A real-time bus arrival time information system using crowdsourced smartphone data: a novel framework and simulation experiments

Piyanit Wepulanon[a*], Agachai Sumalee[a,b], and William H.K. Lam[a]

[a]*Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China;* [b]*Department of Civil Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand*

*Corresponding author. Email: piyanit.wepulanon@connect.polyu.hk

A real-time bus arrival time information system using crowdsourced smartphone data: a novel framework and simulation experiments

This paper proposes a novel framework for developing a real-time bus arrival time information system, using crowdsourced bus information contributed by bus passengers. On one hand, passengers can derive the real-time information via their smartphones. On the other hand, they can provide some bus data in return. Particular characteristics of the participatory-based bus data are introduced. Also, a number of data processing steps are proposed in the framework to handle the data characteristics which pose extra difficulties in real-time bus arrival time prediction. The proposed system is evaluated using simulated bus data sets. Practicality of the system is investigated in terms of prediction accuracy based on different participation percentages of bus passengers.

Keywords: bus arrival time prediction; real-time system; participatory-based bus data; crowdsourced data; smartphone application

1. Introduction

Bus transport could be considered as a major public transport mode in several cities. With the advancement of sensing and communication technologies, real-time bus information systems have been broadly implemented in several cities. The systems are capable of providing beneficial bus information on the real-time basis, especially real-time bus arrival time. Thus, transit agencies can improve their management/service level and bus passengers can better plan their journeys.

Nevertheless, implementation of the real-time systems could be limited by some prohibitive reasons such as budget constraints caused by the needs of bus tracking device installation, and bus data sharing constraints caused by the competitive nature in bus operations. To tackle the limitations, this paper proposes a real-time bus arrival time information system based on crowdsourced bus information. Bus data can be

contributed by bus passengers using their smartphones, instead of installing bus tracking devices or relying on the provision of bus data from bus agencies.

Despite the advantages of participatory-based bus data acquisition, this type of bus data poses extra difficulties in real-time bus arrival time prediction. First, the bus data lacks of bus identification numbers. Second, conflicts among bus location data can be encountered. Third, the frequency of acquiring bus location data is uncertain in both spatial and temporal dimensions. The main objectives of this paper are summarized as follow:

- To introduce an alternative solution which can provide real-time bus arrival time information for some limited situations,
- To introduce essential characteristics in participatory-based bus data and propose a novel framework which could handle the particular challenges, and
- To evaluate the practicality of the real-time system in terms of prediction accuracy based on different participation levels of bus passengers.

The remainder of this paper is organized as follows. Section 2 reviews prior research on bus arrival time information systems. Section 3 presents the system architecture and operational overviews. Basic information of the system is introduced in Section 4. The details of data processing steps including bus location filtering, link travel time estimation, and bus arrival time prediction are described in Section 5, 6, and 7, respectively. Practicality of the proposed system is evaluated and discussed in section 8. Finally, the conclusion and future works are summarized in section 9.

2. Literature reviews

In terms of sensing technologies, bus arrival time information systems have been developed based on Automatic Vehicle Identification System, Automatic Passenger

Counter System, and Automatic Vehicle Location System (AVL). Due to the flexibility of AVL system in collecting vehicle GPS data, the technology has been broadly implemented to develop bus arrival time information systems in several cities (Schweiger 2011).

Apart from bus systems, various sophisticated algorithms have been developed to extract vehicular traffic information using GPS data from private vehicles including map matching algorithms (White, Bernstein, and Kornhauser 2000; Quddus et al. 2003; Lou et al. 2009; Miwa et al. 2012; Hunter, Abbeel, and Bayen 2014; Chen et al. 2014), link travel time estimation (Choi and Chung 2002; Hellinga et al. 2008; Zheng and Zuyley, 2013), and path travel time prediction (Vanitchakornpong, Indra-Payoong, and Sumalee 2013). The algorithms were developed based on the characteristics of AVL data. First, vehicle information is derived from in-vehicle tracking devices which can provide vehicle locations with the device identification number. Second, availability of vehicle information is considerably sufficient which means vehicle information can be obtained in regular time intervals. The average data sampling frequency is usually less than one minute.

Recently, researchers have started to consider a smartphone as a potential sensing device for acquiring vehicle information. Biagioni et al. (2011) proposed an automatic transit tracking system which a smartphone is placed in each vehicle for location tracking. The system makes use of the vehicle traces to predict vehicle arrival time. The results show that passenger waiting time can be significantly reduced with the availability of arrival time information.

Developing a bus information system requires the cooperation of bus agencies which is restrictive in some conditions. In Hong Kong, bus operation is a competitive system. The government does not provide any direct subsidy to the bus agencies.

Hence, they are not willing to share any information since it could affect their revenue. Although bus tracking systems are currently implemented by the bus agencies, the real-time information of all bus routes has not been disseminated within the same system. Thus, bus passengers cannot compare the bus arrival times to make the best decision.

Accordingly, there is an increasing concern about gathering bus information from bus passengers instead. The concept of voluntary participation in data collection has been studied in the field of citizen science, as some research may require intensive data collection over diverse areas for a long time period (Cohn 2008; Silvertown 2009; Hochachka et al. 2012).

In crowdsourcing research, smartphones are used as a tool for solving complex problems especially when location-based applications have become popular (Chatzimilioudis et al. 2012). Google Maps is an example of using crowdsourced mobile devices for estimating traffic conditions. The more participation of the users results in the more accurate estimation results.

For bus systems, a smartphone application can be used to establish two-way data provision systems which gather necessary information from bus passengers and provides beneficial information in return. The concept has been adopted to provide bus arrival time information. Zimmerman et al. (2011) developed a transit information system which passengers can co-produce the transit service by sharing bus information. Zhou, Zheng, and Li (2012) introduced a bus arrival time prediction system based on participatory sensing. Bus locations can be identified using multiple sensors built in passengers' smartphones. Lee and Yim (2014) proposed a system which provides the most updated bus location of individual bus routes. Table 1 summarizes the previous studies on bus arrival time information systems in different dimensions.

The two-way data provision system seems to be a viable solution for providing the real-time bus arrival time information, as bus data can be gathered from substantial amount of bus passengers without the needs of device installation. Nevertheless, essential characteristics of participatory-based bus data have not been addressed in the previous studies. The present study addresses the characteristics of participatory data which should be taken into account when developing a real-time bus arrival time information system. Also, a novel framework is proposed to handle the particular characteristics. Finally, the factors affecting system performance are discussed.

3. System architecture

The system architecture and operational overviews is demonstrated in Figure 1, while Figure 2 shows more details of data processing steps. The system is comprised of two major parts: a smartphone application, and a back-end processing server.

3.1. *Smartphone application*

On the client side, crowdsourced smartphones of bus passengers are considered as a data transmission tool. A smartphone application should be developed to fulfill two primary requirements. First, to gather bus data, participating passengers will be requested to provide some information such as the serving bus route number, and their destination bus stop. The user can simply select the information from a list generated by the system. Other information including bus location data, instantaneous bus speed, and timestamp can be periodically identified by GPS after the user pressed a start button to grant the permission for data collection. The GPS operation should be stopped when the bus is arriving to the passenger's destination. The second requirement is to disseminate predicted bus arrival times. The users can request for bus arrival time information by selecting a bus stop or a bus route number.

3.2. *Back-end processing*

On the server side, the central database is a data storage containing basic information of the system including the bus data provided by participating passengers. The information will be processed with a number of data processing steps, as can be seen from Figure 2. The data processing steps can be considered as core components of the framework. Practical algorithms can be applied in each core component. This study adopts several algorithms in the literature and adjusts some solution functions to tackle the challenges in participatory-based bus data. The results of data processing will also be recorded in the database. Finally, the information in the server can be transmitted to the clients upon smartphone application requests.

4. Basic information

One of the preliminary processes is establishing the data structure of basic information, so as to organize and access the information effectively. This section provides the notation and definition of the basic information which can be classified as fundamental information and historical information.

In order to formulate such information, a number of surveys are required to be carried out before implementing the real-time system. A group of volunteers will be asked to record bus data using a smartphone application. For the surveys, the smartphone application will continuously record a bus data set in every few seconds (1-5 seconds). However, recording bus data with the high sampling frequency is impractical for a participatory-based data provision system due to smartphone battery consumption. Therefore, the optimal frequency should be investigated and applied for the real-time system after survey periods. In this paper, the effects of data sampling

frequencies are examined in Section 8. Also, the factors to be considered for determining the frequency are discussed.

4.1. *Fundamental information*

Fundamental information includes three core elements of the system: road networks, bus routes, and participatory bus data. The manual procedures to establish the information of entire road networks and bus routes are time-consuming. In this paper, the methodologies for transit route and stop extraction proposed by Biagoni et al. (2011) are adopted to formulate such information using raw GPS traces.

4.1.1. *Road network data*

A road network can be represented using link-node representation which nodes identified as intersections or bus stops and edges as the roadways in between. Hence, a road network consists of a set of nodes denoted by $ND = \{\vec{nd}_1, ..., \vec{nd}_n\}$, where $\vec{nd}_i$ is the location vector of node identification number (node ID) i indicated by a vector of its two dimensional coordinates in latitude and longitude $(\vec{nd}_i = [nd_x \quad nd_y]^T)$.

Moreover, to determine the relative position of a GPS data on a link, the positions on a link between each pair of consecutive nodes can be referred by a set of relative locations, $C_{a,b}$. The locations could be approximately measured in an equal distance along the link (such as every 10 meters). The set of locations is denoted by $C_{a,b} = \{\bar{c}_{a,b,1}, ..., \bar{c}_{a,b,n}\}$, where $\bar{c}_{a,b,i}$ is the location vector of the i^{th} relative location on the link between node ID a and b.

4.1.2. *Bus route data*

Generally, a bus route can be recognized by a unique bus route number. As some bus

systems can be operated by multiple agencies, the same bus route number may be assigned to different routes which are operated by different agencies. The duplication can be compromised when the buses of two agencies are serving the same sequence of nodes, since passengers only expect arrival time of the first arriving bus. In other cases, the bus route number will be lacked of a clear distinction between the serving routes. Therefore, the system should be able to distinguish each unique bus route by the route number and the operator. This could avoid the ambiguities when passengers need to request for bus arrival times, and/or to report bus data.

To simplify the notation, let bn represents each bus route number of a particular operator. The trajectory of a bus route number bn operated between an origin (node ID x) and its destination (node y) is defined by a sequence of traversed nodes along the route. The node sequence is represented by a set $RN_{bn,x,y} = \{\overrightarrow{rn}_{bn,x,y,1}, ..., \overrightarrow{rn}_{bn,x,y,n}\}$. The vector $\overrightarrow{rn}_{bn,x,y,i}$ is denoted by $\overrightarrow{rn}_{bn,x,y,i} = [nd \quad st]^T$, where nd is node ID of the node's order i^{th} on the bus route, and st is a binary variable to indicate whether the node is an operating bus stop.

4.1.3. *Participatory-based bus data*

The time-ordered bus information reported by participating passengers is a collection of bus data $P = \{\vec{p}_1, \vec{p}_2, ...\}$. The vector $\vec{p}_i$ is denoted by $\vec{p}_i = [bn \quad rc_x \quad rc_y \quad v \quad t \quad d]^T$, where bn is the bus route number, rc_x and rc_y indicate the bus location, v is instantaneous speed (km/hr), t is the timestamp, and d is node ID of the passenger's destination bus stop.

4.2. *Historical information*

Historical information is necessary for bus arrival time prediction. For preliminary

stages of the system, historical bus information on the entire road network can be extracted from the fine-grained GPS bus traces since bus data sampling frequency is sufficient to provide the detailed bus trajectories. A set of historical bus information could be recorded separately for each link between node ID a and b, also for each time interval τ which is assumed to be 5 minutes in this study.

(1) Average bus travel time on each link $\overline{tt}_{a,b}^{\tau}$ can be extracted.

(2) Spatial and temporal link speed profiles can be constructed based on the average bus travel time (Vanitchakornpong, Indra-Payoong, and Sumalee 2013) in order to facilitate bus arrival time prediction processes.

(3) In the case that buses are deaccelerated and/or stopped at intersection/bus stop, the delay time at intersection/bus stop can be estimated from the time when the buses started to travel at low speed until it passed the intersection/bus stop. Accordingly, average bus delay time on each link $\overline{dt}_{a,b}^{\tau}$ can be calculated.

(4) The bus delay zone on a link $\overline{qz}_{a,b}^{\tau}$ where buses tend to travel at low speeds or stop can be identified. The delay zone can be recognized by a relative location $\bar{c}_{a,b,i}$ which is the starting point of the zone.

(5) If node b representing a bus stop, average time headway of each bus route number at the bus stop $\overline{ht}_{bn,b}^{\tau}$ can be obtained. The average values can be calculated from the difference of consecutive bus arrival times at a bus stop, in the case that survey data is sufficient to track every operating bus. Otherwise, public bus schedules can be used instead.

5. Bus location filtering

In the previous studies, bus arrival time prediction models were developed based on two

facts: bus identification numbers are available in bus location data, and bus location data is considerably sufficient for bus arrival time prediction. However, participatory-based bus data poses particular characteristics in the bus data sets.

First, the quantity and frequency of bus location data are uncertain. Second, conflicts among bus data sets can be encountered due to the lack of bus identification data as well as GPS measurement errors. The conflicts could be occurred in three cases from the data sets reported by passengers on (a) different bus lines which are operating on the same road sections, (b) on different buses of the same bus line, and (c) on the same bus but indicating different bus locations at the same time. It is unable to simply determine that which data set is more reliable.

In this section, a number of data processing steps are introduced to handle the conflicts in participatory-based bus data, filter out the unreliable data, and finally identify the most representative bus data. In this study, the first data conflict case (a), caused by the data sets from different bus lines, can be compromised by the provision of additional bus data including bus route numbers, and passengers' destination bus stops. In the following subsections, the bus sequence assignment process aims to handle the data conflict case (b) caused by the data sets from the same bus line. Whereas, the location matching process is introduced to compromise data the conflict case (c) caused by GPS errors.

5.1. *Bus sequence assignment*

Several buses are usually operated on the same service route in a time period. Suppose that each operating bus of a bus number bn which is operated from node ID x to y can be recognized by its operating sequence identification number r of a day. The objective of bus sequence assignment is to identify a particular bus operating sequence r which

were serving the passenger who reported a bus data set $\vec{p}_i$.

A heuristic approach is implemented in this study. Let $\overrightarrow{rc_i}$ denotes the GPS location of a bus route number, and $\overrightarrow{uc_r}$ denotes the location vector representing the most updated location of a bus sequence r of the same bus route number. A distance measure $S(r)$ between the two locations can be calculated

$$S(r) = dist(\overrightarrow{rc_i}, \overrightarrow{uc_r}) \tag{1}$$

where the distance function $dist(a,b)$ measures the difference between location a and b in a dimension of bus running distance on the route.

A bus operating sequence can be identified as the source of bus data $\vec{p}_i$ when the distance value is the minimum distance D

$$D = \min \sum_{i=m}^{n} S(i) \tag{2}$$

where m and n denote the minimum and maximum bus operating sequences which have not passed the GPS location based on the previous time interval. In the case that m and n is unidentified, it can be assumed that $\vec{p}_i$ is reported from the new bus operating sequence which has not been observed by the system. Hence, the new bus sequence $\max(r)+1$ will be assigned as the source of bus data $\vec{p}_i$.

For each time interval τ, a new data structure of the time-ordered bus data reported from a bus operating sequence r of a route number bn can be denoted by $P^{\tau}_{bn,r,x,y} = \{\vec{p}_{bn,r,x,y,1}, \vec{p}_{bn,r,x,y,2}, ...\}$ where $\vec{p}_{bn,r,x,y,i} = [rc_x \quad rc_y \quad v \quad t]^T$.

Although the heuristic algorithm can quickly assign a bus sequence for each bus data on the real-time basis, the algorithm is based on an assumption that a bus will not be overtaken by the following buses of the same bus line. This could pose drawbacks of

the system according to some common circumstances in bus operations e.g. bus bunching.

The necessity of bus sequence assignment is addressed as a fundamental process of the system. In the future, the methodologies to identify the bus operating sequence can be improved with more available information such as user identification number. The smartphone application may provide a log-in system to obtain an anonymous identity of the users. Therefore, the bus data reported by the same passengers can be identified.

5.2. *Location Matching*

The general map matching algorithms have been developed to identify the best matching route given a sequence of GPS locations from a probe vehicle. For bus systems, searching for the best route may not be a major concern due to the availability of bus route information. The objective of location matching could be more particular in this study: to identify the most representative bus locations on the road network given a set of GPS bus locations reported by various passengers.

This study has adopted the concept of Spatio-Temporal (ST) Matching algorithm proposed by Lou et al. (2009). The algorithm was developed to solve the matching problem for a low-sampling-rate GPS trace derived from an in-vehicle tracking device. Since the characteristics of participatory-based data are more complicated, the algorithm is modified to provide a solution for the data conflict issue. In particular, the major modifications involve the formulation of a candidate graph, as well as spatial and temporal analysis functions.

5.2.1. *Candidate location determination*

Due to GPS measurement errors, a GPS location may not be located on any road

segments. The actual bus position could be any locations within the GPS error region. Given a GPS location $\overrightarrow{rc_i} = [rc_x \quad rc_y]^T$, a set of candidate bus locations within the region is denoted by $CL^i = \{\overrightarrow{cl}^i_1, ..., \overrightarrow{cl}^i_n\}$. The vector $\overrightarrow{cl}^i_j$ is denoted by $\overrightarrow{cl}^i_j = [on \quad dn \quad rloc \quad v \quad t \quad er]^T$ where $rloc$ is the identification number of a relative location vector $\vec{c}_{on,dn,rloc}$ indicating the location on a link between node ID on and dn, v is the instantaneous speed derived with the GPS data, t is the GPS timestamp, er is the GPS errors estimated by the Euclidean distance between the GPS location $\overrightarrow{rc_i}$ and the candidate location $\vec{c}_{on,dn,rloc}$.

The number of all candidate locations in the GPS error region could be superabundant. To minimize the number of candidates, typical location matching problems could consider only one candidate on each link in the error region when the candidate provides perpendicular distance measured from the GPS location. For example, Figure 3a illustrates the corresponding candidate locations of $\overrightarrow{rc_i}$ in the GPS error region (represented by a dotted circle). Three candidates are identified including $\overrightarrow{cl}^i_1$, $\overrightarrow{cl}^i_2$, and $\overrightarrow{cl}^i_3$.

In this study, two more rules are applied to select the candidates of a bus location. First, the candidates can be selected from the links on a specific bus route since the bus route number is included in a reported bus data. Second, a candidate on each corresponding link can be selected when the candidate has the minimum distance from the GPS location. In this case, the minimum distance may not be perpendicular to a link.

Examples of candidate location determination are demonstrated in Figure 3b. Suppose that the bus location $\overrightarrow{rc_j}$ and $\overrightarrow{rc_k}$ were reported from two passengers on the same bus at the same time. The error regions of both GPS locations cover the links between node $E-F$ and $F-G$ which are the road segments on the reported bus route.

Accordingly, $\overrightarrow{rc}_j$ results in candidate locations $\overrightarrow{cl}_1^j$ and $\overrightarrow{cl}_2^j$. Here, an example of data conflict between two reported bus locations is illustrated by the candidate $\overrightarrow{cl}_1^j$ and $\overrightarrow{cl}_1^k$ since the candidates are identifying different bus locations at the same time.

It is noteworthy that the GPS error region should not be too narrow since the system is based on passenger participation. Otherwise, candidate locations could be excessively filtered out and the remaining candidates may not be sufficient to provide the real-time bus arrival time. In this study, the error region is determined based on an assumption that the reported bus data should be retained at least 90%. Distribution of GPS errors can be used to determine the error region. More details of empirical studies on statistical GPS errors are provided in Section 8.

5.2.2. *Candidate location formalization*

Candidate locations are usually sparse over the road segments. Location formalization is proposed for two objectives: to relocate the scattered locations into a comparable index, and to estimate bus arrival time at the new location. Given a candidate location in the middle of a link, the location can be relocated to a node of the link in two ways: the forward or the backward directions.

To select one of the nodes, a set of assumptions on bus travel conditions can be made based on two parameters of the candidate: the location on a link $\overrightarrow{cl}_j^i.rloc$, and instantaneous bus speed $\overrightarrow{cl}_j^i.v$. Suppose that a link can be separated into two sections based on the bus delay zone of the link: (a) regular section, and (b) delay section. The bus travel time spent on a link may not be distributed equally. Therefore, candidate locations can be formalized in four possible cases.

(1) On the regular section, a bus should travel at non-congested speeds. The candidates which are located on a regular section and traveling at non-congested instantaneous speeds will be relocated in the backward direction. The time when the bus was at the node can be estimated using the instantaneous bus speed.

(2) On the delay section, a bus can be delayed due to the next intersection/bus stop. The candidates with such conditions will be relocated in the forward direction. The bus arrival time at the node can be estimated using historical delay time since the actual delay time cannot be estimated by a candidate.

(3) A bus could travel at non-congested speed speeds on a delay section and pass an intersection/a bus stop without deceleration. The candidates will be relocated in the forward direction and the bus arrival time at the node can be estimated using historical bus speed of the link.

(4) Due to GPS errors, a bus could be located on a regular section with congested speeds. The candidates will be relocated in the backward direction, but the bus arrival time at the node should be estimated using historical bus speed of the link.

Figure 4 illustrates above formalization conditions, while Table 2 summarizes the conditions with the parameters used to estimate the bus arrival time at formalized locations. Accordingly, the location of candidates can be represented by a node. There may be the chances that several candidates are relocated to the same node location, and the estimated times at the node are different.

To facilitate further data processing steps, a directed graph can be used to organize the formalization results where each vertex represents an estimated bus arrival time at a node location and each edge represents the transmission between a pair of

node locations. Figure 5 shows an example of a directed graph. The figure is adjusted from the original one proposed by Lou et al. (2009).

The first step to formulate a graph is grouping the vertices by node locations (i.e. $\overrightarrow{nd}_i, \overrightarrow{nd}_j, ..., \overrightarrow{nd}_k$). The nodes are sorted by their sequence on a bus route $RN_{bn,x,y}$ and a new sequence index $sn_1, sn_2, ..., sn_n$ can be used to refer to the sorted nodes. Next, a set of estimated bus arrival times at the same node sequence sn is denoted by $NC^{sn} = \{\overrightarrow{nc}_1^{sn}, ..., \overrightarrow{nc}_n^{sn}\}$ where n is the total vertices representing the estimated times. The vector $\overrightarrow{nc}_k^{sn}$ is described by $\overrightarrow{nc}_k^{sn} = [loc \quad t' \quad seq \quad obs]^T$ where loc is node ID, t' is the estimated bus arrival time at the node, seq is the identification number i of the reported GPS data $\overrightarrow{rc}_i$, and obs is the GPS error inherited from the error of candidate location $\overrightarrow{cl}_j^i.err$ before performing location formalization.

In the next steps, each vertex and each edge on the graph will be associated with observation probability and transmission probability so as to determine the most representative bus arrival time at individual node locations.

5.2.3. *Observation probability*

Observation probability aims to evaluate spatial reliability of each vertex. The probability can be calculated based on the level of GPS errors $\overrightarrow{nc}_k^{sn}.obs$, and the statistical distribution of GPS errors represented by a mean μ and a standard deviation σ.

$$N(\overrightarrow{nc}_k^{sn}.obs) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{\left(\overrightarrow{nc}_k^{sn}.obs - \mu\right)^2}{2\sigma^2}} \tag{3}$$

5.2.4. *Transmission probability*

The objective of transmission analysis is to evaluate temporal reliability which is the

bus travel time between consecutive node sequences. Given the m^{th} vertex of the node sequence $i-1$ and the n^{th} vertex of the node sequence i, the travel time between node sequences can be calculated.

$$ptt_{i-1,i} = \overrightarrow{nc}_n^i.t' - \overrightarrow{nc}_m^{i-1}.t' \tag{4}$$

Transmission probability can be calculated from the likelihood between the estimated travel time, and historical travel time.

$$F_t(\overrightarrow{nc}_m^{i-1} \to \overrightarrow{nc}_n^i) = 1 - \frac{\left| ptt_{i-1,i} - \overline{ptt}_{i-1,i}^{\tau} \right|}{\overline{ptt}_{i-1,i}^{\tau}} \tag{5}$$

Let $\overline{ptt}_{i-1,i}^{\tau}$ be the average travel time from the node sequence $i-1$ to i during the time interval τ. The path travel time can be calculated from the summation of historical link travel times along the path during the same time interval τ.

5.2.5. *Result matching*

The final transmission formulation between a pair of vertices can be denoted.

$$F(\overrightarrow{nc}_m^{i-1} \to \overrightarrow{nc}_n^i) = N(\overrightarrow{nc}_n^i.obs) * F_t(\overrightarrow{nc}_m^{i-1} \to \overrightarrow{nc}_n^i) \tag{6}$$

As can be seen from Figure 5, the transmission between a pair of vertices (i.e. $\overrightarrow{nc}_1^1$ and $\overrightarrow{nc}_1^2$) from a node sequence to the next one is represented by an edge linking the vertices ($\overrightarrow{nc}_1^1 \to \overrightarrow{nc}_1^2$). The transmission formulation can be used to evaluate the transmission probability $F(\overrightarrow{nc}_1^1 \to \overrightarrow{nc}_1^2)$ of the edge.

Accordingly, a chain of vertices can be defined by the linkage of vertices from the first to the last node sequence denoted by $CP_c : \overrightarrow{nc}_{v_1}^1 \to \overrightarrow{nc}_{v_2}^2 \to ... \to \overrightarrow{nc}_{v_n}^n$ where each

vertex $\overrightarrow{nc}_{v_i}^{i}$ on the chain is the vertex number v_i of a node sequence i, n is the total node sequences on the graph, and $\overrightarrow{nc}_{v_1}^{1}.seq \neq \overrightarrow{nc}_{v_2}^{2}.seq \neq ... \neq \overrightarrow{nc}_{v_n}^{n}.seq$.

Suppose that the transmission formulation is the transmission score of a pair of vertices. A chain of vertices CP_c can also be scored by the summation of transmission scores along the chain.

$$F(CP_c) = \sum_{i=2}^{n} F(\overrightarrow{nc}_{v_{i-1}}^{i-1} \rightarrow \overrightarrow{nc}_{v_i}^{i}) \tag{7}$$

Hence, the overall score of a chain $F(CP_c)$ can be used to represent the likelihood that the bus was arrived the node sequences at the time indicated by the chain's vertices. Finally, the chain with the highest score is considered as the best solution to identify the most representative bus data.

$$CP = \arg\max_{CP_c} F(CP_c) \tag{8}$$

Given a solution $CP: \overrightarrow{nc}_{v_1}^{1} \rightarrow \overrightarrow{nc}_{v_2}^{2} \rightarrow ... \rightarrow \overrightarrow{nc}_{v_n}^{n}$, path travel time between consecutive node sequences can be obtained. Let a pair of node sequences $(i-1, i)$ represents by their node ID (a, b), the path travel time between node sequence $i-1$ and i during the time interval τ can be calculated.

$$ptt_{a,b}^{\tau} = \overrightarrow{nc}_{v_i}^{i}.t' - \overrightarrow{nc}_{v_{i-1}}^{i-1}.t' \tag{9}$$

Matching all reported bus locations results in a set of path travel times. The set of estimated path travel times during a time interval τ is denoted by $PTT^{\tau} = \{\overrightarrow{ptt}_1^{\tau}, ..., \overrightarrow{ptt}_n^{\tau}\}$, where $\overrightarrow{ptt}_i^{\tau}$ is a vector describing travel time information tt of a travel path from a node ID on to dn ($\overrightarrow{ptt}_i^{\tau} = [on \quad dn \quad tt]^T$).

Moreover, the last vector $\overrightarrow{nc}_{v_n}^{n}$ of the solution can be used to update 3 types of bus information. Let $UC_{bn,x,y}\{\overrightarrow{uc}_{bn,x,y,1},...,\overrightarrow{uc}_{bn,x,y,n}\}$ be a data set recording the most updated information of individual bus operating sequence r of a route number bn operating from the origin node ID x to destination node ID y. The information of a vector $\overrightarrow{uc}_{bn,x,y,r}$ can be updated including the most updated location cc, estimated bus arrival time at the node location t', and the bus operational status sts ($\overrightarrow{uc}_{bn,x,y,r} = [cc \quad t' \quad sts]^T$). The operational status is initially set to be 'active' and will be updated to be 'terminated' when the bus arrived its terminus.

6. Link travel time estimation

A bus travel path may cover multiple links on the bus route. The general problem of link travel time estimation is to decompose the path travel time into the travel time of individual links on the path.

In this study, the link travel time can be estimated using historical-based travel time information. Let m and n be the node ID representing $\overrightarrow{ptt}_i^{\tau}.on$ and $\overrightarrow{ptt}_i^{\tau}.dn$ respectively. The link travel time from a node ID a to the next node b on a travel path can be estimated by:

$$ltt_{a,b}^{\tau} = \frac{\overline{tt}_{a,b}^{\tau}}{\overline{ptt}_{m,n}^{\tau}} \times \overrightarrow{ptt}_i^{\tau}.tt \tag{10}$$

where $\overline{tt}_{a,b}^{\tau}$ is the historical link travel time between node a and b during time interval τ, and $\overline{ptt}_{m,n}^{\tau}$ is the historical path travel time of the travel path during time interval τ calculated from the summation of historical link travel times along the path.

The estimated travel times on a link may be varied since bus routes are usually overlapped, and several buses could travel on the same road sections during a time

period. The set of estimated travel times of a link can be denoted by $LTT_{a,b}^{\tau}\{ltt_{a,b,1}^{\tau},...,ltt_{a,b,n}^{\tau}\}$. The average link travel time can be determined by applying stratified sampling technique to the data set. In addition, spatial and temporal link speed profiles can be constructed based on the average link travel time.

7. Bus arrival time prediction

Bus arrival time at a bus stop can be predicted using (a) time at the current bus location, (b) predicted travel time between the bus location and the bus stop, and (c) bus delay time at the bus stop. The relationship can be denoted by:

$$arr_{bn,z}^{\tau+1} = \overrightarrow{uc}_{bn,x,y,r}.t' + ptt_{w,z}^{\tau+1} - \overline{dt}_{z-1,z}^{\tau+1} \quad (11)$$

where $arr_{bn,z}^{\tau+1}$ is the predicted bus arrival time for the time interval $\tau+1$ of bus route number bn at the bus stop represented by node ID z.

The most updated bus location and its timestamp can be derived from set $UC_{bn,x,y}$. Next, the path travel time $ptt_{w,z}^{\tau+1}$ between the most updated location (node ID w) and the bus stop (node ID z) during time interval $\tau+1$ needs to be predicted. Finally, bus dwell time at the bus stop will be subtracted from the predicted travel time since the delay time is already included as a part of travel time on the links. This study assumes that dwell time characteristics of all bus lines are not significantly different at the same bus stop during the same time interval. Therefore, historical bus delay time at the bus stop $\overline{dt}_{z-1,z}^{\tau+1}$ can be used to represent the predicted bus dwell time.

In fact, historical bus delay time for individual bus lines can be constructed separately to relax the assumption which could be violated since the variation in bus dwell times is usually observed. Furthermore, the possibility to perform the real-time bus dwell time prediction from participatory-based bus data could be studied.

Providing real-time bus arrival time information involves two major issues. First, travel time prediction using participatory-based bus data is challenging. Second, bus arrival time information of a bus line should be provided for every bus stop with the minimum prediction uncertainty.

7.1. *Travel time prediction*

A path travel time can be predicted by calculating the summation of predicted travel time of the links along the path. As the system will perform the prediction on the real-time basis, it is assumed that the summation of link travel times in the next time step $\tau+1$ can be used to represent the path travel time during the same time step.

In the literature, link travel time prediction algorithms make use of the link travel time in the current time step τ and previous time steps $\tau-n$, in order to predict the travel time in the next time steps ahead $\tau+n$. This means the algorithms require travel time information of a road section in every time step. However, the participatory-based data may not always provide the continuous series of bus travel time. Therefore, parametric prediction models may not be practical for the prediction.

This study applies the traffic pattern matching algorithm proposed by Vanitchakornpong, Indra-Payoong, and Sumalee (2013) to perform the real-time link travel time prediction. The objective function was developed to predict a link travel time by searching for historical traffic patterns which are most similar to the current one. The traffic patterns of a link are recognized by the spatial and temporal correlations of bus speed between the link and its adjacent links.

Without the availability of link travel time information in the current time step, the algorithm could potentially predict the travel time by considering the available

spatial and temporal correlations of the link. Furthermore, the algorithm had verified that the computational time is suitable for massive vehicle information.

The searching space of a traffic pattern can be specified using 3 parameters: spatial correlation level $(level = 1)$, temporal correlation level $(t = 3)$, and number of days $(k = 14)$. Nonetheless, the spatial and temporal correlations derived from the participatory data could be inadequate to predict the travel time of some links. In this case, the average link travel time $\overline{tt}_{a,b}^{\tau+1}$ will be assumed to represent the predicted travel time for the next time interval.

7.2. *Prediction uncertainties of bus arrival time*

For each bus line, the common expectation of bus passengers is bus arrival time of the next arriving bus. Providing the predicted bus arrival time for all operating buses is unnecessary. Therefore, the potential bus operating sequences which could provide the bus arrival time with the minimum prediction uncertainties should be identified.

First, the distance of travel path between the bus location and a bus stop could be considered. The longer distance of a travel path results in the greater prediction uncertainties. Second, the predicted bus arrival time needs to be sufficient for the continuous provision of bus arrival time information during the next time interval. Otherwise, several potential buses should be selected to perform the predictions.

In some cases, the available bus locations may be inadequate to provide bus arrival time information at every bus stop especially the first few stops on a bus route. Historical time headway at a bus stop $\overline{ht}_{bn,b}^{\tau}$ can be alternatively used to estimate the bus arrival time, based on an assumption that the variation of bus arrival times at the first few bus stops is not significant (Biagioni et al. 2011).

8. Experimental studies

To investigate the performance of the proposed system, participatory-based bus data sets are needed. A data set should include the information of multiple buses which are operating on different road sections at the same time period. Thus, microscopic simulation software called VISSIM was used to simulate the bus data in every second.

The virtual environments on a road network can be simulated in 3 dimensions: private vehicles, bus operations, and passenger demands. Private vehicle movements are determined by a car-following model, a lane-changing model, vehicle desired speed, and traffic volume. The models were determined by existing functions, for instance, Widemann's model was selected for car-following behaviors. Vehicle desired speed and traffic volume can be adjusted to include the variation in traffic conditions over multiple time periods. The desired parameters were calibrated using the traffic data from Transport Department of Hong Kong.

Next, bus operations are integrated to the road network including bus stops, bus routes, bus frequencies, and dwell time at bus stops. The distribution of bus frequencies is based on bus information provided by Kowloon Motor Bus Company, whereas the distribution of dwell times is calibrated using the observed data from field surveys. In the simulation, a bus is assigned to dwell at every bus stop on the route.

Finally, passenger demands including origin-destination and passenger arrivals are also calibrated using the observed data of passenger boarding/alighting at each bus stop. Each boarding passenger will be assigned a destination to alight from the bus. The simulation was calibrated and the outputs (e.g. traffic conditions, and passenger demand distribution) were compared with field observations. To this end, the simulation is assumed to replicate bus operations under various traffic environments and passenger demands.

The simulated road network can be represented by 74 nodes and 89 links. The average link distance is 132 meters. Total 20 bus lines were assigned to be operated on the road network. Figure 6 shows the simulated road network with examples of two bus routes. The bus routes are partially overlapped on 3 links. As the simulation provides bus tracking data in every second time unit, additional modifications are required to include smartphone GPS errors in bus locations and to simulate participatory-based bus data sets.

8.1. *Bus data modification*

The distribution of smartphone GPS errors needs to be analysed, in order to determine the magnitude of errors and integrate into bus locations. Measuring the distance error from a GPS location to the actual bus position is complicated. A general approach is analysing spatial proximity - the perpendicular distance measured from a GPS location to the closet road section.

A survey data collection was carried out by a group of volunteers to record GPS bus traces from different bus lines using their smartphones. The data collection was conducted in urban areas of Kowloon, during peak and off-peak time periods for 3 months (November 2013 to January 2013). To avoid the bias in GPS error analysis, 9,500 GPS samples were selected from the entire data sets by considering the equal amount of samples on different road sections, time periods, days of week, and smartphone models.

To this end, it is assumed that the GPS errors can represent complete noise in GPS location data. The distribution of errors can be summarized: 83.02% of total GPS locations restrain the errors within 0-30 meters, 10.32% within 30-50 meters, and the rest of 6.66% contain the errors over 50 meters. Simulated bus locations can be modified by integrating the statistical distribution into the original location data.

It is noteworthy that the error distribution can be used to identify the GPS error region for candidate location determination (Section 5.2.1). In this study, the radius of GPS regions could be 50 meters based on the assumption that at least 90% of the reported bus data should not be filtered out. Also, the mean and variance of GPS errors can be calculated and used in the observation probability function (Section 5.2.3).

The second modification is to generate the bus data reported by participating passengers. The simulated boarding/alighting passengers at each bus stop were used to determine the number of participating passengers. Accordingly, participatory-based bus data sets can be sampled from bus trajectories.

This study aims to investigate the system performance based on two sampling parameters: passenger participation, and bus data sampling frequency. Firstly, four bus data sets are generated from 1%, 3%, 5%, and 10% of the total passengers. Next, four sampling methods are applied to each data set from the first step. The sampling methods consist of both continuous sampling using different frequencies and one-time sampling when the passengers boarded a bus. The details of sampling methods are summarized in Table 3. To sum up, a simulated bus trajectory is used to generate 16 participatory-based bus data sets.

In addition, AVL-based data is also simulated to compare the performance of the proposed system with the conventional one. An AVL-based bus data set is generated in every 30 seconds. The magnitude of GPS errors is determined according to the previous studies on GPS errors in AVL-based bus locations (Lin and Zheng 1999; Jagadeesh, Srikanthan, and Zhang 2004; Jeong 2005).

8.2. *Evaluation results*

The system performance is evaluated in terms of two measures: the mean absolute error (MAE) and the mean absolute percentage error (MAPE).

$$MAE = \frac{\sum \left| ActualArr_y - SystemArr_y \right|}{N} \quad (12)$$

$$MAPE(\%) = \frac{1}{N} \frac{\sum \left| ActualArr_y - SystemArr_y \right|}{traveltime_{x,y}} \times 100 \quad (13)$$

where $ActualArr_y$ is the observed bus arrival time at a node/bus stop represented by node y, $SystemArr_y$ is the estimated/predicted bus arrival time at the node/bus stop, $traveltime_{x,y}$ is the observed travel time between the bus location on x and the bus stop location on y, and N is the number of the estimated/predicted bus arrival time.

The numerical results of two major processes are considered: bus location matching and bus arrival time prediction. Table 4 provides the numerical results derived from 17 simulated bus data sets.

8.2.1 Bus location matching performance

Location matching performance can be evaluated based on the value of (a) MAE of the estimated bus arrival times at node locations. The average MAE derived from participatory-based data is varied from 12 to 18 seconds. It can be observed that the greater number of available bus data results in the more accurate location matching. In addition, the average number of links between consecutive bus locations (b) tends to be decreased when the percentage of participating passengers and the data sampling frequency are increased. The relationship between location matching accuracy and the distance between consecutive bus locations can be explained according to the transmission probability function. The longer distance between bus locations provides the less reliable transmission probability since the variation of path travel time can be larger.

Although the data conflicts in participatory-based data can be compromised by bus location filtering, the estimated bus arrival time at node locations still contains some errors. One of underlying errors could be the GPS measurement errors originally associated with the reported bus locations. The numerical results from AVL-based bus data can be used to support the assumption. It can be observed that the MAE of bus arrival times at node locations is about 10 seconds, even though the average number of links between consecutive bus locations is close to 1.

8.2.2 Bus arrival time prediction performance

The overall performance is evaluated by MAE (d) and MAPE (e) of bus arrival time prediction. In the same way as location matching, the prediction accuracy is improved when more bus data sets are available. The trend of prediction accuracy can be demonstrated by a graph of MAPE in Figure 7. The MAPE values calculated from 17 bus data sets are plotted separately by passenger participation percentages and data sampling frequencies.

For the participatory-based data, MAE and MAPE are varied during 26.58-34.21 seconds and 26.9-34.6% respectively. The performance is not significantly different using 1-minute data sampling frequency. However, the prediction accuracy can be lower than 70% when some sampling methods are applied to the data sets with 1% and 3% of passenger participation. It can be assumed that the reported bus data is not sufficient to provide satisfactory accuracy.

The consequences of insufficient bus data can be described in three aspects. First, the prediction algorithm will rely on historical data instead of real-time data. Simulation results show that the prediction accuracy above 70% can be observed when the availability of the estimated real-time link travel time in each time interval (c) is greater than 30% of the total links. Second, the number of links between consecutive

bus locations (b) can affect prediction accuracy. The longer distance between bus locations may result in the lower accuracy in link travel time estimation/prediction. Third, the reported bus locations may be inadequate to provide the most updated location of individual operating buses for each processing time interval. Therefore, bus arrival time prediction will be based on historical time headway and result in higher prediction errors.

According to the numerical results, the system should fulfill some initial requirements in order to maintain bus arrival time prediction accuracy.

(1) For each time interval, the reported bus data should be sufficient to estimate the real-time link travel time for at least 30% of the links on the road network;
(2) The number of links between consecutive bus locations should not more than 3 to prevent significant errors in link travel estimation; and
(3) The reported bus data should be sufficient to provide the most updated location of individual buses on the road network. For instance, at least a data set is available for individual buses.

It is noteworthy that participatory-based bus data may not meet the requirements under some circumstances. For example, bus data may not be sufficient when the ridership on individual buses is low which could be encountered during off-peak hours, or in the case of high bus service frequencies.

Furthermore, the performance of participatory-based system can be compared with AVL-based bus data. The bus data from 10% of passenger participation with 1-minute sampling frequency is used for the comparison as it provides the maximum amount of reported bus data. According to Table 4, bus arrival time prediction using AVL-based data outperforms the participatory-based data about 3.3% of MAPE.

Moreover, the AVL-based data provides more real-time link travel time on 9.3% of the total links for each time interval.

To this end, it can be concluded that participatory-based bus data can be used to provide the real-time bus arrival time prediction. The prediction accuracy could be limited compared with AVL-based data. A major factor is the quantity of reported bus data which could be affected by two parameters: the number of participating passengers, and bus data sampling frequency.

8.2.3 Discussion of actual implementation

In this study, practicality of the system is investigated using simulation experiments. The results are used to identify a number of fundamental requirements for providing bus arrival time with acceptable prediction accuracy. Such analytical methods can be used as a framework when there is a need to implement the system on a road network.

First of all, participatory-based bus data sets can be simulated by taking account of the road networks, bus operations, and GPS bus data. The availability of actual characteristics is important since it could affect the number of reported bus data in each time interval. Hence, the more available data can provide the more reliable evaluation results. The crucial information consists of average distance of the links on road networks, passenger origin-destination demands, and bus frequencies.

In particular, GPS data could be unreliable in urban canyon environments. The magnitude of GPS errors should be investigated from the bus data collected during survey periods. The radius of GPS error region in this study is 50 meters based on a sample set of smartphone GPS locations in Kowloon, Hong Kong. The error region should be determined for the actual environments. Further investigation on prediction accuracy may be required for the road sections where the radius of GPS error region is larger than 50 meters.

Practicality of the system then can be evaluated using the simulated bus data. Fundamental requirements can be used as the guidelines to determine how many bus data sets are needed to provide reliable bus arrival time prediction for the road network on the real-time basis. Finally, some control parameters - including data sampling frequency and processing time interval - can be optimized so as to maximize the prediction accuracy and minimize smartphone battery consumption. The optimization can be occasionally performed according to the technology acceptance level after the system deployment.

It could be noted that the effects of processing time interval have not been investigated in this study. The parameter could affect the prediction accuracy since more bus data sets can be obtained in the longer time interval. This parameter should be investigated in the actual implementation based on the necessary computational time of adopted algorithms in the system. The time interval should not be too long since bus arrival time should be provided on the real-time basis.

To improve the system performance, another suggestion is the ways to increase user acceptance of the system. It is important to develop the smartphone application features which can provide more benefits and persuade bus passengers to share bus information.

9. Conclusion and future works

This paper proposes a novel framework for developing a real-time bus arrival time information system. The system is based on the two-way data provision concept when a smartphone application is considered as a tool for disseminating the real-time information, and gathering bus data from participating passengers. Without the needs of in-vehicle tracking devices, the system can provide an alternative solution for transit operators or governments.

The characteristics in participatory-based bus data are addressed. A number of data processing steps are proposed as core components of the framework to handle the characteristics, and to compromise GPS errors. The practicality to implement the real-time system is investigated based on the adopted solution algorithms. The results show that participatory-based bus data can be used to provide the real-time bus arrival time information. However, prediction accuracy can be varied depended on the number of reported bus data sets.

It can be noted that the solution algorithms in this study pose a set of assumptions which could be violated in some circumstances in the actual bus operation, such as bus bunching. In the future, algorithms can be developed to provide more realistic assumptions. For example, bus sequence assignment can be developed using probabilistic approaches instead of the heuristic one. Moreover, prediction accuracy can be improved by adopting other estimation/prediction models, i.e. ANN models.

The algorithm development should take account of two issues. First, the algorithms should be developed based on the characteristics of participatory-based data. Second, in cases of complex algorithms, the computational time should be evaluated since bus arrival time information has to be predicted on the real-time basis.

Acknowledgements

The work described in this paper was jointly supported by a Postgraduate Studentship and a research grant from the Research Grant Council of the Hong Kong Special Administrative Region to The Hong Kong Polytechnic University (Project No. PolyU 5242/12E).

References

Biagioni, J., Gerlich, T., Merrifield, T., and Eriksson, J. 2011. “Easytracker: automatic transit tracking, mapping, and arrival time prediction using smartphones.” In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor*

Systems, edited by Lie, J., Levis, P., and Romer, K., 68-81. New York, NY: ACM.

Cats, O., and Loutos, G. 2015. "Real-time bus arrival information system: an empirical evaluation." *Journal of Intelligent Transportation Systems*: 138-151.

Chang, H., Park, D., Lee, S., Lee, H., and Baek, S. 2010. "Dynamic multi-interval bus travel time prediction using bus transit data." *Transportmetrica* 6 (1): 19-38.

Chatzimilioudis, G., Konstantinidis, A., Laoudias, C., and Zeinalipour-Yazti, D. 2012. "Crowdsourcing with smartphones." *IEEE Transactions on Internet Computing* 16 (5): 36-44.

Chen, B. Y., Lam, W. H. K., and Tam, M. L. 2011. "Bus arrival time prediction at bus stop with multiple routes." *Transportation Research Part C: Emerging Technologies* 19 (6): 1157-1170.

Chen, B. Y., Yuan, H., Li, Q., Lam, W. H. K., Shaw, S. L., and Yan, K. 2014. "Map-matching algorithm for large-scale low-frequency floating car data." *International Journal of Geographical Information Science* 28 (1): 22-38.

Chen, M., Liu, X., Xia, J., and Chien, S. I. 2004. "A dynamic bus-arrival time prediction model based on APC data*." Computer-Aided Civil and Infrastructure Engineering* 19 (5): 364-376.

Chien, S., Ding, Y., and Wei, C. 2002. "Dynamic bus arrival time prediction with artificial neural networks." *Journal of Transportation Engineering* 128 (5): 429-438.

Choi, K., and Chung, Y. 2002. "A data fusion algorithm for estimating link travel time." *Journal of Intelligent Transportation Systems* 7 (3-4): 235-260.

Cohn, J. P. 2008. "Citizen science: Can volunteers do real research?." *BioScience* 58 (3): 192-197.

Hellinga, B., Izadpanah, P., Takada, H., and Fu, L. 2008. "Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments." *Transportation Research Part C: Emerging Technologies* 16 (6): 768-782.

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W. K. and Kelling, S. 2012. "Data-intensive science applied to broad-scale citizen science." *Trends in Ecology & Evolution* 27 (2): 130-137.

Hunter, T., Abbeel, P., and Bayen, A. 2014. “The path inference filter: model-based low-latency map matching of probe vehicle data.” *IEEE Transactions on Intelligent Transportation Systems* 15 (2): 507-529.

Jagadeesh, G. R., Srikanthan, T. and Zhang, X. D. 2004. “A map matching method for GPS based real-time vehicle location.” *Journal of Navigation* 57 (3): 429-440

Jeong, R. H. 2005. “The prediction of bus arrival time using automatic vehicle location systems data.” PhD diss., Texas A&M University.

Kuhn, K. 2011. “Open government data and public transportation.” *Journal of Public Transportation* 14 (1): 83-97.

Lee, G., and Yim, J. 2014. “Design of an Android real-time bus location provider.” *Life Science Journal* 11 (7): 619-625.

Lin, W. H., and Zeng, J. 1999. “Experimental study of real-time bus arrival time prediction with GPS data.” *Transportation Research Record: Journal of the Transportation Research Board* (1666): 101-109.

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., and Huang, Y. 2009. “Map-matching for low-sampling-rate GPS trajectories.” In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, edited by Wolfson, O., Agrawal, D., and Lu, C. T., 352-361. New York, NY: ACM.

Miwa, T., Kiuchi, D., Yamamoto, T., and Morikawa, T. 2012. “Development of map matching algorithm for low frequency probe data.” *Transportation Research Part C: Emerging Technologies* 22: 132-145.

Padmanaban, R. P. S., Divakar, K., Vanajakshi, L., and Subramanian, S. C. 2010. “Development of a real-time bus arrival prediction system for Indian traffic conditions.” *IET Intelligent Transport Systems* 4 (3): 189-200.

Quddus, M. A., Ochieng, W. Y., Zhao, L., and Noland, R. B. 2003. “A general map matching algorithm for transport telematics applications.” *GPS Solutions* 7 (3): 157-167.

Schweiger, C. L. 2011. *TCRP Synthesis 91: Use and Deployment of Mobile Device Technology for Real-Time Transit Information*. Transportation Research Board, Washington, D.C.

Silvertown, J. 2009. “A new dawn for citizen science.” *Trends in Ecology & Evolution* 24 (9): 467-471.

Transport Department of Hong Kong. 2016. “Annual Transport Digest 2016.” Accessed 1 February 2017. http://www.td.gov.hk/mini_site/atd/2016/en/section5_0.html

Vanitchakornpong, K., Indra-Payoong, N., and Sumalee, A. 2013. “Siamtraffic2.0: traffic pattern search for travel time prediction in Bangkok road network.” [In Thai.] *Journal of Information Science and Technology* 4 (1): 1-10.

White, C. E., Bernstein, D., and Kornhauser, A. L. 2000. “Some map matching algorithms for personal navigation assistants.” *Transportation Research Part C: Emerging Technologies* 8 (1): 91-108.

Zheng, F., and Zuylen, H. V. 2013. “Urban link travel time estimation based on sparse probe vehicle data.” *Transportation Research Part C: Emerging Technologies* 31: 145-157.

Zhou, P., Zheng, Y. and Li, M. 2012, “ How long to wait?: predicting bus arrival time with mobile phone based participatory sensing.” In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, edited by Davies, N., Seshan, S., and Zhong, L., 379-392. Ambleside, UK: ACM.

Zimmerman, J., Tomasic, A., Garrod, C., Yoo, D., Hiruncharoenvate, C., Aziz, R., Thiruvengadam, N. R., Huang, Y., and Steinfeld, A. 2011. “Field trial of tiramisu: crowd-sourcing bus arrival times to spur co-design.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, edited by Tan, D., Fitzpatrick, G., Gutwin, C., Begole, B., and Kellogg, W. A., 1677-1686. New York, NY: ACM.

Table 1. Summary of existing bus arrival time information systems.

Sensor technology	Prediction algorithm	Authors	Raw data accuracy	Device installation	Vehicle ID	Pattern of bus data
AVI	- SVM, - ANN, - k-NN, - Regression	Chen, Lam, and Tam 2011	High	✓	✓	Fixed location
APC	Kalman Filtering	Chen et al. 2004	Average	✓	✓	Continuous data
AVL	Historical-based	Jeong 2005; Cats and Loutos 2015	Average	✓	✓	Continuous data
	Regression	Jeong 2005; Lin and Zeng 1999				
	Kalman Filtering	Padmanaban et al. 2010				
	k-NN	Chang et al. 2010; Kuhn 2011				
	ANN	Chien, Ding, and Wei 2002; Jeong 2005;				
Smartphone GPS (in-vehicle)	Historical-based	Biagioni et al. 2011	Low	×	✓	Continuous data
Smartphone GPS (participatory)	Historical-based	Zimmerman et al. 2011	Low	×	✓	Random travel time between bus stops
Smartphone sensors (participatory)	Historical-based	Zhou, Zheng, and Li 2012	Low	×	×	Random locations in cell tower areas
Proposed system	Traffic pattern matching		Low	×	×	Random locations

Table 2. Conditions of location formalization and travel time estimation parameters.

Conditions of a candidate location	Relocation direction	Parameters used to estimate bus arrival time at the new location
(1) Regular section & regular speed	Backward	Instantaneous bus speed
(2) Delay section & congested speed	Forward	Average link delay time
(3) Delay section & regular speed	Forward	Average link speed
(4) Regular section & congested speed	Backward	Average link speed

Table 3. Bus location sampling methods.

No.	Sampling method	Sampling frequency
1	Continuous sampling	Check-in every 1 minutes
2	Continuous sampling	Check-in every 2 minutes
3	Continuous sampling	Check-in every 3 minutes
4	One-time sampling	-

Table 4. Numerical results classified by 17 types of bus data sets.

Passenger participation	Sampling method	(a) MAE of bus arrival time at a node location after performing location matching (seconds)	(b) Average no. of links between consecutive bus locations	(c) Availability of link travel time per time interval (%)	(d) MAE of bus arrival time prediction (seconds)	(e) MAPE of bus arrival time prediction (%)
AVL	-	10.34	1.14	67.98	23.14	23.61
10%	1	12.35	1.38	58.68	26.58	26.90
	2	13.38	1.64	55.71	26.91	27.13
	3	14.13	2.01	50.89	27.51	27.79
	4	15.99	2.74	41.66	30.69	28.35
5%	1	13.62	1.77	53.84	27.38	27.98
	2	14.76	2.50	46.27	28.55	28.44
	3	15.35	2.97	40.07	30.89	28.60
	4	16.03	3.57	29.23	31.25	29.88
3%	1	14.84	2.19	48.45	28.63	28.66
	2	15.76	3.49	38.29	31.30	29.49
	3	16.23	4.38	28.76	32.40	32.22
	4	N/A	N/A	<8	N/A	N/A
1%	1	15.48	2.44	34.36	31.56	29.31
	2	16.30	3.92	25.23	32.66	31.18
	3	17.34	5.97	17.53	34.21	34.69
	4	N/A	N/A	<1	N/A	N/A

Figure 1. System architecture and operational overviews.

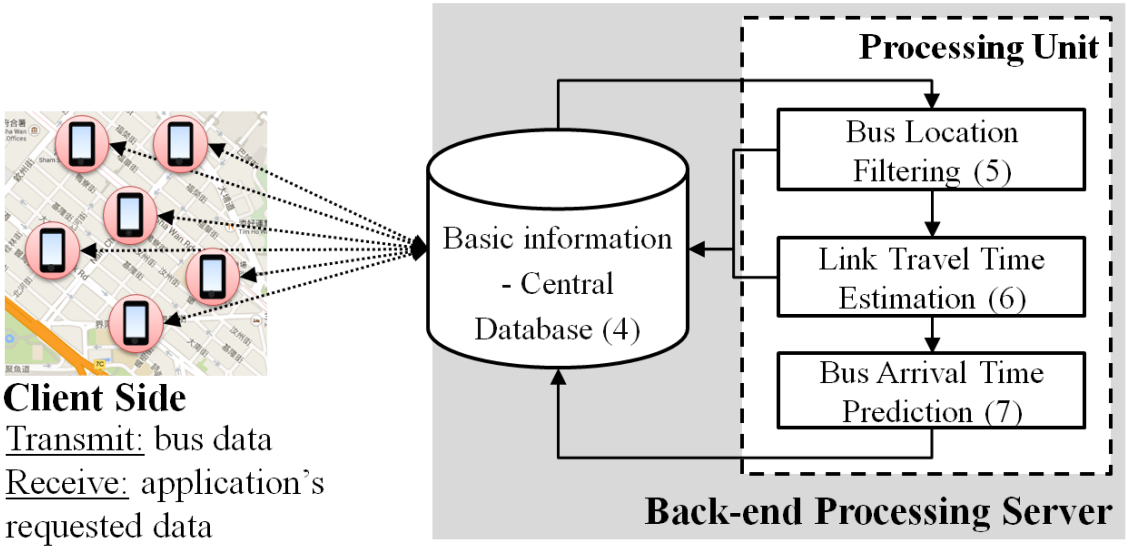

Figure 2. Details of data processing steps.

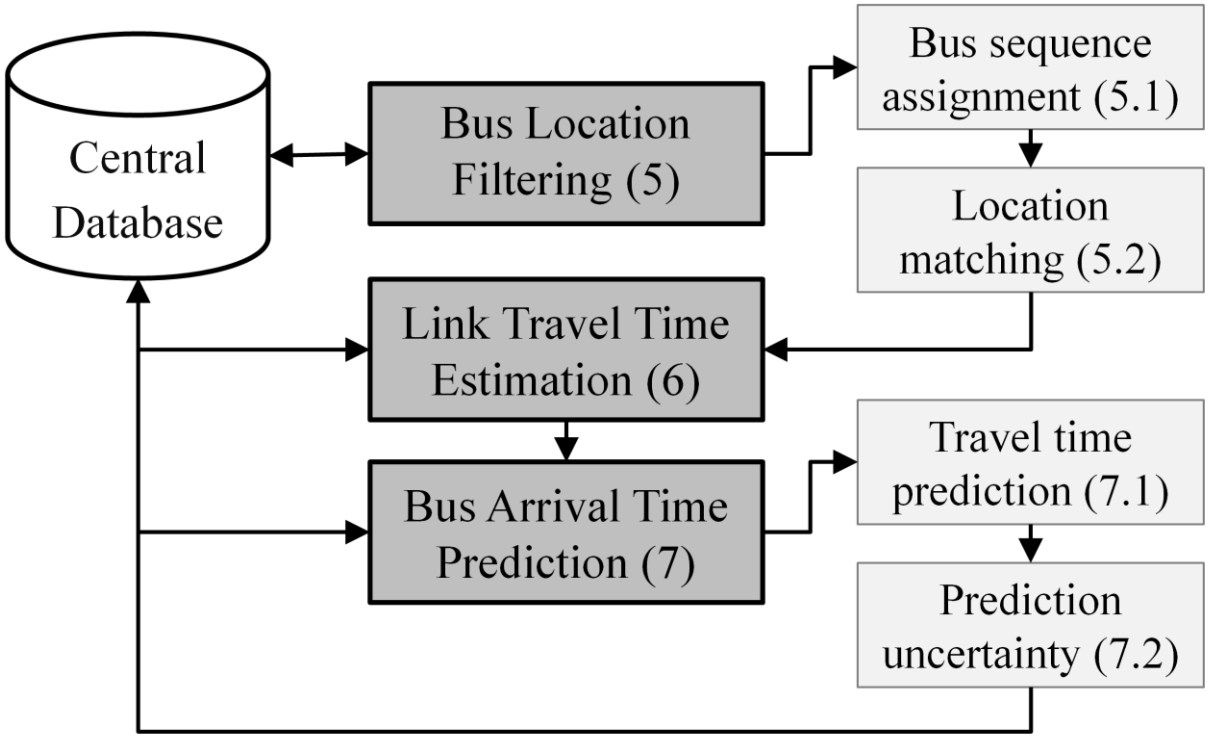

Figure 3. Examples of candidate location determination for the GPS location $\overrightarrow{rc_i}$ (a), $\overrightarrow{rc_j}$ and $\overrightarrow{rc_k}$ (b).

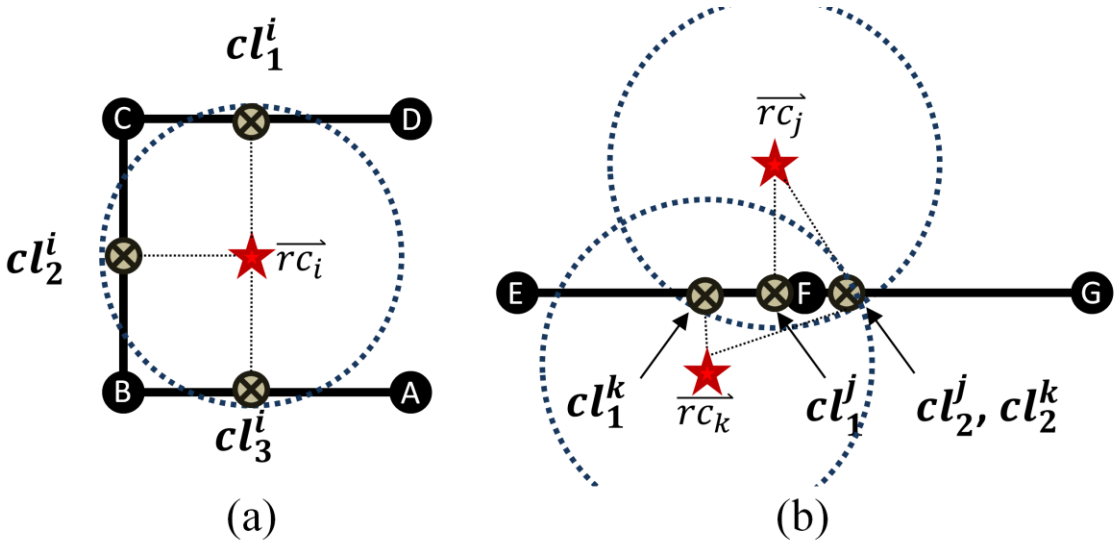

Figure 4. Conditions of location formalization.

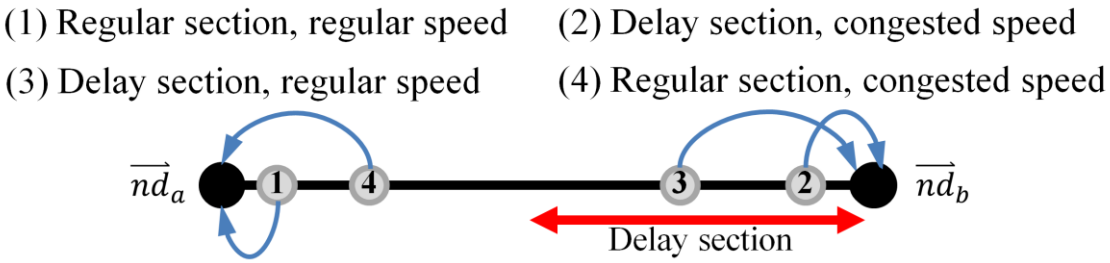

Figure 5. An example of candidate graph (Lou et al. 2009).

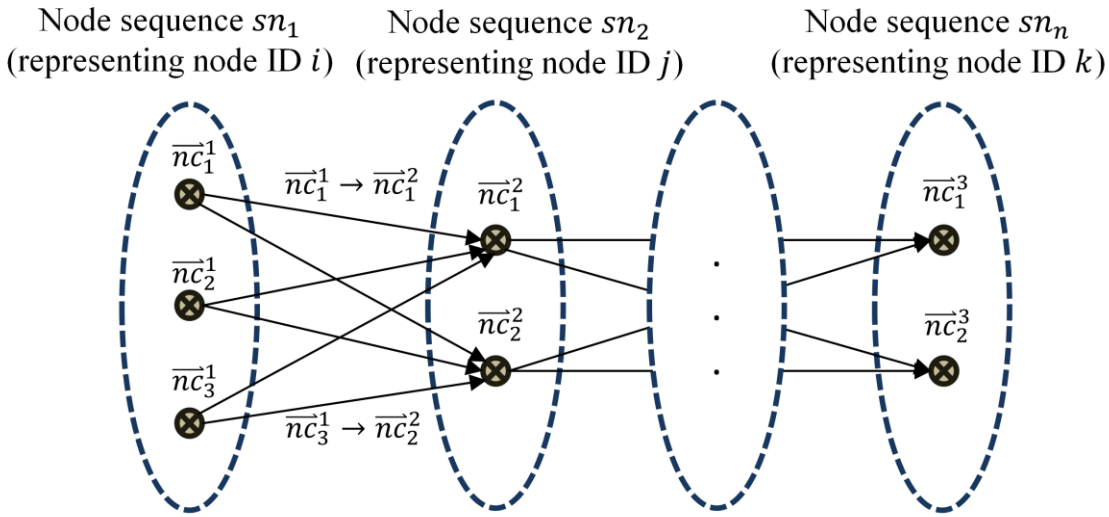

Figure 6. Simulated road network, bus stops, with examples of two bus routes.

Figure 7. MAPE of the predicted bus arrival time plotted by 17 types of bus data sets.