How good are network centrality measures? Longitudinal analysis of traffic in a railway network in the United States

SATYAM MUKHERJEE

Department of Operations Management, Quantitative Methods & Information Systems; Indian Institute of Management, Udaipur; and Research Center for Open Digital Innovation, Purdue University, IN 47906, USA
E-mail: satyam.mukherjee@iimu.ac.in

Abstract. Transportation networks are plagued by frequent delays, which pose crucial challenges to the economic advancement of a country. In previous studies, transportation networks have been envisioned as network structures of vertices with links representing a connection between two vertices. In the present work, the focus is on the delay analysis using tools of complex network analysis of the Amtrak railway network in the US. The Amtrak delay data for a period of 6 years between 2009 and 2014 was analyzed. It was observed that delay distribution varied every year and also for same stations in different years. The effect of conventional network measures on the average daily departure delay at various stations and also on the total daily traffic load on the station were estimated. Contrary to the predictive power of network topology measures in model transportation networks, it was observed that the topology measures had a negligible effect on Amtrak delay, with stations that faced the highest traffic experiencing shorter average delays. The results of this study call for additional realistic network measures and routing schemes, which could capture the features of the real-world transportation networks.

Keywords. Amtrak railways; complex network; delay propagation.

PACS Nos 89.40.-a

1. Introduction

Transportation networks have a strong impact on the economic development of a region [1]. In recent years, transportation networks have been examined from the perspective of complex network analysis [2, 3]. These include road, maritime, public transportation, and airport networks. While some studies have focused on modeling the transportation networks, others have studied real-world transportation networks that include the worldwide air transportation network [4, 5], airport network of China [6], airport network of India, urban road networks [7] and railway networks in India [8–10]. Several different models have been used to analyze transportation networks like the railway network of China [11] and the underground railway networks of Boston and Vienna [12]. These studies give valuable insight into the topological properties of the networks.

All such networks are prone to congestion and traffic delay. A vast amount of research has been devoted to modeling traffic flow and optimizing transport processes in complex networks. For example, earlier studies

have modeled the spread of disease by using airline transportation networks [13]. Delay propagation in the US air transportation network has been studied using a data-driven approach [14, 15]. Although there exist many pieces of research on modeling delay in freight trains [16], in urban networks [17] and in surface transportation networks [18], real-life traffic patterns in the railway transportation networks has been ignored.

The United States railway network is fast emerging as an influential transportation network by facilitating the mobility of people and goods. Amtrak is the only high-speed intercity passenger rail provider of the US, covering more than 21,000 miles and almost 500 destinations in 46 states, the District of Columbia and three Canadian provinces. In the fiscal year, 2011 Amtrak set a new record of carrying nearly 30 million passengers on more than 300 daily intercity trains. The demand for rail transport in the US is projected to increase in the coming decades, with significant consequences in economic development. In the fiscal year 2011, the US Department of Transport granted Amtrak \$450 million to begin the first phase of high-speed rail service between New Brunswick and Trenton,

NJ, which is planned for completion in 2017. Despite attempts to improve the inter-city passenger transport, the railway network is continuously plagued by endemic delays.

All these facts serve as a motivation to characterize the delay dynamics of the railway transportation network from a complex network perspective. The Amtrak railway network in the US was analyzed as a network of stations. Two stations are connected if there exists a physical track connecting the two stations, with the links corresponding to connections within one stop distance. As of 2014, the Amtrak network consisted of ~500 nodes and 1571 edges. The effect of network measures on the traffic load and delay in the stations was predicted. A community analysis of the network was performed and based on the intercommunity and intra-community links, the role of each city was defined. Additionally, the effect of community structure on traffic delay and load was also studied. Whether neighbors of the *connector hubs*—the stations occupying the central position in the network—see significant higher traffic and greater delay than other nodes in the network was also checked.

2. Data and method

Data as obtained from Amtrak status maps available in the dixielandsoftware web page. The web page tracks the position of the trains as described in the Amtrak's web page. This database stores the scheduled departure time, actual departure time, and train number of each train stopping at a particular station. The station names and station codes (as specified by Amtrak) are also mentioned. This information of all trains and stations for years between 2006 and 2014 was extracted. Since Amtrak only shares the information of arrival delays for source and destination stations, the information of scheduled and actual arrival times was ignored. This data is available upon request.

2.1 Network representation

To articulate the network approach to the Amtrak railway network in the US, the methodology of generating the network of train stations is outlined here. Two train stations in the US are linked if there exists a direct Amtrak link between the pair.

2.2 Weighted degree

The degree of a node is defined as the number of nodes it is directly linked to. The degree distribution of the Amtrak network reflects the topology of the network.

The weighted degree (k_i^w) of the node given by the equation was evaluated:

$$K_i^w = \sum_{j=1}^N \alpha_{ij} w_{ij},$$

where w_{ij} is the number of trains passing between stations i and j , and $\sum_{j=1}^N \alpha_{ji}$ is the vertex degree k_i .

2.3 Betweenness centrality

Betweenness centrality measures the extent to which one node is between two other nodes that are not connected to each other. It is defined as

$$C_g^w(i) = \frac{g_{ji}^w(i)}{g_{ji}^w},$$

where w is the weight of link between two nodes j and l , g_{jl} is the number of shortest paths between two nodes and $g_{ji}(i)$ is the number of shortest paths that pass through node i [19].

2.4 k -Shell decomposition

k -core of a network is defined as the maximal sub-graph of the network with at least degree k . The k -shell of the network is defined as the set of all nodes that belong to the k -core but not the $k+1$ core. The algorithm proposed by Carmi *et al.* [20] to measure the k -shell index of the stations in Amtrak railway network was used. First, the degree of each node of the entire network was evaluated. Next, all nodes with degree 1 were removed to get the k -shell index $k = 1$, and the nodes with degree 2 were recursively removed to get k -shell index $k = 2$, and so on. The k -shell index is a measure of embeddedness of a network.

3. Community structure and roles

In the context of complex networks, communities are groups of nodes that more densely packed to one another as compared to the remaining nodes of the network. Community detection algorithms aim to identify functional subunits of networks and have been extensively studied in the last few years [29]. In the context of the Amtrak network, it becomes important to classify the stations according to their positions within the community. The knowledge of node position provides us information about the role played by each station. For example, stations present in the core of a cluster play a role different from stations located at the periphery.

The community structure of the Amtrak network for different years between 2009 and 2014 was analyzed via the random walk algorithm proposed by Rosval and Bergstrom [21].

Post obtaining the community structure of the network, the nodes were classified according to the role they play both within and outside their community as proposed by Guimera and Amaral [22]. First, hubs and non-hubs were differentiated. The within community degree z -score is given by the following equation:

$$z_i = \frac{k_i - k_{s_i}}{\sigma_{k_{s_i}}},$$

where k_i is the number of links of node i to other nodes in its community, s_i is the average of overall nodes in s_i , and $\sigma_{k_{s_i}}$ is the standard deviation of k in s_i . If $z > 2.5$, the nodes are classified as hubs, while non-hubs are identified with $z < 2.5$. The hubs and non-hubs are further classified based on the participation coefficient P , defined as

$$p_i = 1 - \sum_{s=1}^M \frac{K_{i,s}^2}{k_i^2},$$

where k_{is} is the number of links of node i to nodes in community s , and k_i is the degree of the node i . The *connector hubs*—the hub nodes with many connections to most of other communities ($0.30 < P < 0.36$) [22]—were identified.

4. Results

The Amtrak network has been increasing in size since 2006, with the number of station doubling in 2007 and

Figure 1. The growth of the Amtrak railway since 2006. The number of railway stations increased almost two-fold in 2007 as compared to 2006. Since 2009 the growth appears to attain a steady value.

Figure 2. Cumulative degree distribution of complex network of the Amtrak railway for the years 2006–2014.

attaining a steady size of approximately 500 stations in the year 2011 (figure 1). The numbers of stations have been increasing steadily, with the number of stations in 2007 being twice that 2006.

For simplicity, a symmetric Amtrak network was considered. Next, the network topology measures of the network for every year between 2006 and 2014 was evaluated.

First, the degree of a station in the Amtrak network was constructed. The degree of a station is defined as the number of links connecting it with other stations. The complementary cumulative degree distributions are plotted in figure 2.

Next, the traffic on each station was determined by calculating the number of trains passing through a given station (figure 3). It was observed that for the year 2006, the number of trains is very less and the cumulative distribution falls off rapidly. The number of trains increases slowly through 2010, as indicated by the faster decay

Figure 3. The cumulative distribution of traffic at the Amtrak stations. Traffic is defined as the number of trains passing through a station.

Figure 4. The cumulative distribution of average daily departure delay for the years 2006–2014.

of cumulative distribution. Since 2011, the cumulative distribution of the number of trains decays slowly, indicating the increase in number of new train routes and trains.

Next, the departure delay at various stations for years between 2006 and 2014 was studied. It was noticed that few stations showed negative departure delay, indicating departures earlier than scheduled departure (figure 4). Stations that consistently showed negative departure delays include Alexandria, Virginia (ALX), Milwaukee, Wisconsin (MKE), West Palm Beach, Florida (WPB), Pomona, California (POS) and Wilmington, Delaware (WIL). The distribution is broadest in the year 2006. With time, the distribution falls off rapidly, suggesting an improvement in the on-time performance of Amtrak railways. In order to comprehend the effect of network parameters on the departure delay at various stations, the stations with the highest

Figure 5. Stations with the highest weighted degree (marked in parentheses). The cumulative distribution of average departure delay at stations with highest weighted degree is shown.

Figure 6. Stations with highest betweenness centrality (marked in parentheses). The cumulative distribution of average departure delay at stations with highest betweenness centrality is shown.

weighted degree and highest betweenness centrality were identified. In figure 5, it can be seen that Washington, D. C. Union Station (WAS) emerges as the station with the highest weighted degree consecutively for 2006 and 2007. In the year 2008, Chicago Union Station (CHI) had the highest weighted degree. Between 2009 and 2014, Philadelphia, Pennsylvania – 30th Street Station (PHL), is the node with the highest weighted degree. The average daily distribution is much broader for WAS as compared to CHI and PHL, perhaps indicating an efficient performance in the Northeast corridor. Interestingly, CHI emerged as the most central station in 2008 and also between 2009 and 2014. Other centralized stations were Raleigh, North Carolina (RGH) in 2006, WAS in 2007 and New Orleans, Louisiana (NOL) in 2009. A rapid fall of the delay distribution of NOL in 2009 as compared to other stations with highest betweenness centrality in different years (figure 6) was observed.

A visual exploration of delay and traffic in the US map reveals that stations that face the highest traffic are not necessarily the ones that experience highest departure delays (figure 7). Stations in the Northeast corridor experience larger traffic but, interestingly, least delays. Also, Amtrak services outside of the Northeast corridor exhibit comparatively larger delays. The trend is consistent from 2009 through 2014, perhaps indicating a recurring pattern of delay propagation. On the contrary, research on model communication networks assumes nodes with the highest traffic as congestion points. These results suggest that network topology measures fail to reproduce the real-life traffic patterns.

To assess the association between network topology measures and average daily departure delay, a

Figure 7. Delay and traffic in the US map between 2009 and 2014. The size of the nodes is proportional to the traffic at a station. The color code is according to the average delay (in minutes) for a given year.

linear regression model of the following form was performed:

$$DD_i = \alpha_0 + \alpha_1 k_i^w + \alpha_2 B_i + \alpha_3 S_i + \alpha_4 \delta_i^R + \sum_t \beta_t Date_{it},$$

where α_{1-4} are the coefficients we want to estimate, DD_i is the average daily departure delay at station i , k_i^w is the weighted degree of a station i , B is the betweenness centrality of the station, S represents the k -shell index of the station and δ_i^R is a dummy variable that takes value

1 if a station i is a neighbor of connector hubs and is 0 otherwise. The effect of network topology measures on the daily traffic seen by a station was also accessed. The following linear regression was modeled as

$$N_i^T = \alpha_0 + \alpha_1 k_i^w + \alpha_2 B_i + \alpha_3 S_i + \alpha_4 \delta_i^R + \sum_t \beta_t Date_{it},$$

where N_i^T is defined as the number of trains passing through a station i . In both the regressions, the full set of fixed effects for each of dates was taken, where the indicator variables $[(Date)]_{it} \in \{0, 1\}$ are equal to 1 if a train passes through the station i on the date t , while β_t are the coefficients of the indicator variables. Additionally, the explanatory variables were standardized in order to compare the relative strength of the explanatory power of the weighted degree, betweenness centrality, k -core number and neighbor of connector hub.

The explanatory power of the four network topology measures are presented in tables 1 and 2 by estimating the strength and significance of coefficients α_{1-4} . In table 1 the estimates of the linear model on the number of trains (N_i^T) are presented. In Model 1 it was observed that betweenness centrality had a positively significant effect on the number of trains, although the effect-size was low. A one-standard deviation (SD) increase in betweenness centrality yields a 0.085 SD increase in the number of trains. In Models 2–4 the effects of the k -shell index, the neighbor of connector

Table 1. Results of the linear model used for predicting the effect of network parameters in the Amtrak railway network, on number of trains passing through a station including date fixed effects in the year 2014 (***) $p < 0.001$.

	β (SE)					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Betweenness centrality	0.085*** (0.003)				0.086*** (0.002)	0.018*** (0.001)
k -Shell index		0.542*** (0.004)			0.431*** (0.003)	−0.006*** (0.001)
Neighbors of R6			0.452*** (0.005)		0.258*** (0.004)	0.009*** (0.001)
Weighted degree				0.966*** (0.003)		0.964*** (0.003)
Controls: Date fixed effects					Yes	Yes
R ²	0.007	0.29	0.20	0.933	0.36	0.935
BIC					1236947	847220
N	170,233	170,233	170,233	170,233	170,233	170,233

Table 2. Results for the linear model used for predicting the effect of network parameters in the Amtrak railway network on the average daily departure delay at a station including date fixed effects in the year 2014 (***) $p < 0.001$.

	β (SE)					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Betweenness centrality	0.089*** (0.003)				0.084*** (0.003)	0.094*** (0.003)
k -Shell index		−0.209*** (0.001)			−0.188*** (0.002)	−0.121*** (0.002)
Neighbors of R6			−0.121*** (0.001)		−0.042*** (0.001)	−0.004*** (0.001)
Weighted degree				−0.213*** (0.002)		−0.15*** (0.002)
Controls: Date fixed effects					Yes	Yes
R^2	0.007	0.044	0.014	0.045	0.093	0.10
BIC					1872755	1870174
N	170,233	170,233	170,233	170,233	170,233	170,233

hub R6 and weighted degree on the dependent variable (N_i^T) is presented. It was observed that weighted degree had the strongest effect among all the explanatory variables—a one-standard deviation (SD) increase in weighted degree results in 0.966 SD increase in the number of trains. In Model 5, all the explanatory variables except weighted degree, but including data fixed effects, were included. In Model 6, all the explanatory variables were added and the relative strength of weighted degree were compared while accounting for the other three variables. Model 6 provides evidence of an intricate relationship between the traffic on a station (defined as the number of trains passing through the station) and weighted degree of the station for the year 2014. The weighted degree is significantly (***) $p < 0.001$ and positively associated with the traffic at a station (N_i^T), as indicated by the positive coefficient (table 1). The standard R^2 and Bayesian Information Criterion (BIC) statistics was used to interpret the improvement in fit of the regression model when k_i^w was added to Model 5. A drop in the BIC statistic of more than 50 provides “strong” evidence that the variable (weighted degree) added to Model 5 improves the fit of the overall model [30]. An increase in R^2 from 0.36 to 0.935 when k_i^w is added to Model 5 (table 1) was also observed. Table 2 shows the effect of the network parameters on the average daily departure delay (DD_i). In Models, 1–4 the individual explanatory power of betweenness centrality, k -shell index, neighbor of connector hub R6 and weighted degree were estimated. It is interesting

to note that, while in Model 1, betweenness centrality has a weak positive effect on average daily departure delay ($p < 0.001$), the remaining three variables have a negative effect on the average daily departure delay ($p < 0.001$). In Model 5, the betweenness centrality, k -shell index, and neighbor of R6 and control were included for date fixed effects. In Model 6, the weighted degree was added to Model 5. When comparing the R^2 in Model 5 and Model 6, it was observed that they predict only 10% of the observed variance in the data. The negative coefficient of weighted degree signifies that a one-standard deviation (SD) increase in weighted degree yields a 0.15 SD decrease in average daily departure delay. The results obtained in tables 1 and 2 were consistent for every year between 2009 and 2013. The estimates from linear models provide evidence that while network measures, specifically weighted degree significantly, explain the traffic at a station, they fail to explain the real-time delay in the Amtrak railway network.

5. Discussion and conclusion

To summarize, analysis of the Amtrak train delays between 2006 and 2014 using tools of complex network analysis was the focus of this study. This study suggests new conjectures about the nature of delay in rail transportation networks and the need for re-investigating network measures used to study model transportation networks. By connecting pairs of stations, the Amtrak

network was generated if there existed at least one direct link between the two train stops.

The Amtrak network has evolved significantly over the years. As observed in the US air transportation networks, here too long departure delays exist. Interestingly, on an average, negative departure delays for few stations for every year between 2006 and 2014 were seen. The predictive power of various network centrality measures on the average daily-delay of trains were compared. A negligible effect of centrality measures on the train delay at various stations was observed. On the other hand, *weighted degree* of a node (Amtrak station) has a very strong effect on the traffic faced by a station. Most of the literature on model communication networks show a stronger role of betweenness centrality where traffic congestion is the highest [23]. Again, recent research also demonstrated the node spreading influence of the k -core of real networks [24]. The results of the present study thus significantly affect the studies involving susceptibility and tolerance of communication networks.

The results show that the Northeast corridor of the Amtrak faces the highest traffic but the least average departure delays. The effect is persistent for every year between 2009 and 2014. On the contrary, the regions in the US that experience least traffic faces the maximum average departure delays, supporting a previous analysis which suggests host effects and speed restrictions imposed by host railroads as chief causes of Amtrak delays [25]. The next question, which deserves further attention, involves the factors that result in failure of model-network measures to emulate real-network patterns. Kurant and Thiran [26] suggest a layered view of transportation networks, in which the network is split into a physical graph and a logical graph. It would be interesting to distinguish the Amtrak railway network on multiple layers—the logical graph capturing the real-time delay dynamics and the physical graph. One of the key shortcomings of applying network estimators on real-life networks like the Amtrak network is that in real networks one needs to consider external factors like speed restrictions, capacity utilization, and host effects. A possible solution would be to identify potential congestion points in real transportation networks using the concept of routing betweenness centrality (RBC). An earlier study on RBC algorithm demonstrates its advantage in conventional shortest-path algorithms [27].

The results of his study lay the foundation for future research on modeling the spread of delays in rail transportation networks. Even though the Amtrak railway network was considered in the analysis, it would be interesting to see if similar patterns emerge from other busy transportation networks in the world. For example, the Indian railway network, Indian road network,

and European railways face long delays, which have an economic impact in these countries [28]. A possible future effort includes a data-driven approach for modeling and subsequently mimicking the delay propagation patterns in the Amtrak railway network, similar to that undertaken in the case of the US air transportation networks [14].

Acknowledgements

The author thanks the dixielandsoftware website for public availability of on time performance of Amtrak railway in the United States.

References

- [1] N G Rupp, Further investigations into the causes of flight delays. *Working paper*, Department of Economy (East Carolina University 2007)
- [2] A Barrat, R Pastor-Satorras and A Vespignani, *Proc. Natl. Acad. Sci.* **101**, 3747 (2004)
- [3] M Barthelemy, *Phys. Rep.* **499**, 1 (2011)
- [4] G Burghouwt and J de Wit, *J. Air Transport Manage.* **11**, 185 (2005)
- [5] R Guimera, S Mossa, A Turttschi and L A N Amaral, *Proc. Natl. Acad. Sci.* **102**, 7794 (2005)
- [6] W Li and X Cai, *Phys. Rev. E* **69**, 046106 (2004)
- [7] S Colak, A Lima and M C Gonzalez, *Nat. Commun.* **7**, 10793 (2016)
- [8] P Sen, S Dasgupta, A Chatterjee, P A Sreeram, G Mukherjee and S S Manna, *Phys. Rev. E* **67**, 036106 (2003)
- [9] S Ghosh, A Banerjee, N Sharma, S Agarwal, N Ganguly, S Bhattacharya and A Mukherjee, *Acta Physica Polonica B Proc. Suppl.* **4**, 123 (2011)
- [10] P Sen and B K Chakrabarti, *Sociophysics: An introduction* (Oxford University Press, Oxford 2013)
- [11] W Li and X Cai, *Physica A* **382**, 693 (2007)
- [12] K A Seaton and Hackett, *Physica A* **339**, 635 (2004)
- [13] V Colizza, A Barrat, M Barthelemy and A Vespignani, *Proc. Natl. Acad. Sci.* **103**, 2015 (2006)
- [14] P Fleurquin, J J Ramasco and V M Eguíluz, *Sci. Rep.* **3**, 1159 (2013)
- [15] B Monechi, V Loreto and V D P Servedio, *PLoS ONE* **10**, e0125546 (2015)
- [16] A Higgins, L Ferreira and E Kozan, *Trans. Plan. Tech.* **19**, 89 (1995)
- [17] A Higgins and E Kozan, *Trans. Sci.* **32**, 346 (1998)
- [18] D Levinson and B Yerra, *Trans. Sci.* **40**, 179 (2006)
- [19] E R Peay, *Social Networks* **2**, 385 (1980)
- [20] S Carmi, S Havlin, S Kirkpatrick, Y Shavitt and E Shir, *Proc. Natl. Acad. Sci.* **104**, 11150 (2007)
- [21] M Rosvall and C T Bergstrom, *PLoS ONE* **6**, e18209 (2011)
- [22] R Guimera and L A N Amaral, *Nature* **433**, 895 (2005)

- [23] P Holme, *Advs. Complex Syst.* **6**, 163 (2003)
- [24] F D Malliaros, M-E G Rossi and M Vazirgiannis, *Sci. Rep.* **6**, 19307 (2016)
- [25] Federal Railroad Administration *Report CR-2012-148* (2012)
- [26] M Kuran and P Thiran, *Phys. Rev. Lett.* **96**, 138701 (2006)
- [27] S Dolev, Y Elovici and R Puzis, *J. ACM* **57**, 1 (2010)
- [28] S Ghosh, A Banerjee and N Ganguly, *Physica A* **391**, 2917 (2012)
- [29] M Girvan and M E J Newman, *Proc. Natl. Acad. Sci.* **99**, 7821 (2002)
- [30] A E Raftery, *Sociol. Methodol.* **25**, 111 (1995)