Identifying risk profiles in the London's public transport system

Roberto Murcio, Chen Zhong, Ed Manley and Michael Batty

Abstract

Public transport networks are a key element in the survival of modern cities and their inhabitants. It is therefore of paramount importance to be able to identify the most vulnerable parts of a network to disruptions through a prompt and effective risk assessment. Due to the complex nature of these networks, detecting the high risk areas is not always an obvious nor an easy task. This research intends to tap into this problem by examining individual trips in a transport network at a fine spatial and temporal resolution. A novel methodology to the field of Big Data using mathematical models from information is introduced, taking London's Underground transport network as a case study. Our analysis is revealing those high-risk areas given by spatiotemporal correlations, which is in contrast with previous assumptions focussed on crowdedness.

R. Murcio (Corresponding author) • C.Zhong • E.Manley • M.Batty
Centre for Advanced Spatial Analysis, University College London, London W1T 4TJ, UK
Email: r.murcio@ucl.ac.uk

C.Zhong.
Email: c.zhong@ucl.ac.uk

E.Manley
Email: ed.manley@ucl.ac.uk

M.Batty
Email: m.batty@ucl.ac.uk

1. Introduction

Public transport networks are vital in ensuring efficient urban function. In many cities, mass disruption to the public transport network has the potential to cause significant damage to the urban economy and wellbeing. Information about where in the network a disruption would cause the more damage is therefore critical for contingency planning. Several studies [1, 2, 3] associate risk to crowdedness, and hence they only consider the most crowded areas as the most susceptible ones to disruption. Nevertheless, some areas, not affected by crowdedness, can be identified as important hubs or brokers of people travelling through the network. These areas have the potential to create unexpected disruptions in the system, due to their spatial-temporal position. The temporal component would strongly depends on the individual behaviours on route choices since these shape relevant emergent properties of the network. These aspects are essential to understand causation so that well-informed risk management decisions can be taken.

As a case study, we take a subset of the multimodal London's public transport network [4] consisting of the Underground mode. In particular, we focus on the network of passengers that use the Oyster smart travel card [5]. During the financial year 2012/13 [6], around 3820 million journeys were registered in London's public transport network. We will analyse a sample of 762 million trips, made using Oyster cards over 2.5 months between July and September 2012.

Such a figure shows the importance of London's transport network for the city's economic output. Understanding the network's vulnerabilities and risks is therefore crucial for the millions of travellers that rely on its functioning.

In this paper, the high-risk areas are associated with underground stations and the transfer of information between any pair of them, understanding information as the amount of people travelling throughout the network at a particular time. A novel methodology to the field of Big Data through the use of mathematical models from information theory [7], and the application of entropy concepts to the Oyster smart card data is introduced. These tools reveal those high-risk areas given by spatiotemporal correlations in contrast with previous assumptions focussed on crowdedness. Proper measures of disruption and damage to the system are being developed to assess the degree of risk in each of these areas, so, at this point, we only identified these risk areas, not the resilience *per se* of the network, so no redundancy analysis is performed at this stage of the investigation. After analysing the Oyster dataset, two main outputs are being produced:

1) A comprehensive temporal profile of the volume of people that, minute by minute, are in each one of the stations. This corresponds not only to the individuals getting in/out of the transport network, but also its transient population. A route profile for each trip needs to be generated in order to obtain these numbers.

2) A complete spatiotemporal risk profile that highlights the so-called hubs in the system. These provide relevant candidates for testing risk management scenarios. In addition, they lead to important insights about specific activity patterns, other than work related, that could eventually be used to assess particular social mobility risks.

2. Oyster card data set

2.1 Smart-card Data

Smart card data becomes available in recent years since an increasing number of cities regions and countries have adopted the Smart Card Automatic Fare Collection (SCAFC) System. The SCAFC system was originally designed to collect revenue for better managing public transportation system, they also produce large volume of data about aboard and alight transactions [8]. Several advantages arise from the analysis of smart-card data have been identified early by [9], mostly regarding to its high spatiotemporal resolution and embedded information about individuals. Later research make use of the advantages, examples can be found using smart-card data to analyse users' travel behaviour [10]; to improve public transportation service [11]; to estimate OD matrix for evaluating system performance [12]; to infer activity types by travel behaviours and functions of urban areas where people reached via public transportation [13]; and to identify the polycentric structure of cites from urban movement patterns [14, 15].

2.2 Smart-card Data in Great London Area

According to the London Travel Demand Survey (LTDS) in 2011, there are about 30% of total population in Great London area using public transportation (including National Rail, Underground/DLR, and Bus/tram) for their daily commuting [16]. Travellers using Oyster cards account for approximately 90% of all bus passengers and 80% of rail passengers [17]. Oyster is accepted on multiple transit modes, including London buses, the London Underground, the London Overground, the Docklands Light Railway (DLR), and Tramlink. The dataset used for this study contains approximately 18 million transactions per weekday, among which, 9 million

transactions are entry/exit transactions of train systems (Underground, Overground and DLR). Both Entry and exit data are available for train rides at gated stations, while only entry data is available for bus rides.
As a starting point, we chose the Underground line known as Victoria[1] (Figure 1). Risk Profile of all stations on these two lines are generated using Oyster card data in 2014. Though many attributes of a transaction are recorded, only four of them, entry station id, exit station id, entry time, exit time are used here, which we considered enough for constructing a train ride. A ride is generated by simply combining a pair of entry and exit records grouped by card id and sorted by transaction time. Only rides starting and ending on the same underground lines are counted in order to provide a strong indication of the exact route through the network. After a preliminary data processing, there are approximate 0.2 million tube rides on Victoria line on an average weekday. Each ride is in the form of card id, boarding time, boarding station id, alighting time and alighting stations id.

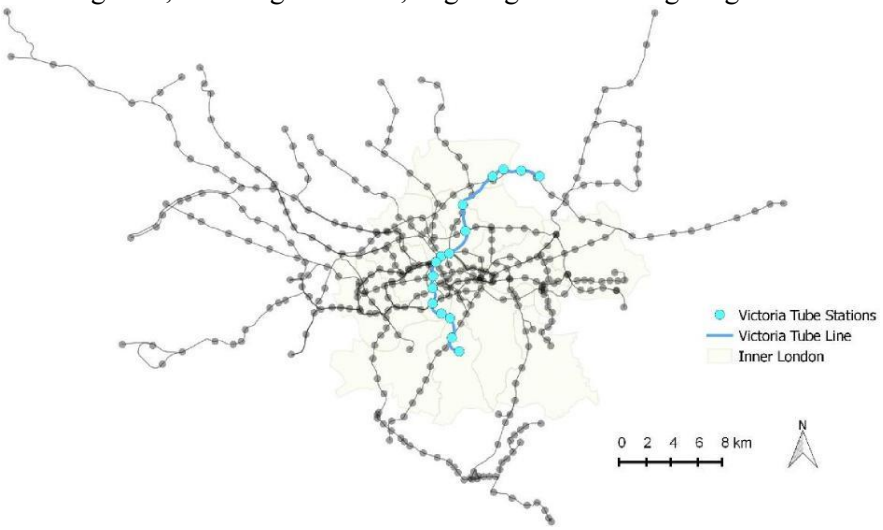

Fig. 1. London's underground network. Enhanced is the Victoria Line

3. Entropy Measures

[1] The Victoria Line, opened in 1968, comprises 16 stations with a total line length of 21km. It runs from Walthamstow Central (North London) to Brixton (South London). It transverses three of the busiest stations in the network: Warren Street, Oxford Circus and Victoria.

In information theory [7], the concept of Shannon entropy (Eq. 1) is the preferred measure to detect the reduction in uncertainty of any measurement *x* of a random variable X.

$$H(X) = -\sum_{x \in X} p(x) \log p(x), \sum_{x \in X} p(x) = 1 \tag{3.1}$$

Extending Shannon entropy to measure the uncertainty between two interacting random variables is accomplished using Mutual Information (MI), defined by:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x) \cdot p(y)} \tag{3.2}$$

One major drawback with MI, is its lack of inherent directionality, MI(X, Y) = MI (Y, X). This shortcoming can be tackled by shifting one of the variables, basing the analysis on the entropy rates and calculating the emerging conditional probabilities of related X and Y. Schreiber [18] defined the concept of Transfer Entropy TE, which use the previous states of X and the next state of Y. Several important statistical properties are found in TE which are useful in analysing systems in which interactions are non-linear. More importantly, it can account for the directional relationships between systems, i.e., it is not symmetrical, unlike mutual information. By now, the use of TE is wide spread [19, 20, 21, 22] and is seen as a robust measure of complexity and interdependence between random variables.

3.1 Transfer Entropy

Given two concurrently sampled spaces of information $X=\{x_1,x_2,\ldots,x_t\}$ and $Y=\{y_1,y_2,\ldots,y_t\}$ the transfer entropy TE from ***X*** to ***Y***, can be obtained from defining the entropy rate between two systems as the amount of additional information to represent the value of the next observation of one of this two systems:

$$h1 = -\sum_{t} p(y_{t+1}, y_t, x_t) \log\big(p(y_{t+1}|y_t, x_t)\big) \tag{3.3}$$

In the second case we define an entropy in which y_{t+1} depends only on y_t thus

$$h2 = -\sum_{t} p(y_{t+1}, y_t, x_t) \log\big(p(y_{t+1}|y_t)\big) \tag{3.4}$$

The Transfer entropy TE between X and Y, is then defined as the difference between these

$$T(X,Y) = h2 - h1 = \sum_{t=1} p(y_{t+1}, y_t, x_t) \log\left(\frac{p(y_{t+1}|y_t, x_t)}{p(y_{t+1}|y_t)}\right) \tag{3.5}$$

Where *t* indicates a given point in time. Basically, (3.5) measures the reduction in uncertainty in y_t , given x_t and y_{t-1} in comparison with given only y_{t-1}, i.e. the amount of information transferred from X to Y.

If this measure is applied directly to our risk detection problem, and X=station A and Y=station B, and t runs for a whole day, the TE would represent the information transferred between A and B in precisely a 24hr period, but this would give none information whatsoever about which stations are candidates for testing risk scenarios. For this, we need a "local" measure that quantify the transfer entropy between station A and station B, but minute by minute of a day.

3.2 Local Transfer Entropy

Following [23], we would like to extract a single element from the summation in (5). In order to do that, the probability (y_{t-1}, x_t) is rewrite in its operation form, i.e. counting the number of tripletsy_{t+1} , y_t, x_t observed (namely ct) and divide it by the total number of points N in the sample:

$$p(y_{t+1}, y_t, x_t) = \frac{ct}{N} = \frac{\sum_{i=1}^{ct} 1}{N} \tag{3.6}$$

Substituting (3.6) in (3.5), we obtain:

$$T(X,Y) = \sum_{t} \frac{1}{N}\left(\sum_{i=1}^{ct} 1\right)\log_b\left(\frac{p(y_{t+1}|\, y_t, x_t)}{p(y_{t+1}\mid y_t)}\right) = \frac{1}{N}\sum_{t=1}^{N}\log_b\left(\frac{p(y_{t+1}|\, y_t, x_t)}{p(y_{t+1}\mid y_t)}\right) \tag{3.7}$$

This last expression represents an average over the weighted probability of observing y_{t+1}
giving y_t, and x_t.Taking out a particular element of (7), the following expression is obtained:

$$LTE\big(t+1, x_{j-1}, y_j\big) = \log_b\left(\frac{p(y_{t+1}|\, y_t, x_t)}{p(y_{t+1}\mid y_t)}\right) \tag{3.8}$$

This measure represents only the information transferred by a particular element xj to a particular element yj at time t+1.Equation (8) is therefore the expression applied to each station at each minute in our Oyster data set. In practice, estimating the conditional probabilities in (8) has been proved to be a very difficult task. Several approaches could be taken to accomplish this calculation [24]. In this research a Kernel density estimation is used to estimate such probabilities (section 4.2).

4. Methods

4.1 Time Series Construction

An 1140 point time series, representing the volume of people that is in a particular station *Si* in a particular time *tj*, is defined from the Victoria Line oyster data as follows:

$$V_{S_i} = \begin{cases} PTin + PWP, no\ train\ on\ S_i \\ PTin + PWP + PIT, train\ on\ S_i \end{cases} \tag{4.1}$$

Where
PTIN = People tapping in,
PWP = People walking to platform
PIT = People on a train align at Si at *ti*
A typical record extracted from the data has the following format:

Table 1. Actual record from a single transaction in the Victoria line

Entry station id	Exit station id	Entry time	Entry time
1	5	583	592

In this context, the station id is assigned as follows:

1. Walthamstow Central; 2. Blackhorse; 3. Tottenham Hale; 4. Seven Sisters; 5. Finsbury Park; 6. Highbury; 7. King's Cross St. Pancras; 8. Euston; 9. Warren Street; 10. Oxford Circus; 11. Green Park; 12. Victoria; 13. Pimlico 14. Vauxhall; 15. Stockwell; 16. Brixton

Times in table 1 are in absolute values, i.e., they represent the cumulative number of minutes from 00:00 hrs of a particular day. In this example, 583 minutes = 9.43am. In general, the number of minutes m is transformed to standard time via a simple arithmetic operation.
In (4.1) we are not taking into account people who are leaving the station. These people, although they are contributing to the total volume in a station at a particular time, are not transferring information to the next station, in the sense specified in this work. We used two additional sources of information besides the Oyster data to precisely calculate (4.1):

1) The official time table for the Victoria Line. With this, we can infer where in the network a person is and
2) An interchanging time survey, which provides the average time that would take a regular person to get into the station platform from the station entrance.

After applying (9.8) to a typical week day in the Victoria line from 4.14am to 11.59pm, we obtain 16 time series (one for each station), each one with 1185 points (one for each minute in the period defined). An extract for these time series in shown in Table 2.

These time series show the amount of people (information) that the system holds at any given minute. Particularly in Table 2 we can observe part of the morning peak. As expected, the endpoint of the line (1. Walthamstow / 16. Brixton) are not particularly crowded in the sense that people do not have to wait too much time to board a train. For example, at series 1 (corresponding of course with Station 1) from 9.00am to 9.01am only 9 people (9 bits of information) entered the system. Then, between 9.02am and

9.03am, the count goes from 59 to 31. This means that a train arrived at 9.02am and the 59 people waiting to get on board, and, at 9.03am a new set of 31 people arrived to station 1. Figure 2 shows the full time series obtained.

Table 2. Total volume of people at each station of the Victoria Line in a 5 min window

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
9am	47	84	25	97	79	80	116	94	125	123	68	211	65	101	30	58
9.01	56	40	40	61	110	75	113	89	88	184	187	169	113	60	69	74
9.02	59	65	82	81	124	81	161	101	128	102	162	179	78	129	26	65
9.03	31	36	31	46	74	83	130	121	78	178	88	201	31	62	46	61
9.04	40	33	64	48	80	102	123	104	113	154	123	193	117	89	46	71
9.05	35	59	45	98	71	82	120	111	140	98	107	119	56	58	23	42

The global maximum (689 people) between the 16 time series is reached at Oxford Circus station at 6.45pm. The morning peak maximum corresponds to Victoria station at 8.34am with 547 people. The general behaviour obtained as was expected. The classic double peaks observed are typical in cities'' transport systems.

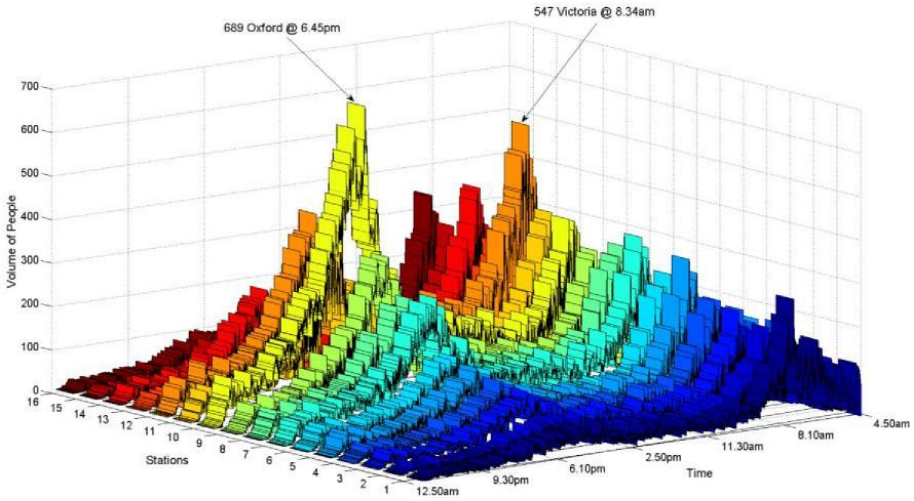

Fig. 2. Volume of people (bits of information), minute by minute, from 4.15am to 11.59pm of a typical week day.

4.2 Kernel Density Estimation (KDE)

One technique to estimate a probability for a given point *xi* of a random variable X={x_1,x_2,…,x_t} is estimate de probability distribution *f* function of X [25]. This estimation is based of approximate the shape of this unknown distribution *f* whit the following expression:

$$f_h(X) = \frac{1}{Nh}\sum_{j=1}^{N} K\left(\frac{x-x_j}{h}\right) \quad (4.2)$$

In our case, we not only need to calculate one probability, but the joint probability at an arbitrary point of the vector, y_{t+1} , y_t, x_t which can be estimated by

$$p(y_{t+1}, y_t, x_t) = \frac{1}{N(h_{yt+1}, h_{yt}, h_{xt})}\sum_{j=1}^{N} K^3\left(y_{t+1} - y_{t+1,j}\right)\cdot\left(y_t - y_{t,j}\right)\cdot\left(x_t - x_{t,j}\right) \quad (4.3)$$

Where j is the index for each of the points in X and Y and h(.) is the bandwidth for each vector

$$h_{(\cdot)} = 1.06\sigma N^{-0.2} \quad (4.4)$$

And the σ is the standard deviation in each vector (yt+1, t). The kernel K, we choose the widely used Gaussian kernel:

$$K(g) = \frac{1}{\sqrt{2\pi}} e^{-0.5g^2} \quad (4.5)$$

The rest of the probabilities in (3.8) are calculated in the same fashion marginalizing (4.3).

5. Results and discussion

Taking Table 2 as an example, if *t+1=9.03am* and *j=7*, then *x6= Highbury* and *y7= King's Cross St. Pancras*. The LTE is the amount of information transferred from y_7 at *9.02am* and *x6* at *9.03am* to y_7 at *9.03am*. In terms of (3.8), we should calculate:

$$LTE(9.03, x_6, y_7) = \log\left(\frac{p(130|161,83)}{p(130\mid 16)}\right)$$

After the calculations we obtain LTE=45.75 bytes of information is the amount of information transferred. In Figures 3 and 4 we show the complete information profile for whole 16 time series.

First, it is important to remark that Figure 3.a is showing the LTE calculation from southbound direction, Station 2 to Station 16. As there is no station before 1, there is no information transferred to it in that direction. The idea behind these profiles is, as stated, that they are able to detect at which time, which stations have the maximum values of information transferred into them, as these can be identified as important brokers in the system. The maximum value in the northbound direction is 417.7 bytes at 7.53am in Tottenham Hale. At this same station/time, the volume of people is just 215 people, while the maximum volume around this time is 547 people at Victoria station, situation that confirms our initial hypothesis that crowed stations are not necessarily key transfers of information in the system. We already mentioned that the maximum volume of people was at Oxford Circus at 6.45pm. At this station, the maximum number of bytes transferred was 265.44 at 5.33pm. More than an hour before the maximum volume of people is reached. This give us an insight on how this complex information patterns are generated in the network.

Analysing the whole profile, the different dynamic between the morning and afternoon peaks is evident. At mornings, destination stations like King's Cross (station 7) have very low LTE values, as they are not transferring information to the system (people exit there); while the end point of the line hold larger LTE values for exactly the opposite reasons: they are transferring large amounts of information into the system. In the afternoons, central stations 6, 7 and 8 are the ones with the greater LTE values in the system at that time period. However, these values are much lower than its morning counterparts. This could have something to do with the train frequency or the different choice route of many people. Further analysis in this dichotomy should be performed.

In Figure 4.a we are presenting the LTE in the northbound direction. The behaviour is very similar to the northbound, morning peak holding the maximum LTE value; morning peak and afternoon peak presenting very different patterns, etc. But, the maximum value shifts from station 3 to station 2 and almost an hour earlier. This certainly is related to the observed fact that people tend to travel early in this direction. In both directions, afternoon peak, reflect very low LTE values at the end points of the line and its surrounding stations. This is due the same reason explained for the low

values at King's Cross at mornings: this stations are exit points, nor exit or transfer points at this particular time.

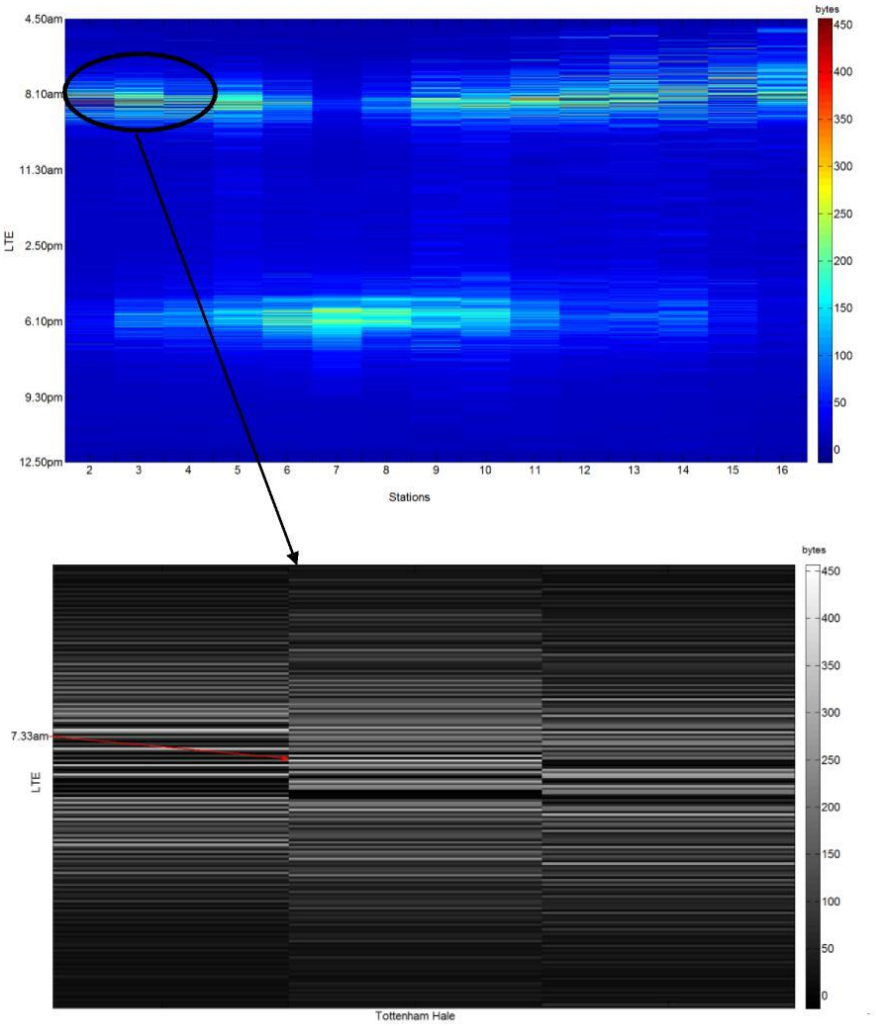

Fig. 3. Information profile of the Victoria line, southbound direction. The areas with larger LTE values are the ones identified as possible spatial temporal points serving as hubs of information in the network. a) LTE values for each minute. The **larger** values are reached at morning peak, around 7 to 10. The different behaviour between morning and afternoon

peaks is evident, showing the different dynamics empirically observed. b) Zoom at the information profile around the maximum value (7.33am at station 3). We can observed how the information transferred from on station to another changes not trivially minute by minute.

The final step is to construct the risk profile for the whole line. As a preliminary processing, we calculated the average bytes transferred at each station at each minute in both directions and then we isolated the positions (station-minute) with the highest LTE value, restringing the calculations to stations 2 to 15, as with the end points no average can be calculated as they appear only in one direction. This procedure is detecting that at mornings (around 7.40am to 8.30) a block of six stations are the most likely to cause a disruption in the line. Interesting enough, only one station at a particular minute in the afternoon was detected. It is important to notice that the volume of people at these block of stations is not too large in the morning.

6. Conclusions and future work

The local information transfer between the end stations at every single minute cannot be constructed as the sum of the transfer information between station 1 and station 2 plus transfer information between station 2 and station 3 ... station 15 and station 16. This is one of the classic footprint for a complex system. The interaction of people travelling in the system increase or decrease the total amount of information in no trivial ways, affecting how the system could respond to possible affectations at different times. Our results support the idea that although crowded stations represents a risk in terms of public security and eventually shutting down such stations could lead to major disruptions in the network, these are not the only points that could provoke a disruption. The stations identified are transferring, at key minutes, much of the information (people) through the system. There are still outstanding questions which we need to explore further. The risk profile proposed, based on the average of the information transferring it in both directions, should be tested and revised with another risk assessment techniques. After that, different disruption scenarios at the identified hubs should be tested.

Transport authorities could take advantages of studies like the one just presented, using it as a support system to evaluate, in real time, risk sensitive stations at particular time windows, and expand their contingency scenarios to these spatiotemporal positions.

Our immediate work, in terms of the Oyster data set, would be to extend the analysis to the whole network and to modify our route definition to represent people travelling via more than one line.

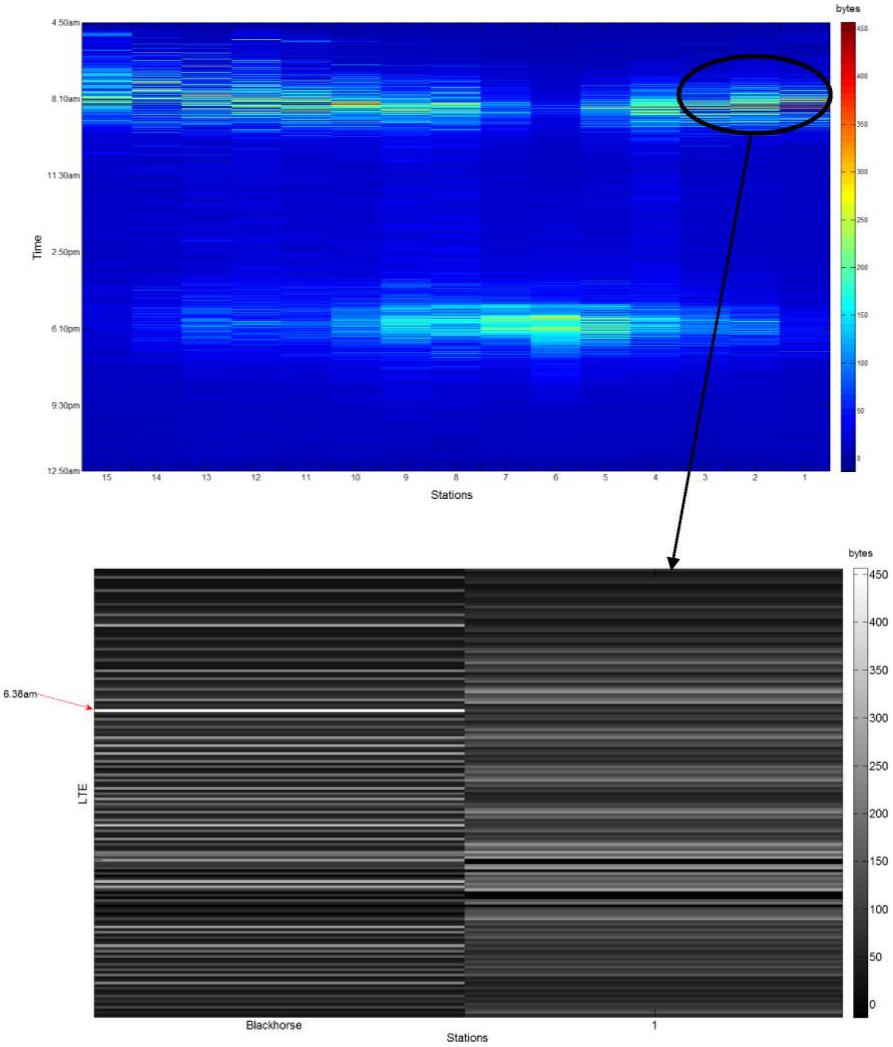

Fig. 4. Information profile the Victoria line, northbound direction. The layout is the same that in the Figure 2. a) LTE values for each minute. Again, the different behaviour between morning and afternoon peaks is evident b) Zoom at the information profile around the maximum value (6.38am at station 2, Blackhorse)

References

Sheridan, T.B. (2000). Risk Human Error, and System Resilience: Fundamental Ideas. Hum Fac Erg Soc 50 (3), 418-426

Countermeasures assessment and security experts, llc. Public Transportation Security (2007). Volume 13. Public Transportation Passenger Security Inspections: A Guide for Policy Decision Makers. Transportation research board. Washington, D.C.

Baumgarten C et al. (2014). A methodology to compare risk management (RM) systems for application and validation of specific threats in public transportation. Risk Analysis IX – A methodology to compare risk management. WIT Press, Brebbia Ed

Transport for London (2014) http://www.tfl.gov.uk. Accessed: 13 Nov. 2014.

What is Oyster? (2014) http://www.tfl.gov.uk/fares-and-ayments/oyster/what-is-oyster. Accessed 13 Nov. 2014.

Annual Report and Statement of Accounts 2013/14. Transport for London, Mayor of London

Shannon, C.E. A mathematical theory of communication (1948). Bell Syst Tech J, 27,379–423

Pelletier M-P, Trépanier M and Morency C. (2011) Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies* 19: 557- 568.

Bagchi M and White P. (2004) What role for smart-card data from bus systems? Municipal Engineer 157: 39-46.

Agard B, Morency C and Trépanier M. (2006) Mining public transport user behaviour from smart card data. 12th IFAC Symposium on Information Control Problems in Manufacturing-INCOM. 17-19.

Park JY, Kim DJ and Lim Y. (2008) Use of smart card data to define public transit use in Seoul, South Korea. Transportation Research Record: Journal of the Transportation Research Board 2063: 3-9.

Munizaga MA and Palma C. (2012) Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies 24: 9-18.

Zhong C, Huang X, Müller Arisona S, et al. (2014b) Inferring building functions from a probabilistic model using public transportation data. Computers, Environment and Urban Systems 48: 124-137.

Roth C, Kang SM, Batty M, et al. (2011) Structure of urban movements: polycentric activity and entangled hierarchical flows. PloS one 6: e15923.

Zhong C, Arisona SM, Huang X, et al. (2014a) Detecting the dynamics of urban structure through spatial network analysis. International Journal of Geographical Information Science 28: 2178-2199.

Transport for London. (2011) Travel in London, Supplementary Report: London Travel Demand Survey (LTDS).

Gordon JB, Koutsopoulos HN, Wilson NH, et al. (2013) Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. Transportation Research Record: Journal of the Transportation Research Board 2343: 17-24.

Schreiber T. (2000). Measuring Information Transfer Phys. Rev. Lett. , 85,461–464

Tung T.A., Ryu T., Lee K.H., Lee D. (2007). Inferring Gene Regulatory Networks from Microarray Time Series Data Using Transfer Entropy. Twentieth IEEE International Symposium on Computer-Based Medical Systems.

Gourévitch B, Eggermont JJ. (2007). Evaluating Information Transfer Between Auditory Cortical Neurons. AJP - JN Physiol 97 (3) 2533-2543

Baek S.K. et al. Transfer Entropy Analysis of the Stock Market. arXiv:physics/0509014

Prokopenko, M., et al. (2014). On the Cross-Disciplinary Nature of Guided Self-Organisation. In Prokopenko, M. (Ed.), Guided Self-Organization: Inception, volume 9 of Emergence, Complexity and Computation, (pp 19-51). Berlin Heidelberg, Springer.

Lizier, J.T., Prokopenko, M., Zomaya, A.Y. (2008). Local information transfer as a spatiotemporal filter for complex systems Phys Rev E 77, 026110

Lee et al. (2012) Transfer Entropy Estimation and Directional Coupling Change Detection in Biomedical Time Series BioMedical Engineering OnLine 11:19

Emanuel Parzen. The Annals of Mathematical Statistics (1962) 33 (3) 065-1076