Chinese Railway Network Analysis

Junlin Zheng
junlin.zheng@mail.mcgill.ca
McGill University
Montreal, Quebec, Canada

ABSTRACT

Transportation system is one of the lifelines of a country. Since railways are the stablest and also exceptionally energy-efficient, yet less flexible in the meantime, it is of high worthiness to understand the railway network, in terms of its structure, properties, strengths and possible weaknesses. China has one of the largest and the most hectic railway networks in the world, which has been tremendously expanding over the recent years, contributing significantly to its economic growth and social development. In this paper, a meticulous study has been conducted on the Chinese railway network, closely inspecting its topological and statistical features such as degree distribution and correlations, clustering coefficients, path lengths, components, centrality under different measures and the overall estimates, communities, and its robustness and attack tolerance.

KEYWORDS

network science, centrality, correlations, community detection, robustness, attack tolerance, datasets

1 INTRODUCTION

Transportation infrastructure is the sinews of a nation. It influences people's travel options and logistics of goods on a daily basis, plays a significant role in its economic growth, and thus can be regarded as a crucial indicator of the development level of a country. Among all the modes of transport, railway system, in particular, is the safest and the most stable one. It was born with high passenger and cargo capacity, with economy of scale, which renders it remarkably energy-efficient. On the other hand, due to the necessity of pre-constructed tracks, it appears unavoidably to be less flexible and more capital-intensive. On account of these traits, it is of cardinal significance to scrutinize the structure and properties of a railway network, and hence to analyze its robustness and resilience in the face of attacks.

Specifically, China has one of the largest and busiest railway systems in the world. By the end of 2018, it had reached 131,000 kilometres operating distance in total – the second longest network in the world, among which 29,000 kilometres were high-speed rail (HSR) – the longest HSR network in the world. In 2018, railways in China delivered 3.375 billion passenger trips, an increase of 9.4% from the year before, generating 1414.658 billion passenger-kilometres, increased 5.1%; and carried 4.026 billion tonnes of freight, an increase of 9.2% from 2017, generating 2882.099 billion cargo tonne-kilometres, increased 6.9% [12]. Freight traffic turnover has increased more than fivefold since 1980 and passenger traffic turnover has increased more than sevenfold over the same period [23]. It has been tremendously expanding over the recent years, with investments of hundreds of billions of dollars from the government, carrying billions of passengers as well as cargo in each year, and generating hundreds of billions of dollars in revenues at the same time, all of which makes it a lifeline of the nation. Consequently, the structure and statistics of the Chinese railway network (CRN) may have changed substantially in the last decade and it will be both meaningful and challenging to probe into this network.

Therefore, in this project, a meticulous study will be conducted on the Chinese railway network, closely inspecting its topological and statistical features such as degree distribution and correlations, clustering coefficients, path lengths, components, and more importantly, centrality of hubs under different measures, communities, its robustness and attack tolerance.

2 RELATED WORK

Network science theory has been widely applied to various real-world transportation networks. Indian railway network was first studied by Sen et al. [18], which disclosed its small-world property. Later in 2011, Ghosh et al. [6] further evaluated Indian railways, not only confirmed that it was a small-world, but also displayed its exponential distributions of both node-degrees and edge-weights, as well as the most important stations with respect to its connectivity and traffic flow. In 2010, the research from Soh et al. [19] highlighted that the rail network in Singapore was almost fully connected and its hub nodes experienced disproportionately large traffic. Pakistan railways were examined by Mohmand et al. [9] in 2014, revealing its hub cities based on the betweenness and closeness centralities of nodes, and that it was also a small-world, with assortativity. These works were mainly concentrating on the metrics of the networks. In more recent years, as vulnerability and robustness of transportation network seem to start attracting more attention, the most critical links of Iran railways were analyzed based on flow interdiction by Bababeik et al. [1]. Later, Pagani et al. [13] focused on the interdependent rail networks in Greater London and around last year, analyzed its cascade dynamics and hence its resilience and robustness.

Some work has been dedicated to Chinese rail networks specifically too. Li et al. [8] looked into Chinese railway network in 2007 and discovered its small-world and scale-free property, as well as its degree correlations and the distributions of clustering coefficients, shortest-path lengths, and real spatial distances. In 2009, inspired by resource-allocation process, Wang et al. [20] designed a method to project a Chinese train-station bipartite network into a weighted station network, and then proposed a new metric to quantify the dependence between pairs of stations, which displayed a shifted power-law distribution.

Recently, due to the surge of high-speed rail (HSR) construction and its immediate prosperity in China, more studies start to pay more attention to this HSR network and especially its evolutionary dynamics. For instance, in 2018, Chen et al. [4] examined the

development of this network over the period 2003–2014 in terms of its accessibility measured by node degree, strength, closeness, and betweenness, and briefly compared its development pattern with other HSR networks in Asia, namely Taiwan, South Korea and Japan. Similarly, Xu et al. [22] presented the evolution process and network characteristics of China's HSR network during 2007–2017, found that the degree and eccentricity of the current major cities increased over time. Moreover, according to China's national railway planning proposal, which sketched out the development prospects of China's HSR network from 2018 to 2030, they also estimated its future statistics and pointed out that the pagerank of the current major cities and their contribution to the network connections would decrease in the long term. Essentially, it revealed the intention to slightly shift the hubs of the current network to some less overcrowded, relatively smaller cities. More specifically, Huang et al. [7] reported the evolving statistical properties and spatiotemporal patterns of China's railway network during each of the four main stages of HSR development over a 10-year period from before August 2008 until July 2017, where they found, in addition to that it was scale-free, the trend of a decreasing average path length and an increasing network clustering coefficient that indicated a more and more outstanding small world characteristic during the evolution of the network. However, Wei et al. [21] later spotlighted only the HSR network with its 704 HSR stations, generated the frequency distributions of its degree centrality, betweenness centrality, and closeness centrality which exhibited highly consistent bimodal-like patterns, and came into conclusion that instead of a scale-free structure, Chinese HSR system was actually more hierarchical. Likewise, a work from Cao et al. [3] that overviewed the entire railway network, although again confirmed its small-world and scale-free characteristics on one hand, also presented its spatial heterogeneity and hierarchy on the other hand. Based on the centralities with respect to degree, strength, betweenness, and closeness, they discovered that, quite counterintuitively, the most connected stations are not necessarily the most central ones in the network, and thus proposed an integrated measure from the four centralities to identify the global role that each node played in this multilayered network.

Although different topological, statistical features of Chinese railway network have been studied by different researchers before, none of them seemed to be thorough enough that had covered a certain number of the informative and interesting ones. Moreover, seldom work has accounted for the network robustness and resilience, which is in fact of extremely high significance to a network that is expected to be as stable and strong as possible. This project will try to integrate and evaluate these properties, so as to further understand the complex structure and distribution of the railway network in China.

3 PROBLEM DEFINITION

Intuitively, railway system can be modelled as a network composed of stations as its nodes and train schedules as its edges, where any pair of stations are connected by any train that calls at them. As proposed, certain relevant measures and metrics will be used to quantify the structure of this network, as presented below.

3.1 Degree distribution

In an undirected network of n nodes, the degree k of a node i is the number of edges attached to it, namely, $k_i = \sum_{j=1}^{n} A_{ij}$, where A_{ij} is the entries in its adjacency matrix. The degree distribution p_k of the network is then defined to be the fraction of nodes in the network with degree k. Thus if there are n_k nodes in the network that have degree k, we then have

$$p_k = \frac{n_k}{n}.$$

The value p_k can also be perceived as the probability of a randomly chosen node in the network having degree k.

It is one of the most fundamental properties of a network, which directly gives rise to *power-law* or so-called *scale-free* property of a network, in which the relation between p_k and k roughly follow a power law:

$$p_k = Ck^{-\alpha}$$

where C and α are constants, and α typically falls into range $2 \leq \alpha \leq 3$.

3.2 Component structure

Completely separate parts in a network are called *components*, a subset of nodes in the network that are connected to each other (directly or indirectly) within this subset and are absolutely disconnected with any other nodes outside this subset. Theoretically, we would expect a railway network to be fully connected – that is all nodes should belong to the same single component.

3.3 Path lengths

One of the most striking and widely discussed network phenomena is the *small-world* effect, the finding that in general, distances between any pair of nodes in numerous networks are astonishingly short. Mathematically, let d_{ij} define the distance between nodes i and j which is essentially the length of the shortest path between them, then the mean distance ℓ_i from node i to every other node is

$$\ell_i = \frac{1}{n} \sum_j d_{ij}.$$

So the mean distance ℓ between all nodes in this network can then be obtained by averaging ℓ_i across all nodes, as follows:

$$\ell = \frac{1}{n} \sum_i \ell_i = \frac{1}{n^2} \sum_{ij} d_{ij}.$$

A network with a small value of ℓ is identified as a *small world*, meaning that only a few steps or hops are needed to jump to any other nodes in the network from any given node. In railway networks, specifically, this would imply that only a small number of transfers between trains are necessary to travel between any pair of cities.

On the other hand, the *diameter* of a network is the length of the longest finite distance between any two nodes, which will also be examined for the railway network.

3.4 Clustering coefficient

Clustering coefficient is defined as the average probability that two neighbors of the same node are themselves neighbors, capturing

the level of the network transitivity. Formally, it can be written as:

$$C = \frac{number\ of\ closed\ paths\ of\ length\ two}{number\ of\ paths\ of\ length\ two}$$

3.5 Degree correlations

Assortativity is the tendency of nodes to connect to others that are like them in some way. When node degree is used as the measure of similarity, we may find distinct behaviors displayed by networks: nodes of comparable degree prefer to link to each other – assortative networks; or on the contrary, hubs (nodes with high degree) tend to connect to small-degree nodes and vice versa – disassortative networks; or there is no significant correlation between neighbors in terms of degree – neutral networks. One way to quantify this property is by correlation coefficient defined as follows [10]:

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)k_i k_j}{\sum_{ij}(k_i\delta_{ij} - k_i k_j/2m)k_i k_j},$$

where A denotes the adjacency matrix of the network, m the total number of edges, k is conventionally the degree of a node, and δ_{ij} is the Kronecker delta.

3.6 Centrality

Centrality is used to identify central or important nodes, or so-called "hubs", in networks. The word "important" may suggest a variety of implications, which leads to a variety of centrality measures, including degree centrality, eigenvector centrality, PageRank, closeness, and betweenness, etc.. Different ones will be covered here for a more complete analysis.

3.7 Communities

A network displays a community structure if its nodes can be easily grouped into subsets of nodes such that each set of nodes is densely connected internally. Detection of communities, if exist, allows us to break a large network apart into smaller clusters so that individual clusters can be studied separately, which fits Chinese railway network perfectly. In addition, communities may exhibit dissimilar features from the average of the whole network. So it is of certain significance to try to cluster the network and see whether it has a community structure or not.

However, dividing a network into different partitions can be a challenging task. Several methods have been developed, such as modularity maximization, the InfoMap method, betweenness-based methods, and hierarchical clustering. Some of them will be applied to the railway network.

3.8 Robustness and attack tolerance

Robustness of a transportation network is of ultimate priority, which ensures its ability to survive and keep functioning despite occasional component failures. Moreover, in the face of deliberate attacks targeting at the hubs, we want the network to hold on as long as possible, before breaking into tiny pieces and becoming practically crippled.

Percolation process will be adopted to inspect the network's robustness, by randomly, gradually removing nodes and all their associated edges from the network, observing the *percolation threshold* at which the giant component falls apart. This definition of critical threshold also applies to the process of hub attack, but it is for sure that a quite different value would be observed, which characterizes a network's vulnerability or resilience under attack.

Theoretically, the *percolation threshold* (also called *critical threshold*) of robustness is given by

$$f_c = 1 - \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$$

where $\langle k^n \rangle$ is the n-th moment of the degree distribution, defined as

$$\langle k^n \rangle = \sum k^n p_k$$

This equation tells us the fraction of nodes that needs to be (randomly) destroyed in the network when a giant cluster ceases to exist.

4 METHODOLOGY

No official graph dataset of the railway network can be ready to use yet. In addition, as the network is currently growing and always subject to changes, it would be better to analyze the most up-to-date data. Fortunately, the latest information of train schedules has been made available on *the official railway ticket website*. Therefore, the metadata can be retrieved from this website by a crawler, which can then be extracted to build the railway network graph.

As briefly described in section 3, in this network, nodes represent the train stations, and edges are trains connecting them. It will be modelled as an undirected network since the trains scheduled between two stations are always bidirectional. As there could be multiple trains passing through the same stations, this network will be constructed into a multigraph.

The network's degree distribution (section 3.1), component structure (section 3.2), lengths of averaged shortest path and diameter (section 3.3), clustering coefficient (section 3.4), degree correlations (section 3.5) and its communities (section 3.7) will be analyzed according to their definitions. Similarly, different measures of centrality (section 3.6), including degree, closeness, betweenness, PageRank [14] and VoteRank [24], will be evaluated separately, and the ranks of nodes' PageRank and VoteRank will then be averaged as the indicator of their overall rankings as hubs. Specifically, since PageRank method involves the eigenvector calculation, which could be solved by different approaches and therefore the results may subject to slight differences, two sets of PageRank values are computed to provide more information, using the NumPy's interface to the LAPACK eigenvalue solvers and the power iteration with a SciPy sparse matrix representation.

Percolation process will be used to examine the network's robustness, and its theoretical critical threshold f_c will be calculated based on its degree distribution (section 3.8) as well. Occasional failures of some of the stations will be simulated by randomly removing the nodes, along with the edges connected to them, from the network, in order to closely monitor and pinpoint the actual critical point – the threshold at which the giant connected component of the network would fall apart. Both the removed fraction when the GCC breaks into two parts and the curve of the new GCC portion during this simulation will be reported. Since this process is stochastic, the result will be averaged over 200 runs. On the other hand, to assess the attack tolerance, a similar but definite procedure will be applied, during which it is the hubs with higher degree that

will be "attacked" accordingly, rather than nodes being removed at random.

5 RESULTS, EVALUATION AND DISCUSSION

China has 22 provinces, five autonomous regions, four direct-controlled municipalities, and two special administrative regions (Hong Kong and Macau). 2884 train stations are officially reported, though only 2526 of them are currently in use (are visited by scheduled trains). More than 8500 trains shuttling to and fro between the stations everyday, and almost 8800 different ones in total, contributing to a large multigraph with 2526 nodes and 568526 edges.

5.1 Degree distribution

The network appears to be roughly scale-free, as both Figure 1 and Figure 2 show, with a fitted power law constant $\alpha = 3.13$. The average degree is about 450, while the maximum degree value is 8588, the minimum is 4, and the mode value is 26.

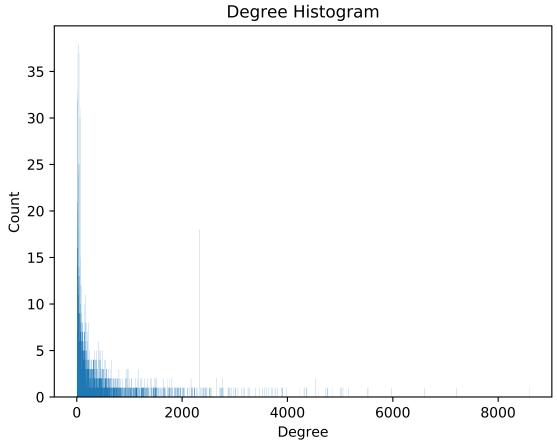

Figure 1: Degree histogram

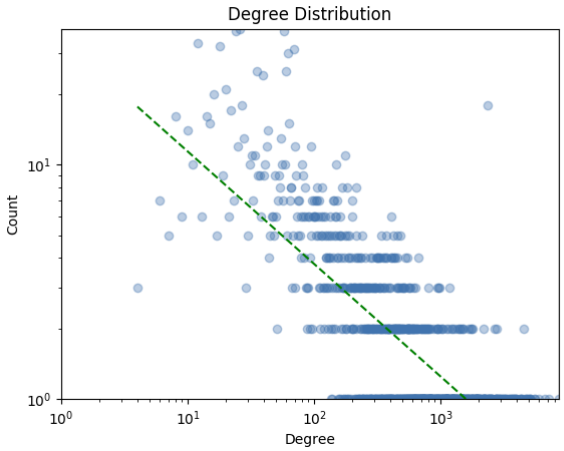

Figure 2: Degree distribution on log scale

5.2 Component structure

There are three components in the network. The giant connected component (GCC) incorporates 2498 nodes , taking up 98.892% of all nodes, whereas the other two have only 17 (0.673%) and 11 (0.435%) nodes, restricted in two small regions respectively in Guangdong and Sichuan province. These two separated components consist of stations that are relatively new and for specific purposes, connected only by high-speed intercity trains (train number starts with C).

5.3 Path lengths

Similar to many real networks, (the GCC of) Chinese railway network also maintains the small-world property. Its averaged shortest path across all nodes is only 2.63, indicating that generally merely less than 2 transfers are necessary for passengers to travel between any places. On the other hand, its diameter is 6, meaning that at most 5 transfers are needed.

5.4 Clustering coefficient

The clustering coefficient of this network is 0.702, which is relatively high, suggesting that if two stations are both directly connected to another station, they themselves are very likely directly connected to each other. In other words, this network is fairly transitive.

5.5 Degree correlations

The degree correlation coefficient is 0.167, implying little assortativity of this network. While it indeed shows a (slight) tendency for stations to link with other stations of similar degree, the tendency is not particularly strong, so the mixing within the network appears to be quite well-distributed.

5.6 Centrality

As mentioned (in section 4), the centrality of each station is assessed with multiple measures, including degree, closeness, betweenness, PageRank (two methods used) and VoteRank centrality. Their overall "importance" is then obtained by averaging their ranks of the PageRank and VoteRank centrality. While degree, closeness and betweenness are in no way trivial at all, when accounting them for the overall ranking, the results do not seem to make as much sense as the current one, as several cities, though would not pop out as the most busiest cities in China traffic-wise, turns out to hold substantially high closeness or betweenness, which would significantly contribute to their rankings and change the final appearance of this list. (For instance, *Shenyang*, *Shenyang North* and *Changchun* perform extremely well when competing closeness and betweenness centrality, most likely because they are the biggest cities in Northeast China, which once was the major industrial base of the country. Although the current economic centres have apparently shifted to other areas in China, since recent years have seen the stagnation and shrinking of its once-powerful heavy-industry sector and a decline of its economic growth, there still are a great number of railway stations in Northeast China. And as it will be demonstrated by the community detection results (section 5.7, Figure 3) later, railways in this area are relatively "isolated" from the rest, rendering

Table 1: Station Centrality Ranks

#	Station	Degree	#	Closeness	#	Betweenness	#	PageRank(N)	#	PageRank(S)	#	VoteRank
1	**Hangzhou East**	8588	0	0.49996	46	0.00715	51	0.00417	1	0.00409	2	0
2	**Guangzhou South**	5970	3	0.44621	341	0.01451	17	0.00382	3	0.00378	4	3
3	**Shenyang**	4315	20	0.52249	13	0.04185	1	0.00445	0	0.00449	0	13
4	**Shenyang North**	4767	11	0.55703	0	0.02295	5	0.00376	4	0.00378	3	7
5	**Zhengzhou**	5154	6	0.54402	4	0.01526	16	0.00358	5	0.00358	5	6
6	**Nanjing South**	7209	1	0.46573	204	0.00232	164	0.00350	8	0.00345	9	1
7	**Shijiazhuang**	5524	5	0.53367	7	0.02018	6	0.00352	7	0.00351	7	4
8	**Changchun**	4327	19	0.54128	5	0.02762	3	0.00409	2	0.00412	1	16
9	Tangshan	4718	13	0.54776	3	0.01289	22	0.00353	6	0.00354	6	10
10	**Shanghai Hongqiao**	6607	2	0.46741	192	0.00279	135	0.00317	11	0.00311	11	2
11	Shanhaiguan	4534	15	0.55082	1	0.01565	15	0.00346	9	0.00348	8	14
12	Suzhou	5013	9	0.50903	30	0.00661	55	0.00297	12	0.00295	12	9
13	**Changsha South**	5541	4	0.46337	223	0.00234	161	0.00282	14	0.00278	14	5
14	**Tianjin**	4238	21	0.55008	2	0.01615	12	0.00329	10	0.00331	10	18
15	Wuxi	4849	10	0.50840	32	0.00645	57	0.00287	13	0.00286	13	15
16	Yiwu	5027	8	0.51854	17	0.00829	43	0.00272	16	0.00269	16	11
17	**Zhengzhou East**	4755	12	0.47323	170	0.00360	99	0.00267	18	0.00265	21	12
18	Shangrao	5054	7	0.45610	266	0.00335	109	0.00261	23	0.00257	25	8
19	Changzhou	4534	16	0.50446	36	0.00591	62	0.00268	17	0.00267	18	24
20	Xuzhou	3781	27	0.52763	11	0.00584	63	0.00265	21	0.00265	20	25
21	**Lanzhou**	2976	46	0.48888	84	0.01604	14	0.00273	15	0.00274	15	38
22	**Xi'an**	3243	39	0.50312	39	0.01002	32	0.00266	19	0.00267	19	33
23	**Wuhan**	4584	14	0.46416	214	0.00426	82	0.00252	27	0.00248	28	17
24	**Nanjing**	3979	22	0.52194	14	0.00882	38	0.00263	22	0.00262	22	35
25	**Changsha**	3604	32	0.50332	38	0.01386	20	0.00248	29	0.00248	29	22
26	Siping	3018	44	0.53195	9	0.01304	21	0.00266	20	0.00268	17	45
27	**Chengdu East**	3526	34	0.46829	187	0.00858	40	0.00243	30	0.00241	30	26
28	**Xi'an North**	3914	23	0.45168	300	0.00274	139	0.00236	33	0.00233	34	21
29	**Beijing West**	3386	36	0.49865	51	0.03210	2	0.00254	26	0.00254	26	39
30	**Jinan**	3573	33	0.54033	6	0.01218	24	0.00236	32	0.00236	32	27
31	Jinhua	4355	18	0.44758	336	0.00255	148	0.00235	34	0.00232	35	23
32	Xuzhou East	4364	17	0.46026	243	0.00123	264	0.00217	37	0.00214	37	20
33	Qinhuangdao	3473	35	0.53333	8	0.00875	39	0.00248	28	0.00249	27	40
34	**Harbin**	2552	63	0.48456	113	0.01901	7	0.00258	25	0.00260	23	53
35	Jinzhou	2872	51	0.53172	10	0.01645	10	0.00258	24	0.00260	24	55
36	Shangqiu	3704	30	0.51953	16	0.00905	35	0.00232	36	0.00232	36	31
37	**Nanchang**	3255	38	0.52272	12	0.00709	52	0.00213	38	0.00212	39	30
38	**Wuchang**	3122	42	0.51177	26	0.01048	30	0.00235	35	0.00235	33	42
39	**Xiamen North**	3905	24	0.47883	140	0.00366	96	0.00196	46	0.00192	48	19
40	Zhuzhou	3162	41	0.50559	34	0.00324	113	0.00212	39	0.00211	40	37
41	**Nanchang West**	3888	25	0.48494	110	0.00411	85	0.00199	44	0.00196	44	29
42	**Beijing**	2522	66	0.52007	15	0.02370	4	0.00238	31	0.00240	31	60
43	**Jinan West**	3718	29	0.45941	248	0.00167	207	0.00196	45	0.00194	47	34
44	**Hankou**	3182	40	0.48762	95	0.01176	26	0.00199	43	0.00197	43	41
45	**Guangzhou**	2766	55	0.51231	25	0.01729	8	0.00205	42	0.00205	42	50
46	Ningbo	3811	26	0.48637	101	0.00343	104	0.00187	54	0.00184	56	28
47	Shenzhen North	3645	31	0.42457	539	0.00143	233	0.00187	53	0.00183	57	32
48	Jiujiang	2942	47	0.51605	19	0.00534	68	0.00188	50	0.00187	51	44
49	Liuzhou	2885	50	0.49555	58	0.00686	54	0.00187	51	0.00186	52	43
50	Zibo	2773	53	0.51746	18	0.01109	27	0.00196	47	0.00195	46	56
51	**Kunming**	2025	116	0.49416	63	0.04778	0	0.00212	40	0.00213	38	76
52	Hengyang	2773	54	0.50549	35	0.00350	102	0.00191	49	0.00191	49	58
53	Qiqihar	1883	134	0.47615	155	0.01436	18	0.00208	41	0.00210	41	79
54	**Chongqing West**	2780	52	0.50169	43	0.01035	31	0.00184	57	0.00183	58	47
55	**Urumqi**	1878	136	0.49376	64	0.01406	19	0.00195	48	0.00195	45	72
56	Wenzhou South	3765	28	0.44293	364	0.00062	389	0.00174	63	0.00171	68	36
57	**Nanning**	2299	96	0.48839	92	0.01285	23	0.00186	55	0.00185	54	68
58	Guangyuan	2516	67	0.50220	42	0.00946	33	0.00183	58	0.00183	59	61

these northeastern cities quite crucial in connecting this area to the others, and thus they would appear to be extraordinarily influential regarding their closeness and betweenness.)

The top 58 hubs are listed in Table 1. Except for VoteRank centrality, which comes out as a rank value directly, other measures are followed by their integer rankings (denoted by "#" in the table) among all the stations. And the whole table is organized by their overall rankings, in descending order.

In Table 1, major cities (municipalities and provincial capitals) are highlighted in bold. It is obvious that most of them are major cities. For example, as two of the first-tier cities (and also the biggest ones) in China, Guangzhou (2nd and 45th stations *Guangzhou South* and *Guangzhou*)[1] and Shanghai (10th station *Shanghai Hongqiao*) both stand very high in the list. While Beijing seems to slightly fall behind, at 29th and 42nd (stations *Beijing West* and *Beijing*), its closeness and betweenness metrics turn out to be rather outstanding. Indeed, as it will also be shown later in Figure 3, Beijing is indeed the heart of China – not only politically, but also with regard to transportation: it is very likely the most all-round city in China. It is at the intersection of almost all railway communities and thus it essentially leads to everywhere within the country.

As the highest-ranking station, *Hangzhou East* has the largest degree, which means it experiences probably the biggest traffic volume. Indeed, it is one of, if not the, largest rail hubs in China (what is for sure is that at least it was the largest one when coming into operation in 2013) – and even within Asia. It serves the whole Hangzhou area almost on its own (the other two stations in Hangzhou are too small even comparing to other stations in other cities). Sitting on the intersecting point of some of the busiest rail corridors in China, *Hangzhou East* station houses the High Speed CRH service to Shanghai, Nanjing, Changsha, Ningbo, and beyond. In fact, most other major cities in China can be reached by direct train service from Hangzhou. Now Hangzhou is the fourth-largest city in China, functioning as the core of the Hangzhou metropolitan area, and has been repeatedly rated as the best commercial city in the mainland of China by Forbes. It is an emerging technology hub and one of the leading representative cities in the new growing cities that became popular in the 2010s. As the primary station in Hangzhou, *Hangzhou East* directly links this area with more than 50 main cities and hence the whole country, and therefore is one of the most important transit hubs in China.

Centrality analysis statistically reveals the significance of each station, further implying the prominence of that city. More intriguingly, it may even uncover some surprising historical or geographical facts that are otherwise subject to oblivion. As critical rail hubs often corresponds to the junctions of multiple subnetworks, serving as the bridges between them, more details will be discussed later, together with the partitioned communities, so as to integrate the visualization of the majority of network and the clusters within it, and thus not only numerically but also graphically investigate its structure and distribution.

[1]There could be more than one station in a city in order to avoid congestions or balance traffic load, or to cater for different uses.

5.7 Communities

Since there is no ground truth for community partitions, several common methods have been adopted to detect communities and to compare their performances in order to get a better solution, including Louvain algorithm [2], greedy modularity maximization algorithm [5], Infomap algorithm [17], Walktrap algorithm [15], label propagation algorithm [16], and leading eigenvector algorithm [11]. The modularity of the resulting partition is used to evaluate its quality. Among them, label propagation and Walktrap method achieve the highest performance, with modularity 0.538 and 0.535, detecting 56 and 86 communities, respectively.

To further examine and analyze the partition results, the major communities (top 10 communities with respect to their sizes, among which the smallest one contains about 40 nodes) clustered by the label propagation algorithm are then plotted onto a map. As shown in Figure 3, the outcome looks sufficiently reasonable and fairly intuitive:

- **Beijing** (29th & 42nd in centrality), the capital of China, sits at the junction of almost all railway communities, and hence should be the most "resourceful" city from which one can easily reach anywhere in the country, confirming its high closeness and betweenness.
- **Shanghai** (10th in centrality), **Hangzhou** (top 1 in centrality), and **Nanjing** (6th & 24th in centrality), the three major cities in East China, initialize the most traffic in this area and hence the stations around them are extremely dense, linking this area with Central China, North China and South China. More precisely, while Shanghai's coverage (roughly marked by orange) is more spreaded (although it is relatively skewed to the north), the concentration of Hangzhou (roughly in turquoise) weighs in more favor of the south.
- **Guangzhou** (2nd & 45th in centrality), the largest and most populous city in South China, also serves as a major port and transportation hub. It is the heart of the most-populous built-up metropolitan area in mainland China that extends into the neighboring cities including Shenzhen (47th), forming one of the largest urban agglomerations on the planet. As a result, it even shares the same number of communities converging on one place, appearing to be as "resourceful" as Beijing in this regard. Similar to Beijing and Shanghai, the area Guangzhou can easily reach also spreads over most East China and Central China. But besides that, it goes further into Southwest China (yellow, purple, green markers), a comparatively less-developed region.
- As mentioned previously in section 5.6, the railway community in Northeast China (in red) are somewhat solitary, which nonetheless extensive as well as detailed enough to cover this area both comprehensively and finely. However, to connect with other areas in the country, it has to pass through Beijing or **Tianjin** (14th), which also scores quite high in closeness and betweenness.
- On North China Plain, except for Beijing and Tianjin, nodes around **Shijiazhuang** (7th), **Taiyuan** (69th & 106th, not shown in Table 1 due to space limit), Zhengzhou (5th & 17th), and **Jinan** (30th & 43rd) are also notably crowded, especially Zhengzhou, who is among the top 5 in centrality ranking.

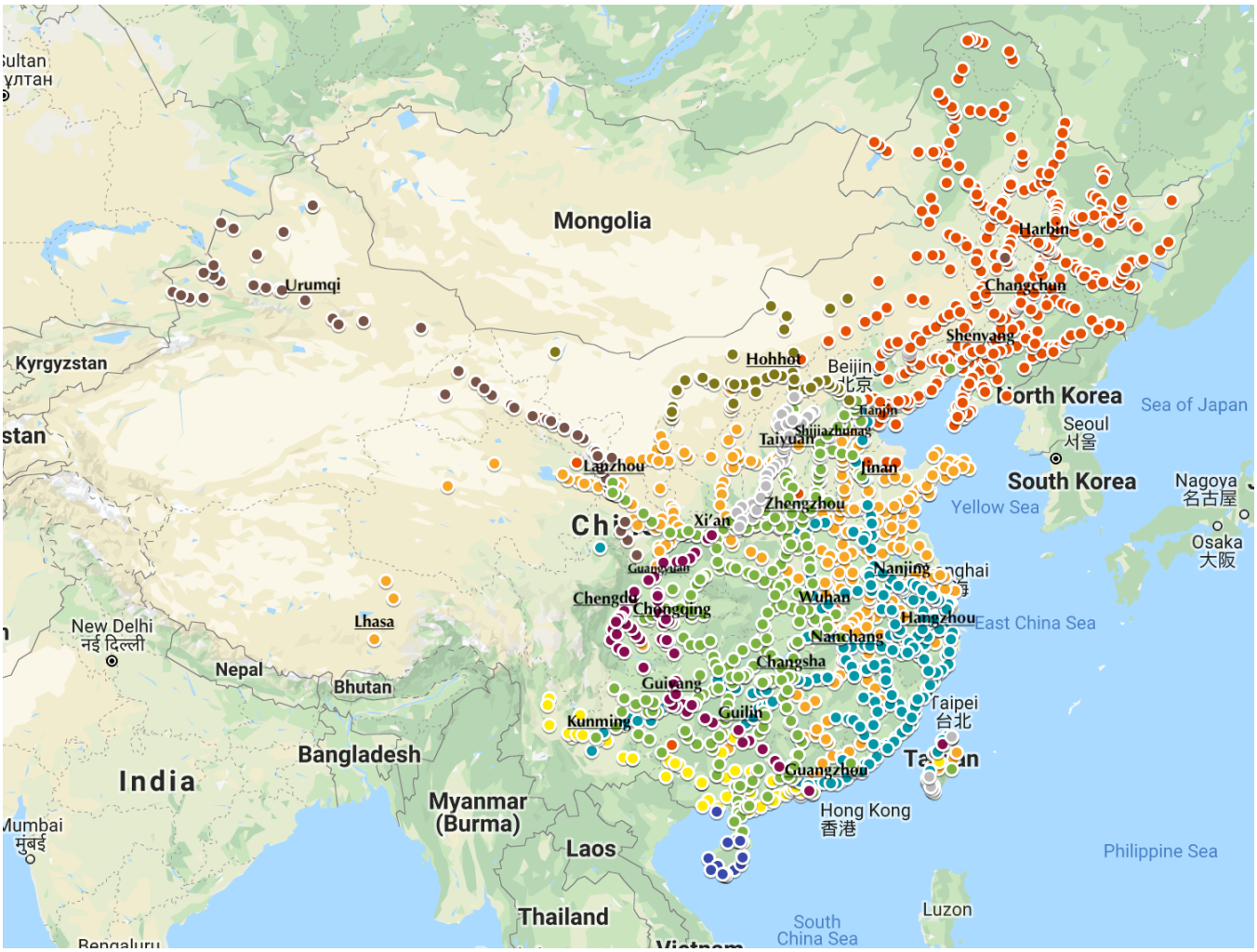

Figure 3: Major partitioned communities

Zhengzhou is at the junction of various nationwide railways which already extends to the winds; meanwhile, even more high-speed railways passing through Zhengzhou have been proposed and are currently under construction, all of which makes Zhengzhou station one of the most important railway stations in China.

- **Changsha** (13th & 25th), **Wuhan** (23rd, 38th & 44th)[2] and **Nanchang** (37th & 41st) are the main cities in Central China. Wuhan Railway Hub is considered one of the key railway hubs of China, served by trains going to all directions.
- The more developed a place is, the more intricate the network around it is. Beijing, Shanghai and Guangzhou all boast much denser and extensive subnetworks surrounding them, radiating in all directions and thus allowing them to conveniently reach any other places. Nevertheless, such transit hubs are also present in less developed areas too. For instance, in Northwest China, apparently **Xi'an** (22nd & 28th) and **Lanzhou** (21st) attracts the most traffic flow; Likewise, in Southwest China, the responsibility is taken by **Kunming** (51st), **Guiyang** (86th & 118th), **Chengdu** (27th) and **Chongqing** (54th).
- The intersection of multiple communities dominantly corresponds to the municipalities and the provincial capitals, as they are evidently the heart of its jurisdiction, wielding enormous impact on the vicinity.

5.8 Robustness and attack tolerance

Since the network consists of three components, and the largest one takes up nearly 99% (2498 nodes) of the whole (section 5.2), in this step, the target analyzing network is replaced by its giant connected component initially, to ensure the observing GCC portion starts from 100%.

[2] All the three railway stations Wuhan, Wuchang and Hankou belong to the city of Wuhan; in fact, the name of "Wuhan" is a portmanteau of the two city names – "Wu" from "Wuchang" and "Han" from "Hankou".

5.8.1 Robustness. To inspect the robustness of the network, nodes and their attached edges are randomly removed from the network, repeated for 200 runs. On average, the network (the original giant connected component) splits into more than one connected components when 7.16% (178.85/2498) of the nodes are disabled. While this value may seem incredibly small, what's worth pointing out is that this value is not the critical point at all – it simply tells us the statistical point at which the network disconnects, meaning that it's no longer possible to reach whatever other places in this network from any given place. However, this does not mean the giant cluster has disappeared; and as it will be illustrated immediately later, it is, in fact, still far from that.

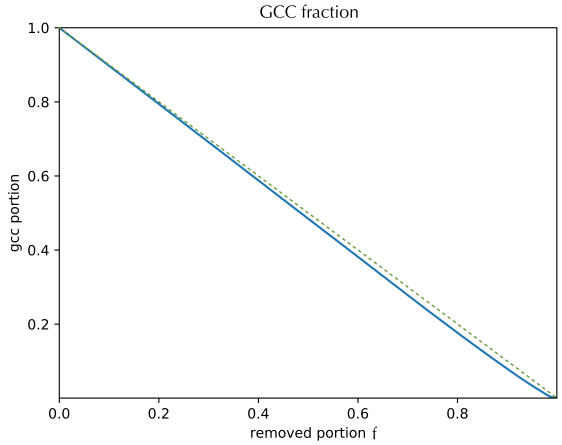

Figure 4: The fraction of stations that belong to the GCC after an f fraction of stations are randomly removed

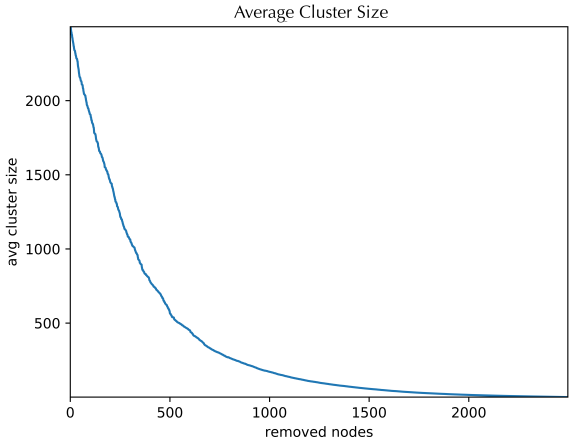

Figure 5: The average cluster size after a number of stations are randomly removed

The theoretical critical threshold f_c (section 3.8) of the Chinese railway network is 99.94%, thanks to its remarkably high second moment of the degree distribution $\langle k^2 \rangle$. It says that almost all stations need to be simultaneously incapacitated, by chance, to fragment the system into many isolated, small pieces, of which the probability is effectively zero. Therefore, this railway network is phenomenally robust, as practically, it would never be paralysed by accident.

Observations from the random station failure simulation process corroborate this analysis. In Figure 4, the actual drop curve of the fraction of nodes in the giant component (blue) almost overlaps with the drop line of the fraction of all left nodes. When the network finally falls apart (exhibiting a very low GCC fraction), almost all nodes have been removed (near $f = 1$), demonstrating a very high critical threshold, $f_c \approx 1$. The decline of the average cluster size in Figure 5 also accords with it: this size, although falls quickly at the beginning, refuses to become small enough until the very end, where more than 2200 nodes are destroyed.

5.8.2 Attack tolerance. Whereas scale-free networks generally tend to be amazingly robust, they are found to be quite vulnerable to attacks, which intentionally go after the hubs, so as to deliberately cripple the network. Normally, the failure of a single hub is unlikely to fragment a network, as the remaining hubs can still hold the network together. After the removal of a few hubs, however, large chunks of nodes would start falling off, and the network could rapidly break into tiny clusters afterwards. So while it is perfectly natural and indeed expected to spot a lower critical threshold of a network under attack, to most scale-free networks, unfortunately, this fraction of removed hubs may be remarkably small – the network would fall apart far before we could imagine.

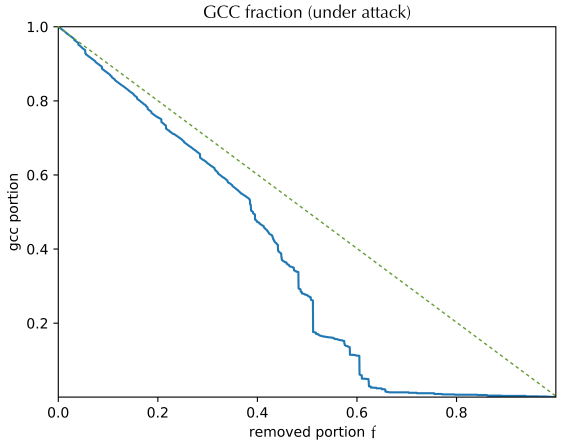

Figure 6: The fraction of stations that belong to the GCC after an f fraction of stations are randomly removed

Outstandingly, however, the simulated critical threshold of the Chinese railway network under malicious strikes on hubs is still sufficiently high. As shown in Figure 6, the network fraction of GCC drops sharply after about 40% of the nodes are removed, which is already rare enough since this essentially corresponds to an improbably large-scale attack on at least 1000 stations. Figure 7

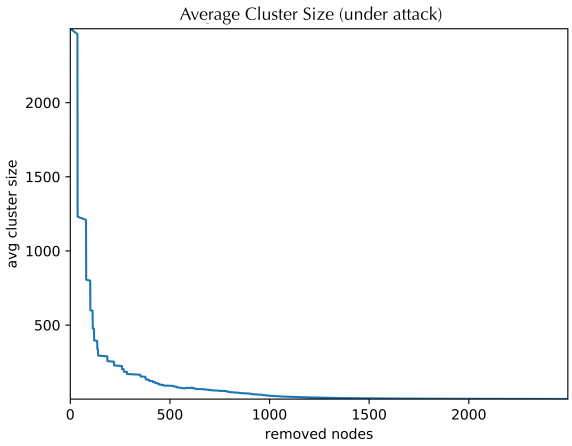

Figure 7: The average cluster size after a number of stations are randomly removed

tells the same story – the average cluster size drops to a very small value after the breakdown of more than 1000 nodes. Therefore, the Chinese railway network is not that susceptible to onslaughts, manifesting an adequately high resilience (tolerance) to attacks.

6 CONCLUSIONS

This paper builds and conducts a comprehensive investigation into the Chinese railway network. A diversity of its statistical and structural properties have been computed and scrutinized, including:

- Scale-free property: this network roughly follows the power law with a fitted constant $\alpha = 3.13$.
- Small-world property: the averaged shortest path across all nodes within this network is as small as 2.63, indicating overall less than 3 transfers are necessary to travel around; and its diameter is 6.
- Component structure: there are three separated components in this network, while 98.892% (2498 nodes) of all nodes belong to the giant connected component.
- The clustering coefficient of the network is 0.702, showing that it is rather transitive.
- The degree correlation coefficient is 0.167, denoting little assortativity in this network.
- The theoretical critical threshold of the network under random failures is 99.94%, suggesting it is phenomenally robust – virtually it would never break down by chance, since it is very unlikely for almost all stations to fail simultaneously.
- The railway network also boasts a sufficiently high attack tolerance, where at least 40% of the hubs (1000 largest stations) need to be destroyed deliberately to cripple the system.

Furthermore, different metrics of the centrality of hubs and detected communities within this network are also evaluated and discussed. By visualizing the partitioned groups together with the calculated hubs, important economic, historical or geographical facts may be disclosed. Specifically, the cities of rail hubs often exhibit higher levels of development, a larger population, a long and glorious history, or crucial geographical location.

In a nutshell, as one of the longest railway networks in the world, the Chinese railways are not only efficiently-operated, but also well-designed. It covers a majority of the vast area of China, allowing a huge amount of passengers and freight to easily move around inside the country, as well as being exceptionally stable and resilient at the same time. Further studies – for example, on its hierarchy, or dynamics – may uncover more information about this network, and hence give rise to even deeper understanding.

REFERENCES

[1] Mostafa Bababeik, Navid Khademi, Anthony Chen, and M. Mahdi Nasiri. 2017. Vulnerability Analysis of Railway Networks in Case of Multi-Link Blockage. *Transportation Research Procedia* 22 (2017), 275 – 284. https://doi.org/10.1016/j.trpro.2017.03.034 19th EURO Working Group on Transportation Meeting, EWGT2016, 5-7 September 2016, Istanbul, Turkey.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (Oct 2008), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

[3] Weiwei Cao, Xiangnan Feng, Jianmin Jia, and Hong Zhang. 2019. Characterizing the Structure of the Railway Network in China: A Complex Weighted Network Approach. *Journal of Advanced Transportation* 2019 (02 2019), 1–10. https://doi.org/10.1155/2019/3928260

[4] Cheng Chen, Tiziana D'Alfonso, Huanxiu Guo, and Changmin Jiang. 2018. Graph theoretical analysis of the Chinese high-speed rail network over time. *Research in Transportation Economics* 72 (2018), 3 – 14. https://doi.org/10.1016/j.retrec.2018.07.030 Long-distance passenger transport market.

[5] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical Review E* 70, 6 (Dec 2004). https://doi.org/10.1103/physreve.70.066111

[6] Saptarshi Ghosh, Avishek Banerjee, Naveen Sharma, Sanket Agarwal, Niloy Ganguly, Saurav Bhattacharya, and Animesh Mukherjee. 2011. Statistical analysis of the Indian Railway Network: A complex network approach. *Acta Physica Polonica B, Proceedings Supplement* 4 (07 2011). https://doi.org/10.5506/APhysPolBSupp.4.123

[7] Yaping Huang, Shiwei Lu, Xiping Yang, and Zhiyuan Zhao. 2018. Exploring Railway Network Dynamics in China from 2008 to 2017. *ISPRS International Journal of Geo-Information* 7 (08 2018), 320. https://doi.org/10.3390/ijgi7080320

[8] W. Li and X. Cai. 2007. Empirical analysis of a scale-free railway network in China. *Physica A: Statistical Mechanics and its Applications* 382, 2 (2007), 693 – 703. https://doi.org/10.1016/j.physa.2007.04.031

[9] Yasir Mohmand and Aihu Wang. 2014. Complex Network Analysis of Pakistan Railways. *Discrete Dynamics in Nature and Society* 2014 (03 2014), 1–5. https://doi.org/10.1155/2014/126261

[10] Mark Newman. 2018. *Networks* (2nd ed.). Oxford University Press, Inc., New York, NY, USA. 800 pages. https://www.oxfordscholarship.com/10.1093/oso/9780198805090.001.0001/oso-9780198805090

[11] M. E. J. Newman. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* 74, 3 (Sep 2006). https://doi.org/10.1103/physreve.74.036104

[12] National Railway Administration of China. 2019. Railway Statistical Communiqué of 2018. (Apr 2019), 10. http://www.nra.gov.cn/xwzx/zlzx/hytj/201904/P020190426367686178375.pdf

[13] Alessio Pagani, Guillem Mosquera, Aseel Alturki, Samuel Johnson, Stephen Jarvis, Alan Wilson, Weisi Guo, and Liz Varga. 2019. Resilience or robustness: identifying topological vulnerabilities in rail networks. *Royal Society Open Science* 6, 2 (2019), 181301. https://doi.org/10.1098/rsos.181301 arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.181301

[14] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/ Previous number = SIDL-WP-1999-0120.

[15] Pascal Pons and Matthieu Latapy. 2005. Computing communities in large networks using random walks (long version). arXiv:physics.soc-ph/physics/0512106

[16] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76, 3 (Sep 2007). https://doi.org/10.1103/physreve.76.036106

[17] M. Rosvall and C. T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 4 (Jan 2008), 1118–1123. https://doi.org/10.1073/pnas.0706851105

[18] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. 2003. Small-world properties of the Indian railway network. *Physical review. E, Statistical, nonlinear, and soft matter physics* 67 (Mar 2003), 036106. Issue 3. https://doi.org/10.1103/PhysRevE.67.036106

[19] Harold Soh, Sonja Lim, Tianyou Zhang, Xiuju Fu, Gary Kee Khoon Lee, Terence Gih Guang Hung, Pan Di, Silvester Prakasam, and Limsoon Wong. 2010. Weighted complex network analysis of travel routes on the Singapore public transportation system. *Physica A: Statistical Mechanics and its Applications* 389, 24 (2010), 5852 – 5863. https://doi.org/10.1016/j.physa.2010.08.015

[20] Yong Li Wang, Tao Zhou, Jian Jun Shi, Jian Wang, and Da Ren He. 2009. Empirical analysis of dependence between stations in Chinese railway network. *Physica A: Statistical Mechanics and its Applications* 388, 14 (2009), 2949 – 2955. https://doi.org/10.1016/j.physa.2009.03.026

[21] Sheng Wei, Shuqing N. Teng, Hui Jia Li, Jiangang Xu, Haitao Ma, Xia-li Luan, Xuejiao Yang, Da Shen, Maosong Liu, Zheng Y. X. Huang, and Chi Xu. 2019. Hierarchical structure in the world's largest high-speed rail network. *PLOS ONE* 14, 2 (02 2019), 1–11. https://doi.org/10.1371/journal.pone.0211052

[22] Wangtu (Ato) Xu, Jiangping Zhou, and Guo Qiu. 2018. China's high-speed rail network construction and planning over time: a network analysis. *Journal of Transport Geography* 70 (2018), 40 – 54. https://doi.org/10.1016/j.jtrangeo.2018.05.017

[23] Hong Yu. 2015. Railway Sector Reform in China: controversy and problems. *Journal of Contemporary China* 24, 96 (2015), 1070–1091. https://doi.org/10.1080/10670564.2015.1030957 arXiv:https://doi.org/10.1080/10670564.2015.1030957

[24] Jian-Xiong Zhang, Duan-Bing Chen, Qiang Dong, and Zhi-Dan Zhao. 2016. Identifying a set of influential spreaders in complex networks. *Scientific Reports* 6, 1 (2016), 27823. https://doi.org/10.1038/srep27823