions according to $k_{out}$ can be summarized by the equations:

$$q(k_{out}) = \text{conts}, \quad b(k_{out}) = Ae^{-\alpha k_{out}}.$$  

The parameters for equations [3] are obtained by fitting the empirical data as shown in the main text. Table [3] shows the values obtained with the fit: To conclude the investigation about departure delays, it is necessary to study the occurrences of positive and negative ones as $k_{out}$ varies. Fig. [15] shows these dependencies for the Italian and German Railway Networks. Similarly to what we have found for
Figure S13: Distributions of the negative departure delays according to $k_{out}$ of the links in the Italian Railway Network. Dotted lines represents the $q$-exponential fit of the distribution. The parameters obtained with the fits are shown in the legend.

Figure S14: Distributions of the positive departure delays according to $k_{out}$ of the links in the German Railway Network. Dotted lines represents the $q$-exponential fit of the distribution. The parameters obtained with the fits are shown in the legend.
Table 3: Parameters for the equation (3), governing the behavior of the parameters $q$ and $b$ of the $q$-exponential distribution as out-degree $k_{out}$ of the nodes varies.

<table>
<thead>
<tr>
<th></th>
<th>$q$</th>
<th>$A$</th>
<th>$a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITA positive</td>
<td>1.28</td>
<td>0.014</td>
<td>1.87</td>
</tr>
<tr>
<td>ITA negative</td>
<td>1.01</td>
<td>0.004</td>
<td>0.99</td>
</tr>
<tr>
<td>GER positive</td>
<td>1.20</td>
<td>0.026</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Figure S15: Probability of having a positive (dashed line) and negative (continuous line) departure delay as function of the out-degree of the nodes in the Italian (left) and German (right) Railway Networks. Dashed and continuous lines corresponds to linear or constant fits of the data. We assume that there is no dependency with $k_{out}$ in every case but the positive departure delays in the Italian data where we found the linear relation: $P = ak_{out} + b$ (parameters shown in the legend).

the dependency of the exogenous delay with the length of the links, in the German Railway Network no dependence can be observed and the probabilities of having a positive or negative departure delay can be considered constant in every station. However, this is not true for the Italian Railway Network where just the probability of having a negative delay can be considered constant. On the contrary, the probability of having a positive departure delay increases linearly with $k_{out}$. 

11
5 Optimal Choice of $\beta$

In order to determine the optimal value of the $\beta$, we tune our model to reproduce with the highest probability the delay that each train gets whenever it crosses a station during its path. Considering a train $i$ arriving at a station $n$ on a given day, we call $\delta t_{i,n}^{\text{emp}}$ its measured arrival delay at that station as recorded in the dataset. Hence, we perform 200 simulations of the schedule of the considered day in order to compute the distribution $P(\delta t_{i,n})$ of the corresponding $\delta t_{i,n}$. Hence, considering the null hypothesis that $\delta t_{i,n}^{\text{emp}}$ is produced by our model (i.e. it is extracted by $P(\delta t_{i,n})$), we calculate the double tailed $p$-value for such hypothesis for every pair $(i, n)$, i.e. for every train and for every station. For each day we average the $p$-values of all the train-station pairs to obtain the performance metrics for $\beta$. Fig. 16 shows the dependence of the average $p$-value as a function of $\beta$. The curves have been computed from the simulation of a week of daily schedules. The values of $\beta = 0.15$ for Italy and $\beta = 0.10$ for Germany allows the model to maximize the probability of reproducing the correct arrival delay for each train at a particular station. These values will be used in all the numerical simulations in the following, if not otherwise specified.

Figure S16: Average $p$-value statistics for the arrival delay for each train at each station in their route as a function of the diffusion parameter $\beta$. The curves have been obtained as the average of the single curve of different days (from the 1st to the 6th of March 2015). The highlighted regions correspond to the range of values of $\beta$ where the average maximum $p$-value has been observed;
6 Predictive limits of the model

By means the definition of the $p$-value of the previous paragraph, we can explore a bit which are the predictive limits of the model, i.e. in which part of the Railway Network it is more likely to not reproduce correctly the delays. In other words, we computed for each station $n$ the average $p$-values, by averaging all the $p$-values assigned to the couple $(i, n)$. Fig. [17] shows the distribution of these average $p$-values for the stations in the networks, obtained with the optimal value of $\beta$. The largest parts of the stations have $p$-values centred around a typical value of $\sim 0.6$ for Italy and $\sim 0.7$ for Germany, yet there is a large percentage ($\sim 11\%$ for the Italian and $\sim 20\%$ for the German case) with a $p$-value smaller than $0.05$. In these stations the predictive power of the model is particularly unsatisfactory and it is interesting to understand something about their features.

Fig. [18] shows the distribution of the average weight of the links (in terms of the number of trains that have travelled in the links in our whole datasets) and the distribution of the length of the links connected to a station, in the case of stations with $p$-value larger and smaller than $0.05$. We can see that in the latter, the distribution of the weight is considerably narrower, indicating that in these stations the traffic is usually low. Hence, the disagreement might be the result of a poor sampling of the exogenous disturbances around these stations leading to poor predictions or to a dependence of the transmission parameter $\beta$ on the traffic conditions that have been ruled out when we assumed them to be constant in time and uniform all over the network.

For the Italian case, these stations also are usually connected to links with a shorter distance. This fact can be interpreted as the effect of discrepancies in the fitted exogenous delay model when the links are not sufficiently long.

![Figure S17](image-url)
Figure S18: Distributions of the average weight of the in-going (A-E) and out-going (B-F) of the links of the Railway Networks. Weights are computed as the total number of trains that have travelled over a link during the whole 2015. Distributions of the average length of the of the in-going (C-G) and out-going (D-H) of the links of the Railway Networks. Red and Blue lines correspond to the Italian and German railway networks, dotted lines are the distributions for the stations with a p-value smaller than 0.05.
7 Funding

This work has been supported by the KREYON Project, funded by the John Templeton Foundation under contract n. 51663; VDPS acknowledges financial support from the Austrian Research Promotion Agency FFG under grant #857136. RDC as Newton International Fellow of the Royal Society acknowledges support from The Royal Society, The British Academy and the Academy of Medical Sciences (Newton International Fellowship, NF170505).

References


