A data-driven approach to infer spatial characteristics and service reliability of public transport hubs

Ir. Menno Yap
Dr. ir. Niels van Oort
Dr. Oded Cats
Prof. dr. ir. Serge Hoogendoorn

M.D.Yap@TUDelft.nl
https://nielsvanoort.weblog.tudelft.nl/

May 22nd, 2017

Introduction (1)

- Public transport hubs have a central role in the network
- Public transport hub characteristics (analogy airports):
  - High connectivity *(Pels, 2001)*
  - Network centrality *(Shaw 1993, Lohmann et al. 2009)*
  - Limited number of hubs in network *(Alderighi et al. 2005)*
  - Concentration of different OD-passenger flows in time and space transferring via hub *(Burghouwt, 2007)*
- Hubs important in relation to passenger reliability

Introduction (2)

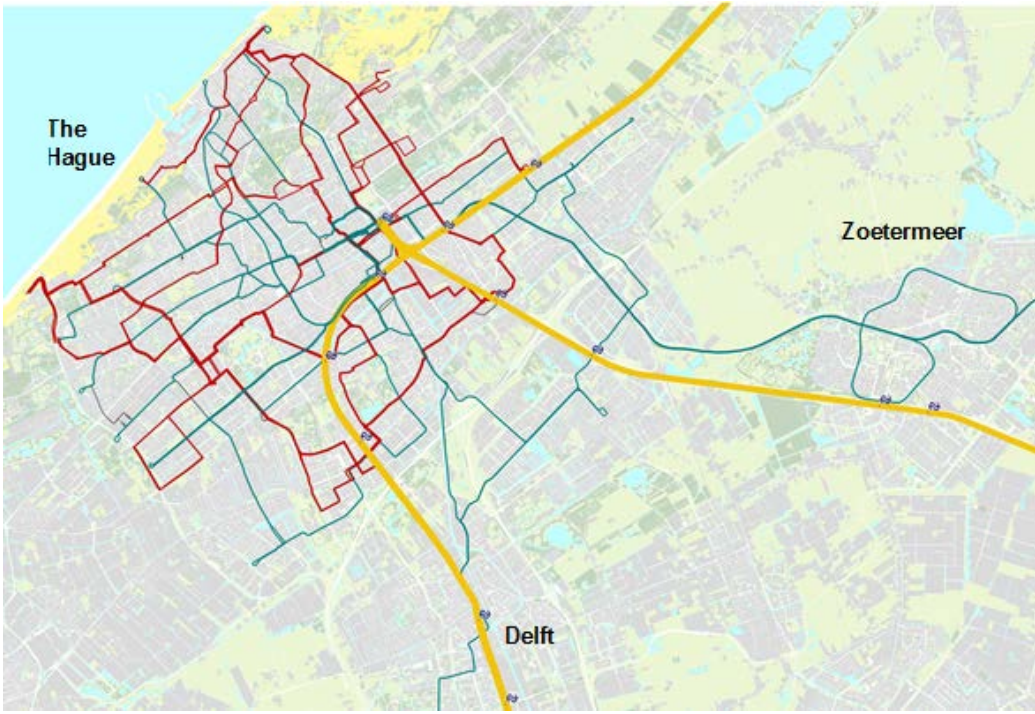
- Public transport reliability measures: from vehicle-based to passenger-based metrics
  - Punctuality
  - Regularity
- Passenger-oriented reliability measures: from trip to journey level; use of passive data sources
  - Additional passenger waiting time per line
  - Journey excess time
- Despite importance of hubs in affecting passenger reliability, no measures focusing specifically on hub reliability

Research goal

- Development of measures to quantify and compare hub reliability from a passenger perspective
  - Based on passive data sources
  - General applicable, independent of the case study network
- Research consists of three steps:
  - Infer spatial characteristics of potential hubs: which stops form a coherent cluster of transfer stops
  - Hub identification: which cluster of transfer stops concentrate substantial transfer flows in the network
  - Hub reliability: quantify and compare reliability of identified hubs
- Focus on urban public transport hubs only

Case study: network

- The Hague metropolitan area: ≈800.000 inhabitants
  - 2 light rail lines, 10 urban tram lines, 8 urban bus lines
  - 500 urban public transport stops (1650 Stop IDs), 8 train stations
  - ≈250.000 journeys per average working day (light rail + tram + bus)
  - 80% of these journeys by light rail / tram, 20% by bus



Case study: passive data sources

- Automated Fare Collection (AFC) data: entry-exit system

Tap-in date + time	Tap-in stop-ID	Tap-in line	Tap-out date + time	Tap-out stop-ID	Trip-ID	Vehicle ID	Smart-card ID
4-3-2014 11:42:37	35309	6	4-3-2014 12:03:19	34997	3423	3050	81675688
4-3-2014 12:15:57	30091	18	4-3-2014 12:23:04	32857	6545	187	81675688

- Automated Vehicle Location (AVL) data

Stop-ID	Trip-ID	Order-nr	Nominal arr	Realized arr	Nominal dep	Realized dep
1119	4464	28	19:22:35	2016-01-06 19:23:25	19:22:35	2016-01-06 19:23:49
1119	4465	28	18:23:48	2016-01-06 18:26:26	18:23:48	2016-01-06 18:26:44

- Infer vehicle occupancy by integrating AFC+AVL data
- Stop data

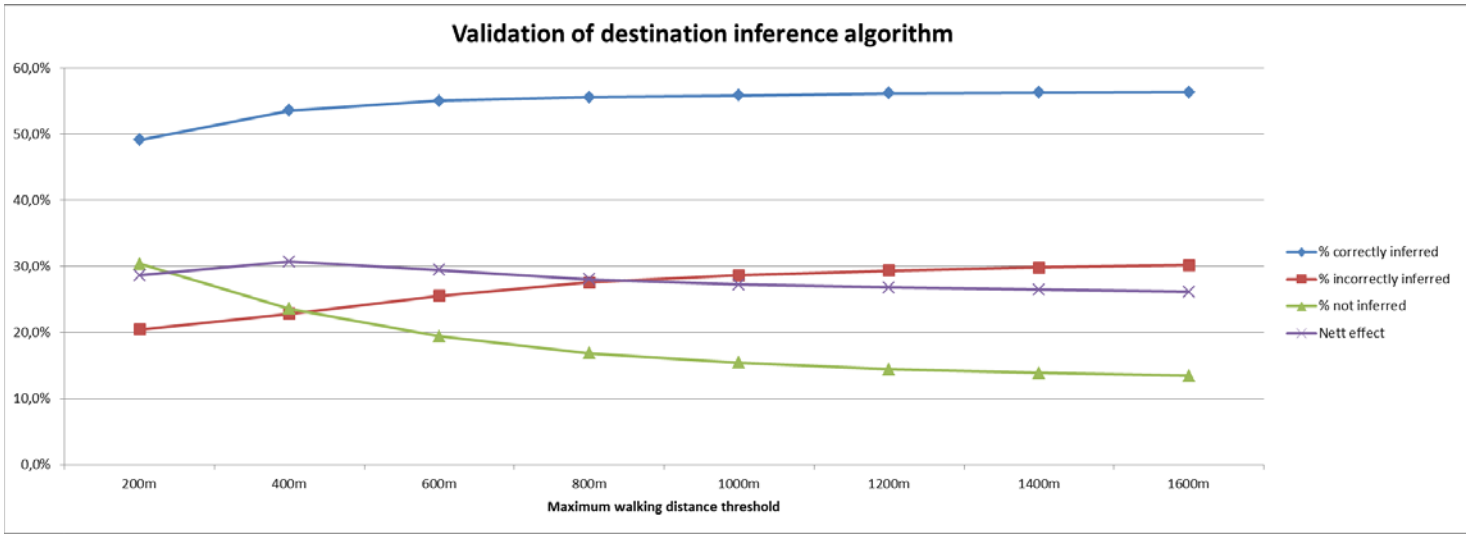
Stop-ID	RD x-coordinate	RD y-coordinate	Passenger stop name
35309	81962	450867	Dr. H. Colijnlaan
30091	82188	455213	Central Station

- For this study: data used of 1 week (Nov 23 – Nov 27 2015)

TU Delft

Data processing: destination inf.

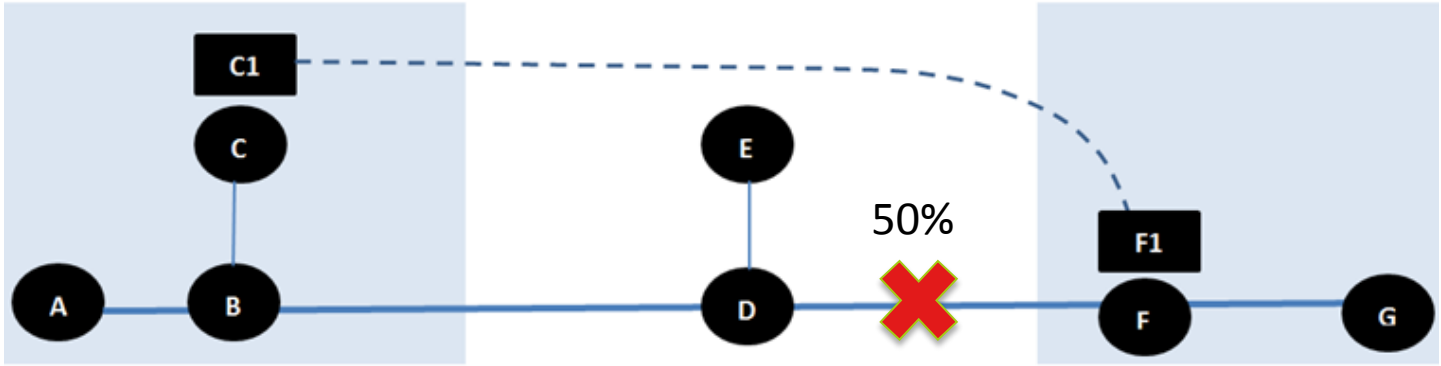
- Data cleaning (0.05 – 0.5% of daily transactions)
  - Delete system error transactions / unrealistic CoTime / missing trip ID
- Missing check-outs (1.4%): destination inference (Trépanier)
  - If $m > 1$ and $j \neq m$, alighting location of j is closest to $s^b_{p(j+1)k}$.
  - If $m > 1$ and $j = m$, alighting location of j is closest to $s^b_{p(j=1)k}$.
  - If $m = 1$, trip chaining is not possible: remove from dataset
- $d_{walk} = argmax\left(\hat{s}^{a,c}_{pjk} - \hat{s}^{a,w}_{pjk}\right)$, $d_{walk}\{d_{200}, d_{400}..d_{1600}\}$: 400 Euclidean meter



TU Delft

Data processing: transfer inference

- State-of-the-practice: $\widetilde{t_{dp(j+1)k}} \leq \widetilde{t_{apjk}} + t_{t,max}$ (e.g. 35 min)
- State-of-the-art: alighting + boarding is transfer if:
  - $l_{p(j+1)k} \neq l_{pjk}$ → what in case of short-turning, deadheading?
  - If first vehicle run $r_{lp(j+1)k}$ is taken after alighting → denied boarding?
  - $d\left(s^b_{p(j+1)k}, s^a_{pjk}\right) \leq d_{walk}$ → use intermediate PT on other network level?
- Improved transfer inference algorithm: transfer if:
  - $l_{p(j+1)k} \neq l_{pjk}$ or $l_{p(j+1)k} = l_{pjk}$ if first run after alighting r_{lpjk} is taken
  - If first vehicle run $r_{lp(j+1)k}$ is taken after alighting where $q_{lr} < capacity$
  - If first vehicle run is taken given intermediate level AVL data, $d > d_{walk}$



Spatial demarcation of potential hubs (1)

- Cluster transfer stops which form a coherent set of stops between which passenger transfer flows occur
- Clusters of transfer stops form potential hubs
- Determining clustering technique

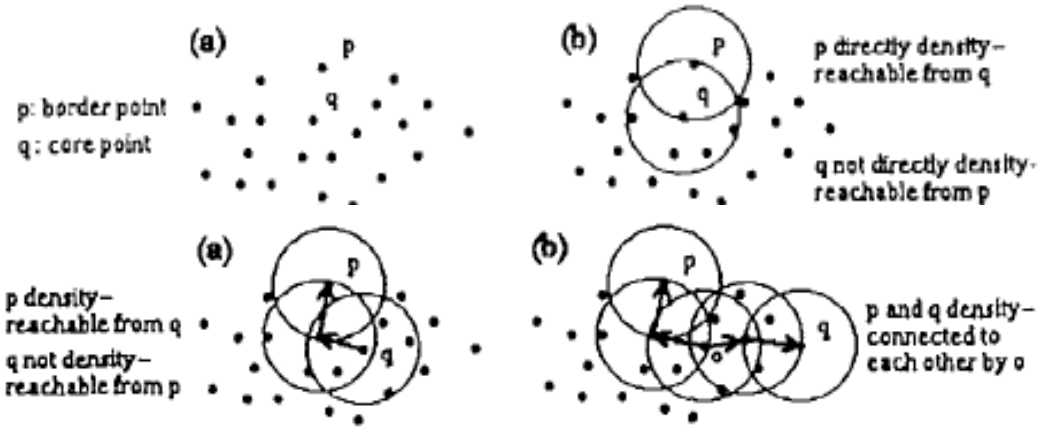
Technique Characteristics	K-means/ K-medoid	Hierarchical agglomerative clustering	DBSCAN
Pre-defined k	Pre-defined	Not pre-defined	Not pre-defined
Complete / partial	Complete	Complete	Partial
Exclusive / overlap	Exclusive	Exclusive	Overlap

- DBSCAN clustering technique applied

Spatial demarcation of potential hubs (2)

- Determination of distance measure DBSCAN:
  - Not distance based, but passenger-oriented: transfer flow based
    - $F(i,j) = F(i,j) + F(j,i) \rightarrow F(j,i) = F(i,j)$ → symmetric distance mat
    - $F(i,j) = \max(F) - F(i,j)$ → inversed, non-negative distance matrix

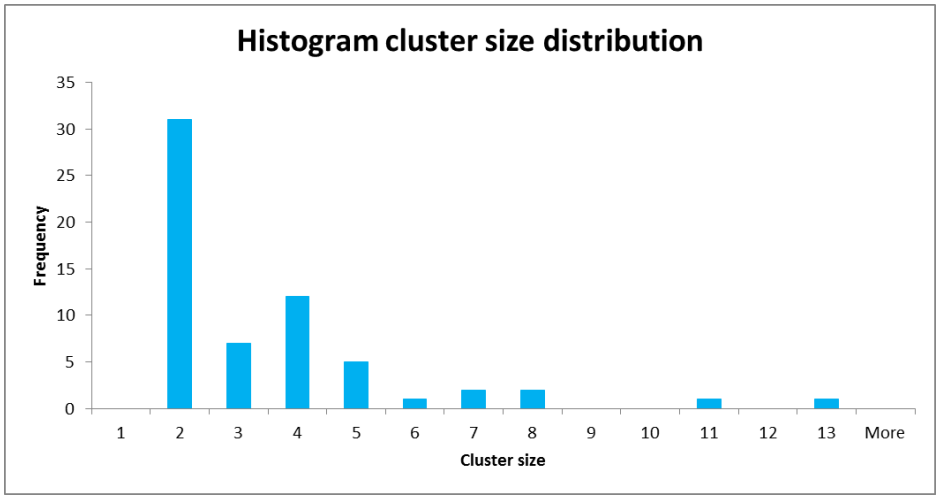
- Determination of DBSCAN parameters:
  - The neighborhood of a given radius *Eps* contains at least *MinPoint*
    - *MinPoint*: context-derived. Hub min. 2 Stop IDs → *MinPts* = 1
    - *Eps:* experiment values to check external validity → *Eps = max(F)-100*



Ester, Kriegel, Sander, Xu (1996)

Spatial demarcation of potential hubs (3)

- Resulting stops clustered by DBSCAN algorithm:
  - From 1650 StopIDs → transfers occurred between 910 StopIDs
  - 694 (76%) of these StopIDs is not clustered → 'noise'
  - Remaining 216 (24%) StopIDs clustered in 62 clusters



- Resulting transfer flows clustered by DBSCAN algorithm:
  - Maximize within-cluster transfer flows / minimize between-cluster flows
  - 86% of all network transfer flows: within-cluster transfer flows
  - 98% of transfer flows from/to clustered StopIDs: wiithin-cluster flows

Hub identification (1)

- From 62 clusters of potential hubs: which clusters concentrate substantial transfer flows to be considered a hub

- Analogy airline industry to apply economic metrics *(Costa et al. 2010; Rodriguez-Deniz et al. 2013)*
  - Use Herfindahl-Hirschman Index (HHI) to calculate market concentration based on market share of cluster i P_i:
    $HHI = \sum_{i=1}^{I} {P_i}^2$
  - Number of 'effective' market players (= hubs) $n_e = HHI^{-1}$

- Results case study network:
  - $HHI = 0.0889$, $n_e = 11.3$ → 11 hubs identified from 62 clusters of potential hubs

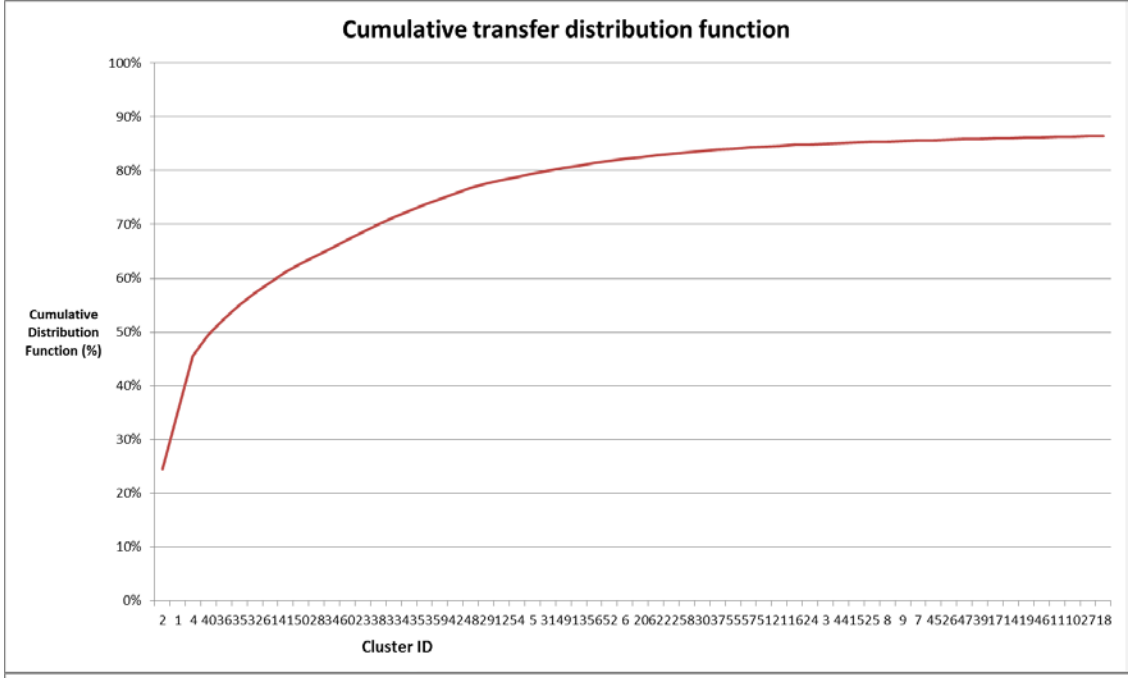
Hub identification (2)

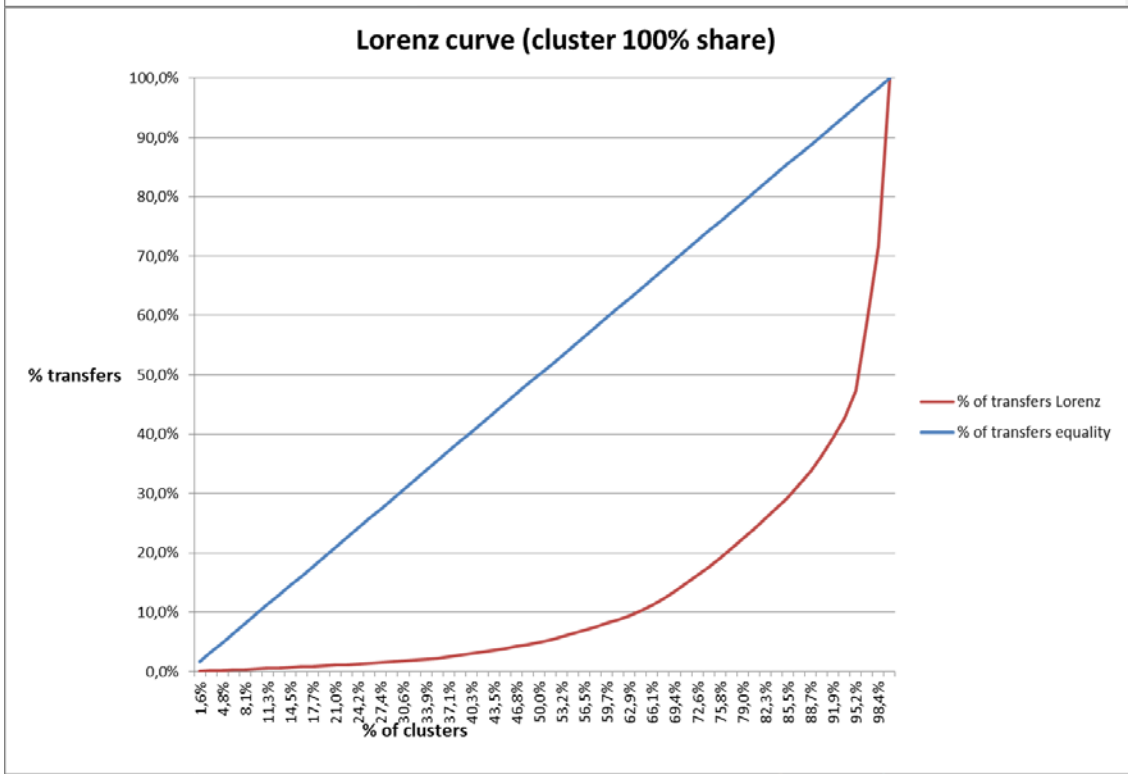
Cumulative distribution function

Lorenz curve

Gini coefficient = 0.745

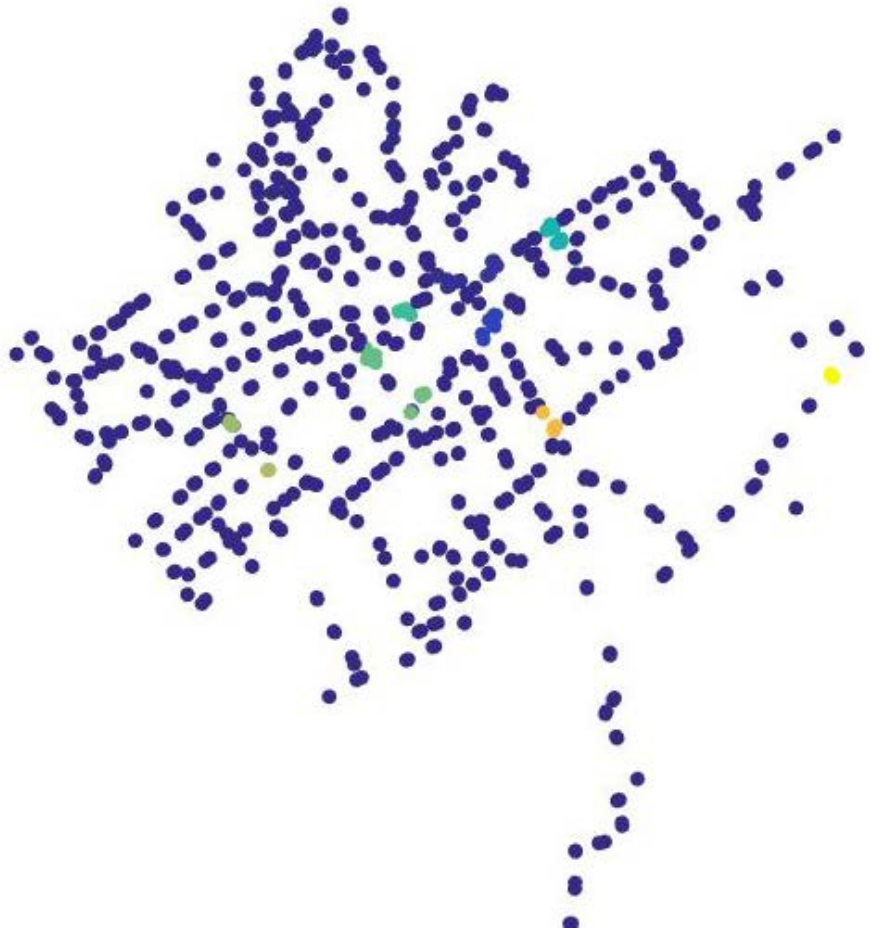
(cluster market share = 100%)





TU Delft

Hub identification (3)



Cluster ID	Hub name
1	Centrum / Spui / Kalvermarkt
2	Central Station
4	Station Hollands Spoor
28	Laan van NOI
32	Brouwersgracht
35	The Hague Market
36	Wouwermanstraat
40	Leyenburg
41	Leyweg
50	Herenstraat
61	Leidschenveen

TU Delft

Hub reliability (1)

- Hub-level passenger-oriented reliability indicators:
  - % transferring passengers missing their connection Q_{mc} at hub s_{hub}

$$Q_{mc} = \frac{\sum_l^L \sum_r^R q_{r_{l1}-r_{l2}} * MC_{q_{r_{l1}-r_{l2}}}}{\sum_l^L \sum_r^R q_{t,r_{l1}-r_{l2}}} \quad \forall\, s_{hub} \in S_{hub}$$

  - Perceived journey excess time to due to lost connection at hub

$$PJET_{mc} = \frac{\sum_l^L \sum_r^R q_{r_{l1}-r_{l2}} * MC_{q_{r_{l1}-r_{l2}}} * (T^a - T^s)}{\sum_l^L \sum_r^R q_{r_{l1}-r_{l2}} * MC_{q_{r_{l1}-r_{l2}}}} \quad \forall\, s_{hub} \in S_{hub}$$

  - Societal unreliability costs due to lost connection at the hub

$$C_{mc} = \sum_l^L \sum_r^R q_{r_{l1}-r_{l2}} * MC_{q_{r_{l1}-r_{l2}}} * (T^a - T^s) * VoT \quad \forall\, s_{hub} \in S_{hub}$$

with MC $\begin{cases} 1\ if\ r_{l2}^a > r_{l2}^s \\ 0\ if r_{l2}^a \leq r_{l2}^s \end{cases}$

Hub reliability (2)

- Example hub reliability quantification: hub Leyweg
- Average: 5.3% lost connections with on average 12 minutes additional perceived journey travel time --> yearly societal costs ≈ €18.000

Arriving line	Departing line	Lost transfer flow	Total transfer flow	Q_{mc} (%)	$PJET_{mc}$ (min)	C_{mc} (€ / year)
21	23	16	318	5%	13	1450
21	25	6	269	2%	3	146
23	21	26	477	5%	15	2664
23	25	108	1344	8%	10	7784
25	21	16	441	4%	18	1415
25	23	46	1253	4%	15	4136
Total		*218*	*4102*	*5.3%*	*12.3*	*€18.000*

Hub reliability (3)

- Yearly societal costs due to hub unreliability at all hubs (accounting for 86% of all transfers) for case study network: €386.000

Cluster ID	Hub name	Q_{mc} (%)	$PJET_{mc}$ (min)	C_{mc} (€ / year)
2	Central Station	3.6%	13.3 min	**€ 114.000**
4	Station Hollands Spoor	5.2%	11.9 min	€ 84.000
1	Centrum / Spui / Kalvermarkt	5.1%	12.1 min	€ 80.000
40	Leyenburg	3.6%	12.9 min	€ 23.000
41	Leyweg	5.3%	12.3 min	€ 18.000
50	Herenstraat	**5.5%**	12.1 min	€ 15.000
35	The Hague Market	3.1%	13.7 min	€ 15.000
61	Leidschenveen	2.1%	**24.3 min**	€ 11.000
28	Laan van NOI	4.3%	10.7 min	€ 10.000
32	Brouwersgracht	2.2%	12.8 min	€ 8.900
36	Wouwermanstraat	1.1%	14.0 min	€ 6.700

Conclusions & further research

- Conclusions:
  - Generic, data-driven methodology developed
  - To identify urban public transport network hubs
  - To quantify and compare hub (un)reliability
  - To express hub unreliability in monetary terms → SCBA

- Further research:
  - Incorporate hub connectivity / complexity explicitly in hub identification
  - Incorporate perceived in-vehicle time due to crowding as consequences of hub unreliability in $PJET_{mc}$ and C_{mc}
  - Incorporate hub unreliability in explaining passenger route choice

A data-driven approach to infer spatial characteristics and service reliability of public transport hubs

Ir. Menno Yap
Dr. ir. Niels van Oort
Dr. Oded Cats
Prof. dr. ir. Serge Hoogendoorn

M.D.Yap@TUDelft.nl
https://nielsvanoort.weblog.tudelft.nl/

May 22nd, 2017