PAPER

# Heterogeneous Delay Tomography for Wide-Area Mobile Networks*

**Hideaki KINSHO**[†a)], **Rie TAGYO**[†], **Daisuke IKEGAMI**[†], ***Members***, **Takahiro MATSUDA**[††], ***Senior Member***, **Jun OKAMOTO**[†], ***and*** **Tetsuya TAKINE**[†††], ***Members***

**SUMMARY** In this paper, we consider network monitoring techniques to estimate communication qualities in wide-area mobile networks, where an enormous number of heterogeneous components such as base stations, routers, and servers are deployed. We assume that average delays of neighboring base stations are comparable, most of servers have small delays, and delays at core routers are negligible. Under these assumptions, we propose *Heterogeneous Delay Tomography* (HDT) to estimate the average delay at each network component from end-to-end round trip times (RTTs) between mobile terminals and servers. HDT employs a *crowdsourcing* approach to collecting RTTs, where voluntary mobile users report their empirical RTTs to a data collection center. From the collected RTTs, HDT estimates average delays at base stations in the *Graph Fourier Transform* (GFT) domain and average delays at servers, by means of *Compressed Sensing* (CS). In the crowdsourcing approach, the performance of HDT may be degraded when the voluntary mobile users are unevenly distributed. To resolve this problem, we further extend HDT by considering the number of voluntary mobile users. With simulation experiments, we evaluate the performance of HDT.

***key words:*** *delay tomography, graph Fourier transform (GFT), compressed sensing (CS), crowdsourcing, mobile network*

## 1. Introduction

Increased traffic volume due to the spread of mobile communication technologies and services leads to degradation of Quality of Service (QoS) such as throughput, packet loss rate, and delay in mobile networks. In such a situation, network monitoring is an important technique to maintain and design mobile networks. We consider network monitoring in wide-area mobile networks such as Long Term Evolution (LTE) networks, composed of an enormous number of heterogeneous network components such as base stations, routers in the core network, and servers that mobile terminals connect. In order to identify which components affect QoS degradation, we utilize *network tomography* [1]–[11], where network internal characteristics such as packet loss rates and delays are estimated from end-to-end measurements. Network tomography is a promising approach in the wide-area mobile network monitoring because it does not need to implement a measurement tool in each network component. Although a lot of network tomography schemes have been proposed, these schemes aim at estimating QoS of homogeneous network components such as link delays. In other words, the existing schemes are inapplicable in wide area mobile networks with heterogeneous components.

In this paper, we consider delay tomography, the estimation of the average delay at each network component from end-to-end round trip times (RTTs). The relationship between delays and RTTs is formulated by a system of linear equations, and delays will be estimated from it. When applying delay tomography to a wide-area mobile network, however, we have to consider two technical issues: *how to collect empirical RTTs* and *how to estimate network internal characteristics*. This paper proposes *Heterogeneous Delay Tomography* (HDT) to resolve these issues.

As for the collection of empirical RTTs, we have two approaches: active and passive measurements. In the active measurement, to collect empirical RTTs, probe packets are injected into the network from a measurement node deployed at the edge of the network. On the other hand, in the passive measurement, RTTs of user traffic are collected. The active measurement may not be suitable in wide-area mobile networks because an enormous measurement nodes have to be deployed so as to collect measurement data from the whole network. HDT utilizes a *crowdsourcing* approach, which is a kind of passive measurement methods with the help of voluntary mobile users. In the crowdsourcing approach, each voluntary mobile user implements a measurement tool in its mobile terminal, and reports RTTs measured during its sessions to a data collection center.

As for the estimation of the average delay, we have to resolve the *rank deficiency* problem and the *heterogeneity* problem of delays. The rank deficiency problem means that a system of linear equations for average delays is not full-rank even if all possible measurement paths are used. In other words, there are infinitely many candidates of the solution (i.e., an underdetermined linear inverse problem). On the other hand, the heterogeneity problem means that different types of network components have different statistical characteristics of delays. To the best of our knowledge, there have been no tomography schemes for the heterogeneity problem.

This paper considers a simple mobile network model

Manuscript received February 28, 2018.
Manuscript revised October 18, 2018.
Manuscript publicized February 6, 2019.

[†]The authors are with the NTT Network Technology Laboratories, NTT Corporation, Musashino-shi, 180-0012 Japan.
[††]The author is with the Graduate School of Systems Design, Tokyo Metropolitan University, Hino-shi, 191-0065, Japan.
[†††]The author is with the Graduate School of Engineering, Osaka University, Suita-shi, 565-0871, Japan.

*This paper was presented in part at the 2017 IEEE International Communication Quality and Reliability Workshop (CQR 2017) [17].

a) E-mail: hideaki.kinsho.fw@hco.ntt.co.jp

DOI: 10.1587/transcom.2018EBP3056

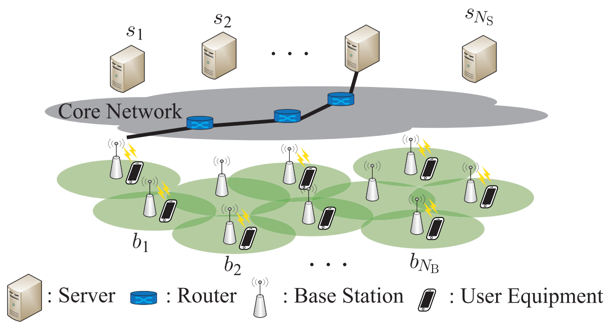



**Fig. 1** Network model.

composed of base stations, core routers, and servers (see Fig. 1) under the following assumption.

**Asspumption 1** *We assume that*

*(i) delays at base stations are spatially dependent and average delays at neighboring base stations are comparable,*
*(ii) delays at core routers are negligible, and*
*(iii) only a few servers have large average delays.*

Assumption (i) indicates that the geographical distribution of mobile users connecting to base stations are spatially dependent. Readers may refer to [12] for traffic characteristics in large-scale mobile networks. By assumption (iii), we consider that there are only a few popular servers in the mobile network.

Under these assumptions, HDT utilizes *Compressed Sensing* (CS) [13], [14] and *Graph Fourier Transform* (GFT) [15], [16]. CS can resolve the underdetermined linear inverse problem with a prior knowledge that the true solution is a sparse vector. GFT is an emerging signal processing technique for signals defined on graphs [16]. By assumption (i), we can represent average delays at base stations as an approximately sparse vector in GFT domain. Further, by assumption (iii), we can represent average delays at servers as an approximately sparse vector. Based on these facts, HDT estimates average delays at base stations in the GFT domain and average delays at servers with CS. Although many network tomography schemes have been proposed, to the best of our knowledge, there have been no methods to solve the heterogeneity problem except for HDT.

In the crowdsourcing approach to collecting empirical RTTs, measurement paths with more voluntary mobile users contribute to finer estimation of average delays. The performance of HDT, however, may be degraded due to the *user heterogeneity* problem. In [17], HDT without considering the user heterogeneity is proposed, which is referred to as HDT with *unweighted estimation* in this paper. In order to improve robustness against the user heterogeneity, we extend HDT with unweighted estimation to HDT with *weighted estimation*, where the number of voluntary mobile users in each path is taken into account.

The rest of this paper is organized as follows. We review related work in Sect. 2. In Sect. 3, we explain CS and GFT briefly. In Sect. 4, we explain HDT with unweighted and weighted estimation. In Sect. 5, we evaluate the performance of the proposed scheme with simulation experiments. Finally, we conclude the paper in Sect. 6.

## 2. Related Work

There have been several types of network tomography such as traffic matrix estimation [18], network topology estimation [19], estimation of network-internal link or node level characteristics such as packet loss rates and delays [1], [2], [4]–[11], [19]. Network tomography for estimating packet loss rates is referred to as *loss tomography* and network tomography for estimating delay characteristics is referred to as *delay tomography*. In this paper, we focus on delay tomography schemes.

Some delay tomography schemes aim at estimating the probability distributions of delays and can be classified into parametric [2], [9] and non-parametric schemes [5], [11]. In the parametric schemes, network internal delays are modeled with a certain probability distribution with the finite number of parameters. On the other hand, the non-parametric schemes do not presume a specific probability distribution. In [2], link delays are modeled with a mixture of probability distributions and it is estimated by the *General Method of Moments*. In [9], link delays are also modeled with a mixture of probability distributions and it is estimated with an Expectation Maximization (EM) algorithm. In [5], link delays are modeled with a non-parametric discrete distribution and it is estimated by using multicast measurements. In [11], link delays are modeled with a non-parametric multinomial distribution and it is estimated with an EM algorithm.

CS is applied to both loss tomography [6], [7], [10] and delay tomography [4], [8]. Delay tomography schemes with CS aim at estimating average delays, rather than their distributions. In [4], identifiability of a delay tomography scheme with CS is discussed. In [8], the synchronization-free delay tomography based on CS is proposed so as to resolve the synchronization problem between source and receiver measurement nodes. In HDT we propose, CS is used to estimate average delays. HDT, however, is different from existing schemes because it can estimate average delays with different statistical characteristics, i.e., HDT can estimate delays at base stations and servers with CS by applying graph signal processing to base station delays.

## 3. Preliminary

In this section, we explain CS and GFT briefly. These methods are utilized in HDT to estimate delays at base stations and servers. With regard to HDT, we explain the detail in Sect. 4.

### 3.1 Graph Fourier Transform (GFT) [16]

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ denote an undirected, connected graph with $N$ nodes, where $\mathcal{V} = \{v_n \mid n = 1, 2, \ldots, N\}$ and $\mathcal{E} \subseteq \mathcal{V} \times$

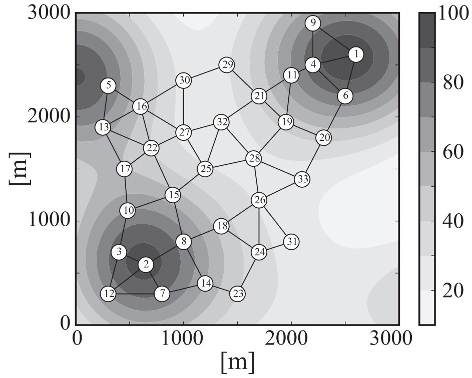


**Fig. 2** An example of spatially dependent graph signals.

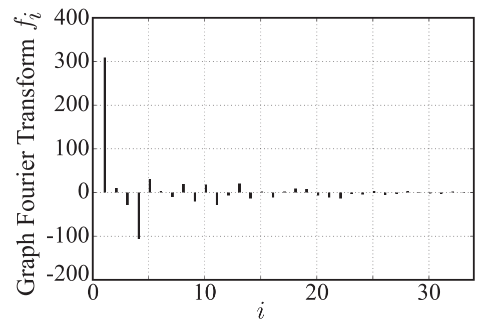


**Fig. 3** Graph Fourier Transform of the graph signal in Fig. 1.

$\mathcal{V}$ denote the sets of nodes and links, respectively. We define $\boldsymbol{A}$ and $\boldsymbol{D}$ as an $N \times N$ adjacency matrix of $\mathcal{G}$ and an $N \times N$ diagonal matrix, respectively, where the $(i, j)$-th $(i, j = 1, 2, \dots, N)$ component $a_{i,j}$ of $\boldsymbol{A}$ and the $i$-th $(i = 1, 2, \dots, N)$ diagonal component $d_i$ of $\boldsymbol{D}$ are given by

$$a_{i,j} = \begin{cases} 1, & \text{if } (v_i, v_j) \in \mathcal{E}, \\ 0, & \text{otherwise,} \end{cases} \qquad d_i = \sum_{j=1}^{N} a_{i,j}.$$

Graph Laplacian is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$. Let $\lambda_i$ and $\boldsymbol{u}_i$ $(i = 1, 2, \dots, N)$ denote the $i$-th eigenvalue of $\boldsymbol{L}$ and an eigenvector associated with $\lambda_i$. Because $\boldsymbol{L}$ is a real, symmetric, irreducible matrix, we can set $0 = \lambda_1 < \lambda_2 \le \cdots \le \lambda_N$ [20] and choose $\boldsymbol{u}_i$ $(i = 1, 2, \dots, N)$ to be orthonormal. We define an orthonormal matrix $\boldsymbol{U} = (\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_N)$, that is, $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_N$, where $\boldsymbol{I}_N$ denotes an $N \times N$ identity matrix.

We define $\boldsymbol{x} = (x_1\ x_2\ \cdots\ x_N)^\top$ as a graph signal on $\mathcal{V}$, where the $n$-th $(n = 1, 2, \dots, N)$ component $x_n$ is a signal associated with node $v_n \in \mathcal{V}$. With the orthonormal matrix $\boldsymbol{U}$, $\boldsymbol{x}$ can be represented uniquely by

$$\boldsymbol{x} = f_1 \boldsymbol{u}_1 + f_2 \boldsymbol{u}_2 + \cdots + f_N \boldsymbol{u}_N = \boldsymbol{U}\boldsymbol{f},$$

where $\boldsymbol{f} = (f_1\ f_2\ \cdots\ f_N)^\top$ is referred to as *Graph Fourier Transform* (GFT) of $\boldsymbol{x}$ and $\boldsymbol{U}$ is referred to as an *IGFT* (Inverse GFT) matrix.

Suppose graph signal $\boldsymbol{x}$ is spatially dependent and nodes within a smaller distance have more comparable values. In such a case, the GFT of $\boldsymbol{x}$ is known to be sparse, i.e., most of the components in $\boldsymbol{x}$ are zeros exactly or can be regarded as zeros approximately. Figure 2 shows an example of spatially dependent graph signals, where nodes in darker area have larger values, and Fig. 3 shows the GFT of the graph signal, which can be represented as an approximately sparse vector.

### 3.2 Compressed Sensing (CS) [13], [14]

Suppose a vector $\boldsymbol{y} \in \mathbb{R}^M$ is given by a linear transformation of vector $\boldsymbol{x} \in \mathbb{R}^N$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{1}$$

where $M$ and $N$ are natural numbers and $\boldsymbol{A}$ denotes an $M \times N$ real matrix. We consider a linear inverse problem to estimate $\boldsymbol{x}$ from given $\boldsymbol{y}$ and $\boldsymbol{A}$. Note that if rank $\boldsymbol{A} < N$, the solution of (1) is not unique. CS is applicable in such a situation, where CS selects a sparse solution, assuming $\boldsymbol{x}$ is a sparse vector.

For $p \ge 1$, we define $\ell_p$-norm $\|\boldsymbol{x}\|_p$ of $\boldsymbol{x}$ as $\|\boldsymbol{x}\|_p = (\sum_{i=1}^{N} |x_i|^p)^{1/p}$. In CS, if $\boldsymbol{x}$ is exactly sparse, we obtain an estimation $\hat{\boldsymbol{x}}_{\ell_1}$ of $\boldsymbol{x}$ by solving the following $\ell_1$ optimization problem:

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.$$

On the other hand, if $\boldsymbol{x}$ is approximately sparse, we obtain an estimation $\hat{\boldsymbol{x}}_{\ell_1\text{-}\ell_2}$ of $\boldsymbol{x}$ by solving the following $\ell_1$-$\ell_2$ optimization problem [21]:

$$\hat{\boldsymbol{x}}_{\ell_1\text{-}\ell_2} = \arg\min_{\boldsymbol{x}} \left( \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \eta \|\boldsymbol{x}\|_1 \right),$$

where $\eta$ denotes a positive parameter.

## 4. Heterogeneous Delay Tomography in Mobile Network

### 4.1 Network Model

We introduce some notations/quantities related to our network model in Fig. 1. Let $\mathcal{V}_\mathrm{B} = \{b_1, b_2, \dots, b_{N_\mathrm{B}}\}$ denote a set of base stations, where $N_\mathrm{B}$ denotes the number of base stations. We also define $\mathcal{V}_\mathrm{S} = \{s_1, s_2, \dots, s_{N_\mathrm{S}}\}$ as a set of servers, where $N_\mathrm{S}$ denotes the number of servers. Let $M$ denote the number of paths, and let $\mathcal{P} = \{p_m \mid m = 1, 2, \dots, M\}$ denote the set of paths, where $p_m = (i_m, j_m)$ $(m = 1, 2, \dots, M)$ is given by a pair of a base station $b_{i_m} \in \mathcal{V}_\mathrm{B}$ and a server $s_{j_m} \in \mathcal{V}_\mathrm{S}$. For later use, we introduce functions $\mathrm{BS}(m)$ and $\mathrm{Server}(m)$ for path $p_m = (i_m, j_m)$ $(m = 1, 2, \dots, M)$.

$$\mathrm{BS}(m) = i_m, \quad \mathrm{Server}(m) = j_m.$$

Let $N_m$ $(m = 1, 2, \dots, M)$ denotes the number of voluntary mobile users transmitting packets on path $p_m$ and we call the $k$-th $(k = 1, 2, \dots, N_m)$ mobile user on $p_m$ $(m = 1, 2, \dots, M)$ user $(m, k)$. User $(m, k)$ transmits a packet to server $s_{j_m}$ via base station $b_{i_m}$ and the server transmits its reply packet to the user. RTTs are measured at mobile terminals of the mobile users. We assume that delays in the core network and in mobile terminals are negligibly small. RTT $z$ between a base station and a server is then given by

$$z = delay_{\mathrm{up}} + delay_{\mathrm{server}} + delay_{\mathrm{down}}, \tag{2}$$

where $delay_{\mathrm{up}}$ and $delay_{\mathrm{down}}$ denote delays at the base station on the uplink and downlink directions, respectively, and $delay_{\mathrm{server}}$ denotes a delay at the server.

By setting $delay_{\mathrm{BS}} = delay_{\mathrm{up}} + delay_{\mathrm{down}}$, (2) is rewritten to be

$$z = delay_{\mathrm{BS}} + delay_{\mathrm{server}},$$

where $delay_{\mathrm{BS}}$ and $delay_{\mathrm{server}}$ are referred to as *BS delay* and *server delay*, respectively. Let $X_i$ ($b_i \in \mathcal{V}_{\mathrm{B}}$) and $Y_j$ ($s_j \in \mathcal{V}_{\mathrm{S}}$) denote random variables for BS delay at base station $b_i$ and server delay at server $s_j$, respectively. Also, let $\mu_{b_i}$ and $\sigma^2_{b_i}$ denote the mean and the variance of $X_i$, respectively, and let $\mu_{s_j}$ and $\sigma^2_{s_j}$ denote the mean and the variance of $Y_j$, respectively.

## 4.2 Crowdsourcing for Collecting Empirical RTTs

To collect empirical RTTs, we utilize *crowdsourcing*, which is a concept of outsourcing tasks to an undetermined crowd of people [22]. In the proposed delay tomography scheme, each voluntary mobile user implements a measurement tool in its mobile terminal and reports empirical RTTs to a data collection center. In what follows, voluntary mobile users are referred to as *measurement users*.

Crowdsourcing is a promising technique for data collection and it is utilized in many applications such as QoE assessment systems [22], [23] and mobile crowdsensing [24]. Note that crowdsourcing is cost-effective because any measurement nodes need not to be deployed in the network. It, however, has several issues such as privacy protection, increased traffic, battery consumption and user incentives [25]. Due to these issues, the number of measurement users is not always large, and then it causes user heterogeneity, i.e., measurement users are unevenly distributed over base stations. Therefore the number of data collected by using crowdsourcing has bias. In this paper, in order to estimate delays under the condition, we focus on the unreliability problem due to user heterogeneity, which will be discussed in Sect. 4.5.

## 4.3 Problem Formulation

Let $z_{m,k}$ denote an RTT measured by measurement user $(m, k)$.

$$z_{m,k} = x^{(p_m)}_{i_m,k} + y^{(p_m)}_{j_m,k},$$

where $x^{(p_m)}_{i_m,k}$ and $y^{(p_m)}_{j_m,k}$ denote a BS delay at base station $b_{i_m}$ and a server delay at server $s_{j_m}$, respectively, of user $(m, k)$. We define $z_m$ ($m = 1, 2, \ldots, M$) as the average of empirical RTTs on path $m$.

$$z_m = \frac{1}{N_m}\sum_{k=1}^{N_m} z_{m,k} = x^{(p_m)}_{i_m} + y^{(p_m)}_{j_m}, \tag{3}$$

where $x^{(p_m)}_{i_m}$ and $y^{(p_m)}_{j_m}$ are given by

$$x^{(p_m)}_{i_m} = \frac{1}{N_m}\sum_{k=1}^{N_m} x^{(p_m)}_{i_m,k}, \quad y^{(p_m)}_{j_m} = \frac{1}{N_m}\sum_{k=1}^{N_m} y^{(p_m)}_{j_m,k}.$$

To proceed further, we introduce some notations. We define $N_{b_i}$ and $N_{s_j}$ as the numbers of measurement users that pass through $b_i$ and $s_j$, respectively.

$$N_{b_i} = \sum_{p_m \in \mathcal{P}_{b_i}} N_m, \qquad N_{s_j} = \sum_{p_m \in \mathcal{P}_{s_j}} N_m,$$

where $\mathcal{P}_{b_i}$ and $\mathcal{P}_{s_j}$ denote sets of paths that pass through $b_i \in \mathcal{V}_{\mathrm{B}}$ and $s_j \in \mathcal{V}_{\mathrm{S}}$.

$$\mathcal{P}_{b_i} = \{p_m \in \mathcal{P} \mid \mathrm{BS}(m) = i\},$$
$$\mathcal{P}_{s_j} = \{p_m \in \mathcal{P} \mid \mathrm{Server}(m) = j\}.$$

We then define $x_i$ ($i = 1, 2, \ldots, N_{\mathrm{B}}$) and $y_j$ ($j = 1, 2, \ldots, N_{\mathrm{S}}$) as

$$x_i = \frac{1}{N_{b_i}}\sum_{p_m \in \mathcal{P}_{b_i}}\sum_{k=1}^{N_m} x^{(p_m)}_{i_m,k} = \sum_{p_m \in \mathcal{P}_{b_i}} \frac{N_m}{N_{b_i}} \cdot x^{(p_m)}_{i_m}, \tag{4}$$

$$y_j = \frac{1}{N_{s_j}}\sum_{p_m \in \mathcal{P}_{s_j}}\sum_{k=1}^{N_m} y^{(p_m)}_{j_m,k} = \sum_{p_m \in \mathcal{P}_{s_j}} \frac{N_m}{N_{s_j}} \cdot y^{(p_m)}_{j_m}.$$

Note here that $x^{(p_m)}_{i_m}$, $y^{(p_m)}_{j_m}$, $x_{i_m}$, and $y_{j_m}$ are unbiased, i.e.,

$$\mathrm{E}(x^{(p_m)}_{i_m}) = \mathrm{E}(x_{i_m}) = \mu_{b_{i_m}}, \quad \mathrm{E}(y^{(p_m)}_{j_m}) = \mathrm{E}(y_{j_m}) = \mu_{s_{j_m}}.$$

We define $\nu_m$ ($m = 1, 2, \ldots, M$) as

$$\nu_m = x^{(p_m)}_{i_m} + y^{(p_m)}_{j_m} - x_{i_m} - y_{j_m}. \tag{5}$$

(3) is then rewritten to be

$$z_m = x_{i_m} + y_{j_m} + \nu_m. \tag{6}$$

Let $\boldsymbol{z} = (z_1\ z_2\ \cdots\ z_M)^\top$, $\boldsymbol{x} = (x_1\ x_2\ \cdots\ x_{N_{\mathrm{B}}})^\top$, $\boldsymbol{y} = (y_1\ y_2\ \cdots\ y_{N_{\mathrm{S}}})^\top$, and $\boldsymbol{\nu} = (\nu_1\ \nu_2\ \cdots\ \nu_M)^\top$, which we call measurement vector, BS delay vector, server delay vector, and noise vector, respectively. We then have

$$\boldsymbol{z} = \boldsymbol{R}_{\mathrm{B}}\boldsymbol{x} + \boldsymbol{R}_{\mathrm{S}}\boldsymbol{y} + \boldsymbol{\nu}, \tag{7}$$

where $\boldsymbol{R}_{\mathrm{B}}$ denotes an $M \times N_{\mathrm{B}}$ matrix whose $(m, i)$-th component is set to be 1 if $i = \mathrm{BS}(m)$, and otherwise 0, and $\boldsymbol{R}_{\mathrm{S}}$ denotes an $M \times N_{\mathrm{S}}$ matrix whose $(m, j)$-th component is set to be 1 if $j = \mathrm{Server}(m)$, and otherwise 0.

The problem studied in this paper is to estimate BS delay vector $\boldsymbol{x}$ and server delay vector $\boldsymbol{y}$ from measurement vector $\boldsymbol{z}$. As mentioned in Sect. 1, (7) has the rank deficiency problem [17]. To see this, we consider an ideal situation that $\boldsymbol{\nu} = 0$, i.e.,

$$\boldsymbol{z} = \boldsymbol{R}_\mathrm{B}\boldsymbol{x} + \boldsymbol{R}_\mathrm{S}\boldsymbol{y} = \begin{pmatrix} \boldsymbol{R}_\mathrm{B} & \boldsymbol{R}_\mathrm{S} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}. \tag{8}$$

Obviously, $M \le N_\mathrm{B} N_\mathrm{S}$, and the equality holds if all pairs of base stations and servers are used as measurement paths.

**Theorem 1 (Theorem 1 in [17])** *Let* $\boldsymbol{R}$ *denote an* $M \times (N_\mathrm{B} + N_\mathrm{S})$ *matrix given by* $\boldsymbol{R} = (\boldsymbol{R}_\mathrm{B}\ \boldsymbol{R}_\mathrm{S})$. *We then have* rank $\boldsymbol{R} < N_\mathrm{B} + N_\mathrm{S}$ *for all* $M = 1, 2, \ldots, N_\mathrm{B} N_\mathrm{S}$.

*Proof.* By definition, the $m$-th ($m = 1, 2, \ldots, M$) row vector of $\boldsymbol{R}_\mathrm{B}$ (resp, $\boldsymbol{R}_\mathrm{S}$) is a unit vector with one at the BS($m$)-th (resp. Server($m$)-th) position. We then have $\boldsymbol{R}_\mathrm{B}\boldsymbol{1}_M = \boldsymbol{R}_\mathrm{S}\boldsymbol{1}_M = \boldsymbol{1}_M$, where $\boldsymbol{1}_M$ denotes an $M \times 1$ vector whose components are all equal to one. This implies that $(N_\mathrm{B} + N_\mathrm{S})$ column vectors of $\boldsymbol{R}$ are linearly dependent, from which the theorem follows. □

Theorem 1 indicates that we cannot determine $\boldsymbol{x}$ and $\boldsymbol{y}$ uniquely from (8) even if all measurement paths $p_m$ ($m = 1, 2, \ldots, N_\mathrm{B} N_\mathrm{S}$) are used. In what follows, we propose two estimating schemes: unweighted estimation and weighted estimation, under Assumption 1 stated in Sect. 1.

### 4.4 HDT with Unweighted Estimation [17]

We present a naive delay estimation using GFT and CS explained in Sect. 3.

#### 4.4.1 Delay Estimation Using Graph Fourier Transform and Compressed Sensing

Let $\mathcal{G}_\mathrm{B} = (\mathcal{V}_\mathrm{B}, \mathcal{E}_\mathrm{B})$ denote an undirected graph for base stations, where $\mathcal{V}_\mathrm{B} = \{b_i \mid i = 1, 2, \ldots, N_\mathrm{B}\}$ and $\mathcal{E}_\mathrm{B} \subset \mathcal{V}_\mathrm{B} \times \mathcal{V}_\mathrm{B}$. Let $e \in \mathcal{E}_\mathrm{B}$ denote a *virtual link* between two base stations. We define $\boldsymbol{\alpha}_i \in \mathbb{R}^2$ as the location of base station $b_i$ and $\mathrm{dist}(\boldsymbol{\alpha}_{i_1}, \boldsymbol{\alpha}_{i_2}) = \|\boldsymbol{\alpha}_{i_1} - \boldsymbol{\alpha}_{i_2}\|_2$ as the Euclidean distance between base stations $b_{i_1}$ and $b_{i_2}$. For $b_{i_1}, b_{i_2} \in \mathcal{V}_\mathrm{B}$, $(b_{i_1}, b_{i_2}) \in \mathcal{E}_\mathrm{B}$ if and only if $\mathrm{dist}(\boldsymbol{\alpha}_{i_1}, \boldsymbol{\alpha}_{i_2}) \le D_\mathrm{th}$.

Let $\boldsymbol{U}_\mathrm{B}$ denote an IGFT matrix of $\mathcal{G}_\mathrm{B}$. (7) is then rewritten to be

$$\boldsymbol{z} = \boldsymbol{R}_\mathrm{B}\boldsymbol{U}_\mathrm{B}\boldsymbol{f} + \boldsymbol{R}_\mathrm{S}\boldsymbol{y} + \boldsymbol{\nu}, \tag{9}$$

where $\boldsymbol{f}$ denotes the GFT of $\boldsymbol{x}$.

Under Assumption 1 (iii), server delay vector $\boldsymbol{y}$ are approximately sparse. Furthermore, Assumption 1 (i) implies that the GFT $\boldsymbol{f}$ of BS delay vector $\boldsymbol{x}$ is approximately sparse. We thus have

$$\boldsymbol{f} = \boldsymbol{f}_\mathrm{sparse} + \Delta\boldsymbol{f}, \quad \boldsymbol{y} = \boldsymbol{y}_\mathrm{sparse} + \Delta\boldsymbol{y}, \tag{10}$$

where $\boldsymbol{f}_\mathrm{sparse}$ and $\boldsymbol{y}_\mathrm{sparse}$ are sparse vectors, and $\Delta\boldsymbol{f}$ and $\Delta\boldsymbol{y}$ denote error terms such that $\|\Delta\boldsymbol{f}\|_2 \ll 1$ and $\|\Delta\boldsymbol{y}\|_2 \ll 1$.

Using (10), we re-define $\boldsymbol{f}$ and $\boldsymbol{y}$ as $\boldsymbol{f} := \boldsymbol{f}_\mathrm{sparse}$ and $\boldsymbol{y} := \boldsymbol{y}_\mathrm{sparse}$. It then follows from (9) that

$$\boldsymbol{z} = \boldsymbol{R}_\mathrm{B}\boldsymbol{U}_\mathrm{B}\boldsymbol{f} + \boldsymbol{R}_\mathrm{S}\boldsymbol{y} + \boldsymbol{\omega}, \tag{11}$$

where $\boldsymbol{\omega} = \boldsymbol{R}_\mathrm{B}\boldsymbol{U}_\mathrm{B}\Delta\boldsymbol{f} + \boldsymbol{R}_\mathrm{S}\Delta\boldsymbol{y} + \boldsymbol{\nu}$. By regarding $\boldsymbol{\omega}$ as a noise vector, we can estimate them from $\boldsymbol{z}$ by using CS. Let $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{y}}$ denote estimated vectors of $\boldsymbol{f}$ and $\boldsymbol{y}$, respectively. We obtain $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{y}}$ by solving the following $\ell_1$-$\ell_2$ optimization problem:

$$(\hat{\boldsymbol{f}}, \hat{\boldsymbol{y}}) = \mathop{\arg\min}_{\boldsymbol{f}, \boldsymbol{y}} \left\{ \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{R}_\mathrm{B}\boldsymbol{U}_\mathrm{B}\boldsymbol{f} - \boldsymbol{R}_\mathrm{S}\boldsymbol{y}\|_2^2 + \eta_1\|\boldsymbol{f}\|_1 + \eta_2\|\boldsymbol{y}\|_1 \right\}, \tag{12}$$

where $\eta_i$ ($i = 1, 2$) are positive parameters. Note that an estimation $\hat{\boldsymbol{x}}$ of BS delay vector $\boldsymbol{x}$ is obtained by $\hat{\boldsymbol{x}} = \boldsymbol{U}_\mathrm{B}\hat{\boldsymbol{f}}$. We solve (12) by Algorithm 1, where we estimate $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{y}}$ iteratively until the convergence criteria are met. In this paper, we use FISTA (First Iterative Shrinkage-Thresholding Algorithm) [14], [26] to estimate $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{y}}$.

**Algorithm 1:** Iterative algorithm for (12).

**Input :** $\boldsymbol{R}_\mathrm{B}, \boldsymbol{R}_\mathrm{S}, \boldsymbol{U}_\mathrm{B}, \eta_1, \eta_2$, measurement vector $\boldsymbol{z} \in \mathbb{R}^M$, and the stopping criteria $\epsilon$ ($\epsilon > 0$).
**Output:** Estimated BS delay vector in the GFT domain $\hat{\boldsymbol{f}} \in \mathbb{R}^{N_\mathrm{B}}$ Estimated server delay vector $\hat{\boldsymbol{y}} \in \mathbb{R}^{N_\mathrm{S}}$.

$\hat{\boldsymbol{y}} := \boldsymbol{0}_{N_\mathrm{S}}$, $\hat{\boldsymbol{f}} := \boldsymbol{0}_{N_\mathrm{B}}$.
**repeat**
  $\boldsymbol{f}_\mathrm{prev} := \hat{\boldsymbol{f}}$, $\boldsymbol{y}_\mathrm{prev} := \hat{\boldsymbol{y}}$.
  $\hat{\boldsymbol{f}} := \mathop{\arg\min}_{\boldsymbol{f}} \left( \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{R}_\mathrm{B}\boldsymbol{U}_\mathrm{B}\boldsymbol{f} - \boldsymbol{R}_\mathrm{S}\hat{\boldsymbol{y}}\|_2^2 + \eta_1\|\boldsymbol{f}\|_1 \right)$.
  $\hat{\boldsymbol{y}} := \mathop{\arg\min}_{\boldsymbol{y}} \left( \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{R}_\mathrm{B}\boldsymbol{U}_\mathrm{B}\hat{\boldsymbol{f}} - \boldsymbol{R}_\mathrm{S}\boldsymbol{y}\|_2^2 + \eta_2\|\boldsymbol{y}\|_1 \right)$.
**until**

$$\left( \frac{\|\hat{\boldsymbol{f}} - \boldsymbol{f}_\mathrm{prev}\|_2}{\|\boldsymbol{f}_\mathrm{prev}\|_2} \le \epsilon \right) \wedge \left( \frac{\|\hat{\boldsymbol{y}} - \boldsymbol{y}_\mathrm{prev}\|_2}{\|\boldsymbol{y}_\mathrm{prev}\|_2} \le \epsilon \right).$$

### 4.5 HDT with Weighted Estimation

In general, measurement users are unevenly distributed over base stations and popular servers are connected from many mobile users. Therefore it is natural that the numbers $N_m$'s of measurement users on different paths would be different from each other, which we call user heterogeneity.

Note that the unweighted estimation in Sect. 4.4 treats the average delays $z_m$'s of all paths equally, regardless of the values of $N_m$, as shown in (3), (9), and (12). We first show that user heterogeneity yields large variance in the unweighted estimation, and then we propose a weighted estimation method so as to alleviate the effect of user heterogeneity.

#### 4.5.1 Effect of the Number of Measurement Users

We demonstrate that user heterogeneity degrades the performance of the unweighted estimation. For simplicity, we assume that $\{x_{i_m,k}^{(p_m)}\}$ and $\{y_{j_m,k}^{(p_m)}\}$ are independent identically distributed samples from their respective distributions. It then follows from (5) that the first two moments of noise factor $\nu_m$ of path $m$ are given by $\mathrm{E}[\nu_m] = 0$ and

$$\begin{aligned}\mathrm{E}[\nu_m^2] &= \mathrm{E}[(x_{i_m}^{(p_m)} - x_{i_m})^2] + \mathrm{E}[(y_{j_m}^{(p_m)} - y_{j_m})^2] \\ &= \frac{1-\beta_m^{\mathrm{B}}}{\beta_m^{\mathrm{B}}} \cdot \frac{\sigma_{b_{i_m}}^2}{N_{b_{i_m}}} + \frac{1-\beta_m^{\mathrm{S}}}{\beta_m^{\mathrm{S}}} \cdot \frac{\sigma_{s_{j_m}}^2}{N_{s_{j_m}}}, \end{aligned} \tag{13}$$

where $\beta_m^{\mathrm{B}}$ (resp. $\beta_m^{\mathrm{S}}$) denotes the ratio of the number $N_m$ of measurement users on path $m$ to the total number $N_{b_{i_m}}$ (resp. $N_{s_{j_m}}$) of measurement users accessing base station $i_m = \mathrm{BS}(m)$ (resp. server $j_m = \mathrm{Server}(m)$).

$$\beta_m^{\mathrm{B}} = \frac{N_{b_{i_m}}}{N_m}, \qquad \beta_m^{\mathrm{S}} = \frac{N_{s_{j_m}}}{N_m}.$$

Note that factor $\sigma_{b_{i_m}}^2/N_{b_{i_m}}$ (resp. $\sigma_{s_{j_m}}^2/N_{s_{j_m}}$) in (13) is common for all paths with base station $b_{i_m}$ (resp. server $s_{j_m}$), and the $(1-\beta_m^{\mathrm{B}})/\beta_m^{\mathrm{B}}$ and $(1-\beta_m^{\mathrm{S}})/\beta_m^{\mathrm{S}}$ determine the contribution of the common factors to the variance $\mathrm{E}[v_m^2]$. For example, consider two paths $m_1$ and $m_2$ such that $\mathrm{BS}(m_1) = \mathrm{BS}(m_2) = b_1$ and $\beta_{m_1}^{\mathrm{B}} < \beta_{m_2}^{\mathrm{B}}$ (i.e., $N_{m_1} < N_{m_2}$). We then have

$$\frac{1-\beta_{m_1}^{\mathrm{B}}}{\beta_{m_1}^{\mathrm{B}}} \Bigg/ \frac{1-\beta_{m_2}^{\mathrm{B}}}{\beta_{m_2}^{\mathrm{B}}} > \frac{\beta_{m_2}^{\mathrm{B}}}{\beta_{m_1}^{\mathrm{B}}} = \frac{N_{m_2}}{N_{m_1}}.$$

Note that a similar observation can be made for server delay. The above discussion shows that the average of empirical RTTs on a path with relatively small numbers of measurement users has relatively large variance. Therefore, if we deal with all paths equally, the accuracy of the estimation may be degraded due to large variance of the noise term in paths with small numbers of measurement users. In what follows, we propose a weighted estimator so as to alleviate the effect of user heterogeneity.

#### 4.5.2 Weighted Estimation According to the Number of Measurement Users

We first identify the theoretically optimal weights. Let $X$ denote a random variable with mean $\mu$ and variance $\sigma^2$ and there are $N$ samples of $X$ (which may represents a BS delay or a server delays). We divide the $N$ samples into $K$ groups $\boldsymbol{Q}_k = \{q_{k,1}, q_{k,2}, \ldots, q_{k,N_k}\}$ $(k = 1, 2, \ldots, K)$, where $q_{k,n}$ $(n = 1, 2, \ldots, N_k)$ denotes the $n$-th sample in $\boldsymbol{Q}_k$. Let $q_k = (q_{k,1} + q_{k,2} + \ldots, +q_{k,N_K})/N_k$ $(k = 1, 2, \ldots, K)$. Note that $q_k$ is a random variable with mean $\mu$ and variance $\sigma^2/N_k$. For a set of $q_k$ $(k = 1, 2, \ldots, K)$, an estimator $\hat{\mu}$ of $\mu$ is called Best Linear Unbiased Estimator (BLUE) if (i) $\hat{\mu}$ is a linear function of $q_k$ $(k = 1, 2, \ldots, K)$, (ii) $\hat{\mu}$ is unbiased (i.e., $\mathrm{E}[\hat{\mu}] = \mu$), and (iii) the variance of $\hat{\mu}$ is minimum among all estimators satisfying (i) and (ii).

**Theorem 2** *The BLUE $\hat{\mu}$ of $\mu$ is given by*

$$\hat{\mu} = \sum_{k=1}^{K} \frac{N_k}{N} \cdot q_k.$$

*Proof.* Due to the shortage of space, we only provide an outline of the proof. It is clear that a linear estimator $\mu_{\mathrm{a}} = \sum_{k=1}^{K} a_k q_k$ is unbiased iff $\sum_{k=1}^{K} a_k = 1$. Furthermore, the variance $\sigma_{\mathrm{a}}^2$ of $\mu_{\mathrm{a}}$ is given by $\sigma_{\mathrm{a}}^2 = \sigma^2 \sum_{k=1}^{K} a_k^2/N_k$. Therefore $\{a_k;\ k = 1, 2, \ldots, K\}$ for the BLUE is given by the solution of the following linear optimization problem.

$$\min \sum_{k=1}^{K} \frac{a_k^2}{N_k} \quad \text{subject to} \quad \sum_{k=1}^{K} a_k = 1.$$

It is easy to verify that the solution of the above problem is $a_k = N_k/N$ $(k = 1, 2, \ldots, K)$. □

Theorem 2 suggests that if $x_{i_m}^{(p_m)}$ $(p_m \in \mathcal{P}_{b_i})$ and $y_{j_m}^{(p_m)}$ $(p_m \in \mathcal{P}_{s_j})$ were given, it would be reasonable to adopt the BLUEs of $x_i$ and $y_j$:

$$x_i = \sum_{p_m \in \mathcal{P}_{b_i}} \frac{N_m}{N_{b_i}} \cdot x_i^{(p_m)}, \quad y_j = \sum_{p_m \in \mathcal{P}_{s_j}} \frac{N_m}{N_{s_j}} \cdot y_j^{(p_m)}.$$

Note here that the contributions of $x_i^{(p_m)}$ and $y_j^{(p_m)}$ to $x_i$ and $y_j$ in the BLUEs are proportional to the number $N_m$ of measurement users of path $p_m$. Keeping this in mind, we propose weighted estimation as follows.

We define $z_m^*$ $(m = 1, 2, \ldots, M)$ as $z_m^* = \sum_{k=1}^{N_m} z_{m,k}$. It then follows from (3) and (6) that

$$\begin{aligned} z_m^* &= \sum_{k=1}^{N_m} \left( x_{i_m,k}^{(p_m)} + y_{j_m,k}^{(p_m)} \right) = N_m x_{i_m}^{(p_m)} + N_m y_{j_m}^{(p_m)} \\ &= N_m x_{i_m} + N_m y_{i_m} + N_m \nu_m. \end{aligned} \tag{14}$$

Let $\boldsymbol{N}^*$ denote an $M \times M$ diagonal matrix whose $m$-th $(m = 1, 2, \ldots, M)$ diagonal component is given by $N_m$. (14) is then rewritten to be

$$\begin{aligned} \boldsymbol{z}^* &= \boldsymbol{N}^* \boldsymbol{R}_{\mathrm{B}} \boldsymbol{x} + \boldsymbol{N}^* \boldsymbol{R}_{\mathrm{S}} \boldsymbol{y} + \boldsymbol{N}^* \boldsymbol{\nu} \\ &= \boldsymbol{R}_{\mathrm{B}}^* \boldsymbol{U}_{\mathrm{B}} \boldsymbol{f} + \boldsymbol{R}_{\mathrm{S}}^* \boldsymbol{y} + \boldsymbol{N}^* \boldsymbol{\nu}, \end{aligned}$$

where $\boldsymbol{z}^* = (z_1^*\ z_2^*\ \cdots\ z_M^*)^\top$ and

$$\boldsymbol{R}_{\mathrm{B}}^* = \boldsymbol{N}^* \boldsymbol{R}_{\mathrm{B}}, \qquad \boldsymbol{R}_{\mathrm{S}}^* = \boldsymbol{N}^* \boldsymbol{R}_{\mathrm{S}}.$$

As in unweighted estimation, $\boldsymbol{f}$ and $\boldsymbol{y}$ are estimated by solving the following $\ell_1$-$\ell_2$ optimization problem:

$$\min_{\boldsymbol{f}, \boldsymbol{y}} \frac{1}{2} \|\boldsymbol{z}^* - \boldsymbol{R}_{\mathrm{B}}^* \boldsymbol{U}_{\mathrm{B}} \boldsymbol{f} - \boldsymbol{R}_{\mathrm{S}}^* \boldsymbol{y}\|_2^2 + \eta_1^* \|\boldsymbol{f}\|_1 + \eta_2^* \|\boldsymbol{y}\|_1, \tag{15}$$

where $\eta_1^*$, $\eta_2^*$ denote positive parameters. Note that (15) is equivalent to (12) if $N_1 = N_2 = \cdots = N_M$.

## 5. Simulation Experiments

We validate the proposed scheme with unweighted/weighted estimation by conducting simulation experiments.

### 5.1 Experimental Setup

In all simulation experiments, we assume $N_{\mathrm{B}} = 100$ and

$N_S = 10$. Figure 4 shows base stations deployed on the grid in 3000 $[m] \times$ 3000 $[m]$ area. We set $D_{th} = 300$ $[m]$ to establish virtual links.

At base station $b_i \in \mathcal{V}_B$, delays are generated according to an exponential distribution with mean $\mu_{b_i}$. Delays at base stations are spatially dependent, as shown in Fig. 4 where base stations on darker areas have larger average delays. Delays at server $s_j \in \mathcal{V}_S$ are also generated according to an exponential distribution with mean $\mu_{s_j}$, where $\mu_{s_1} = \mu_{s_2} = 100$ [msec] and $\mu_{s_j} = 5$ [msec] for $s_j \in \mathcal{V}_S \setminus \{s_1, s_2\}$. We define $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ as *true BS delay vector* $\boldsymbol{x}_0 = (\mu_{b_1}\ \mu_{b_2}\ \cdots\ \mu_{b_{N_B}})^\top$ and *true server delay vector* $\boldsymbol{y}_0 = (\mu_{s_1}\ \mu_{s_2}\ \cdots\ \mu_{b_{N_S}})^\top$, respectively.

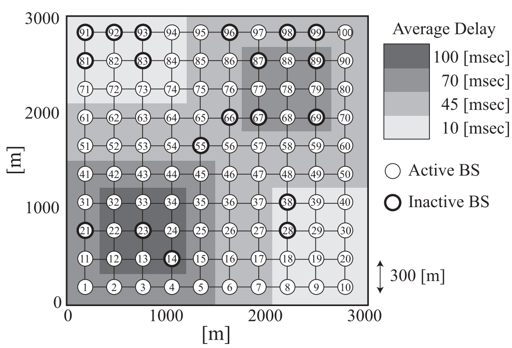
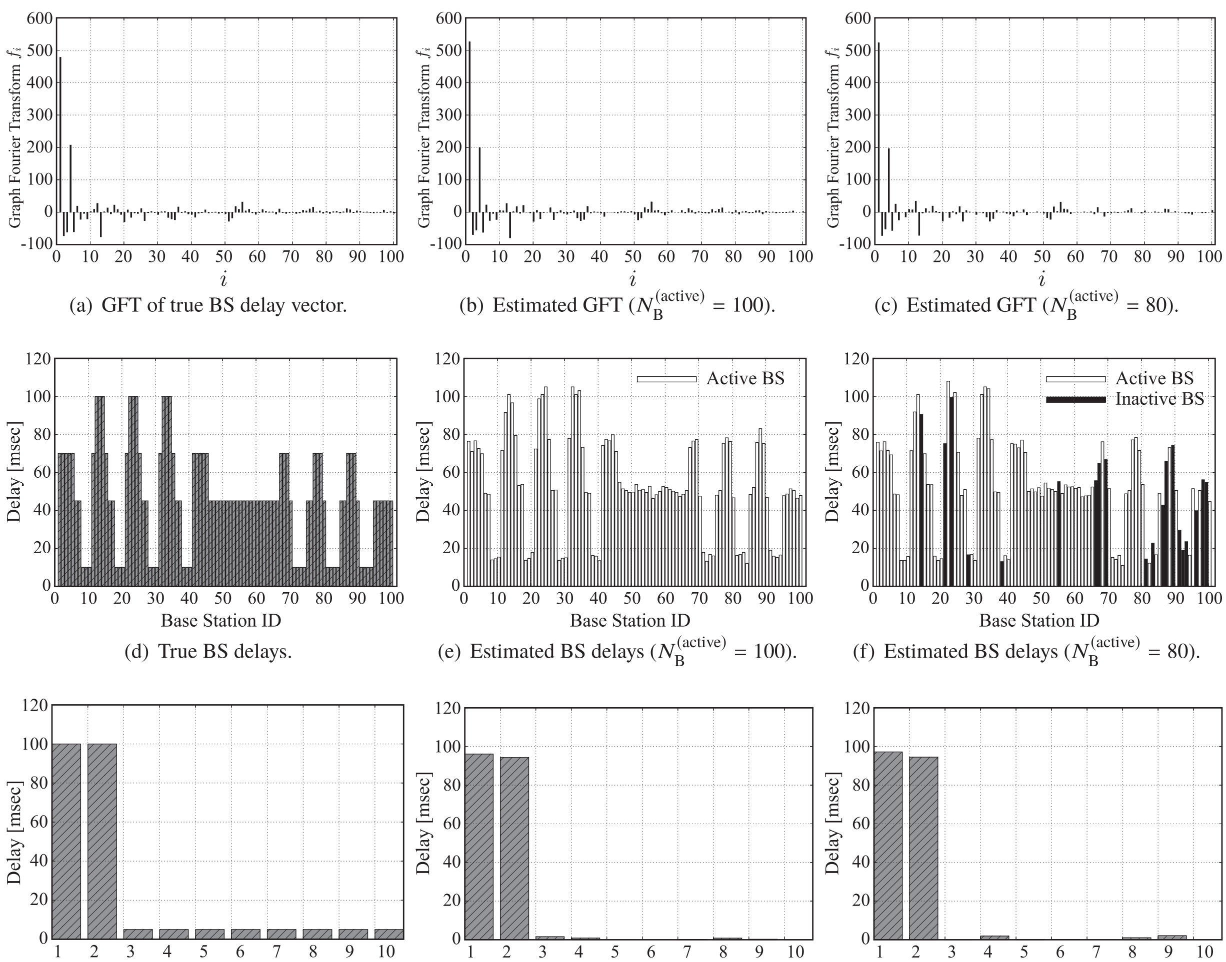



**Fig. 4** Graph structure of base stations and a spatially dependent delay distribution.

Parameters in the iterative algorithm are set to be: $\epsilon = 10^{-5}$, $\eta_1 = 5$, and $\eta_2 = 10$ in unweighted estimation, and $\epsilon = 10^{-5}\overline{N}$, $\eta_1^* = 5\overline{N}$ and $\eta_2^* = 10\overline{N}$ in weighted estimation, where $\overline{N} = (N_1 + N_2 + \cdots + N_M)/M$. Although the parameter optimization is important for the proposed scheme, we leave it as future work.

## 5.2 Interpolation Effect of GFT

It is known that GFT has the *interpolation effect* [27], i.e., missing values of signals on a graph are interpolated by smoothness of eigenvectors. In network monitoring techniques based on crowdsourcing, this effect is important because measurement users may not exist at some base stations. In this section, we evaluate the interpolation effect in




(a) GFT of true BS delay vector. (b) Estimated GFT ($N_B^{(\text{active})} = 100$). (c) Estimated GFT ($N_B^{(\text{active})} = 80$).

(d) True BS delays. (e) Estimated BS delays ($N_B^{(\text{active})} = 100$). (f) Estimated BS delays ($N_B^{(\text{active})} = 80$).

(g) True server delays. (h) Estimated server delays ($N_B^{(\text{active})} = 100$). (i) Estimated server delays ($N_B^{(\text{active})} = 80$).

**Fig. 5** GFTs of BS delays (upper), BS delays (middle), and server delays in the proposed scheme with unweighted estimation.

the proposed scheme, where unweighted estimation is employed. We consider two types of base stations: *active base stations* and *inactive base stations*, where an active base stations has some measurement users and an inactive base station has no measurement users. We choose $N_{\mathrm{B}}^{(\mathrm{active})}$ active base stations randomly and other $(N_{\mathrm{B}} - N_{\mathrm{B}}^{(\mathrm{active})})$ base stations are set to be inactive. Figure 4 shows active/inactive base stations for $N_{\mathrm{B}}^{(\mathrm{active})} = 80$. We assume that there are $N_{\mathrm{user}} = 500$ measurement users in each active base station. Each measurement user chooses one server from 10 servers randomly.

Figure 5 shows a typical result in the above scenario, where Fig. 5(a) shows the GFT $\boldsymbol{f}_0$ of the true BS delay vector $\boldsymbol{x}_0$, and Figs. 5(b) and 5(c) show estimated GFTs for $N_{\mathrm{B}}^{(\mathrm{active})} = 100$ and 80, respectively. We observe that estimated GFTs are sparsified as compared with that of the true BS delay vector. Figure 5(d) shows a true BS delay vector $\boldsymbol{x}_0$, and Figs. 5(e) and 5(f) show estimated BS delay vectors for $N_{\mathrm{B}}^{(\mathrm{active})} = 100$ and 80, respectively. From these figures, we observe that the proposed scheme can capture the spatial dependence of delays at base stations, even if some base stations have no measurement users. Moreover, Fig. 5(g) shows a true server delay vector $\boldsymbol{y}_0$, and Figs. 5(h) and 5(i) show estimated server delay vectors for $N_{\mathrm{B}}^{(\mathrm{active})} = 100$ and 80, respectively. These figures clearly indicate servers with larger delays.

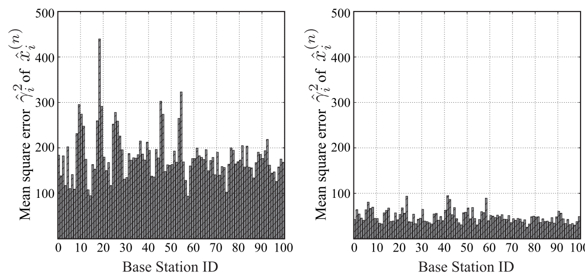
### 5.3 Performance of Weighted Estimation

We now demonstrate the robustness of weighted estimation against user heterogeneity. For this purpose, we assume $M = 2N_{\mathrm{B}}$, and for each base station $b_i$ $(i = 1, 2, \ldots, N_{\mathrm{B}})$, we set two paths $p_{2i-1} = (i, j_{i,1})$, $p_{2i} = (i, j_{i,2})$, where two different servers $s_{j_{i,1}}$ and $s_{j_{i,2}}$ are chosen randomly. To represent user heterogeneity, we set $N_{2i-1} = 50(1-\rho)$ and $N_{2i} = 50(1+\rho)$ $(0 \le \rho \le 1)$. Note that when $\rho = 0$, $N_m = 50$ for all $m = 1, 2, \ldots, M$, and the degree of user heterogeneity increases with $\rho$. We define $\hat{\gamma}_i^2$ $(i = 1, 2, \ldots, N_{\mathrm{B}})$ and $\hat{\gamma}^2$ as the mean square error (MSE) of estimated delay $\hat{x}_i^{(n)}$ at base station $b_i$ and mean of $\hat{\gamma}_i^2$.

$$\hat{\gamma}_i^2 = \frac{1}{N_{\mathrm{run}}} \sum_{n=1}^{N_{\mathrm{run}}} (\hat{x}_i^{(n)} - x_i)^2, \qquad \hat{\gamma}^2 = \frac{1}{N_{\mathrm{B}}} \sum_{n=1}^{N_{\mathrm{B}}} \hat{\gamma}_i^2,$$

where $x_i$ is given by (4) and $N_{\mathrm{run}} = 200$.

Figure 6 shows $\hat{\gamma}_i^2$ $(i = 1, 2, \ldots, N_{\mathrm{B}})$ for $\rho = 0.9$ (i.e., $N_{2i-1} = 5$ and $N_{2i} = 95$). From Figs. 6(a) and 6(b), we observe that unweighted estimation has much larger MSE than weighted estimation. Figure 7 shows $\hat{\gamma}^2$ of unweighted estimation and weighted estimation as a function of $\rho$. While $\hat{\gamma}^2$ in unweighted estimation increases rapidly with $\rho$, $\hat{\gamma}^2$ does not increase so much in weighted estimation even if $\rho$ approaches 1. Therefore, the proposed scheme with weighted estimation is more robust against user heterogeneity.




(a) unweighted scheme (b) weighted scheme

**Fig. 6** Mean square error $\hat{\gamma}_i^2$ of the estimated $\hat{x}_i^{(n)}$ $(\rho = 0.9)$.

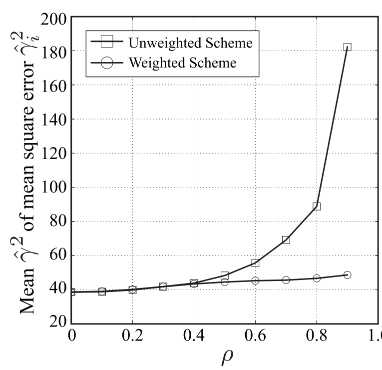



**Fig. 7** Mean $\hat{\gamma}^2$ of mean square error $\hat{\gamma}_i^2$ vs. $\rho$ $(0 \le \rho \le 1)$.

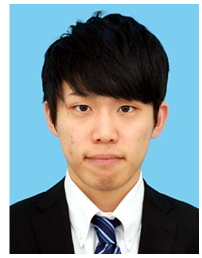
## 6. Conclusion

In this paper, we proposed heterogeneous delay tomography to estimate delays at base stations and servers in mobile networks by means of graph Fourier transform and compressed sensing. Simulation results validate the proposed scheme. We still have some remaining issues with the proposed scheme such as the parameter optimization of the proposed scheme and the performance evaluation in various network environments. Further, the graph construction is an interesting problem in GFT. While a binary adjacency matrix is used in this paper, it can be defined by a real matrix [16]. We will leave these issues to future research.

### References

[1] R. Cáceres, N.G. Duffield, J. Horowitz, and D.F. Towsley, "Multicast-based inference of network-internal loss characteristics," IEEE Trans. Inf. Theory, vol.45, no.7, pp.2462–2480, Nov. 1999.
[2] A. Chen, J. Cao, and T. Bu, "Network tomography: Identifiability and Fourier domain estimation," IEEE Trans. Signal Process., vol.58, no.12, pp.6029–6039, Dec. 2010.
[3] M. Coates, A. Hero, R. Network, and B. Yu, "Internet tomography," IEEE Signal Process. Mag., vol.19, no.3, pp.47–65, May 2002.
[4] M.H. Firooz and S. Roy, "Link delay estimation via expander graphs," IEEE Trans. Commun., vol.62, no.1, pp.170–181, Jan. 2014.
[5] F. Lo Presti, N.G. Duffield, J. Horowitz, and D. Towsley, "Multicast-based inference of network-internal delay distributions," IEEE/ACM Trans. Netw., vol.10, no.3, pp.761–775, Dec. 2002.
[6] T. Matsuda, M. Nagahara, and K. Hayashi, "Link quality classifier with compressed sensing based on $\ell_1$-$\ell_2$ optimization," IEEE Commun. Lett., vol.15, no.10, pp.1117–1119, Oct. 2011.

[7] T. Morita, T. Matsuda, and T. Takine, "Spatially dependent loss tomography for multihop wireless networks," Proc. 2016 International Conference on Information Networking (ICOIN 2016), pp.29–34, Jan. 2016.
[8] K. Nakanishi, S. Hara, T. Matsuda, K. Takizawa, F. Ono, and R. Miura, "Synchronization-free delay tomography based on compressed sensing," IEEE Commun. Lett., vol.18, no.8, pp.1343–1346, Feb. 2014.
[9] M.-F. Shih and A.O. Hero, III, "Unicast-based inference of network link delay distributions with finite mixture models," IEEE Trans. Signal Process., vol.51 no.8, pp.2219–2228, Aug. 2003.
[10] K. Takemoto, T. Matsuda, and T. Takine, "Sequential loss tomography using compressed sensing," IEICE Trans. Commun., vol.E96-B, no.11, pp.2756–2765, Nov. 2013.
[11] Y. Tsang, M. Coates, and R.D. Nowak, "Network delay tomography," IEEE Trans. Signal Process., vol.51, no.8, pp.2125–2136, Aug. 2003.
[12] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," IEEE Commun. Surveys Tuts., vol.18, no.1, pp.124–161, First Quarter, 2016.
[13] E.J. Candès and M.B. Wakin, "An introduction to compressed sampling," IEEE Signal Process. Mag., vol.25, no.2, pp.21–30, March 2008.
[14] K. Hayashi, M. Nagahara, and T. Tanaka, "A user's guide to compressed sensing for communications systems," IEICE Trans. Commun., vol.E96-B, no.3, pp.685–712, March 2013.
[15] A. Sandryhaila, and J.M.F. Moura, "Discrete signal processing on graphs: Frequency analysis," IEEE Trans. Signal Process., vol.62, no.12, pp.3042–3054, June 2014.
[16] D.I. Shuman, S.K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," IEEE Signal Process. Mag., vol.30, no.3, pp.83–98, May 2013.
[17] H. Kinsho, R. Tagyo, D. Ikegami, T. Matsuda, A. Takahashi, and T. Takine, "Heterogeneous delay tomography based on graph Fourier transform in mobile networks," Proc. 2017 IEEE ComSoc International Communications Quality and Reliability Workshop (CQR2017), May 2017.
[18] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and Internet traffic matrices (extended version)," IEEE/ACM Trans. Netw., vol.20, no.3, pp.662–676, June 2012.
[19] N.G. Duffield, J. Horowitz, F. Lo Presti, and D.F. Towsley, "Multicast topology inference from measured end-to-end loss," IEEE Trans. Inf. Theory, vol.48, no.1, pp.26–45, Jan. 2002.
[20] F.R.K. Chung, Spectral Graph Theory, American Mathematical Society, May 1997.
[21] M. Zibulevsky and M. Elad, "L1-L2 optimization in signal and image processing," IEEE Signal Process. Mag., vol.27, no.3, pp.76–88, May 2010.
[22] C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," IEEE Trans. Multimedia, vol.15, no.5, pp.1121–1137, Aug. 2013.
[23] M.S. Mushtaq, B. Augustin, and A. Mellouk, "Crowd-sourcing framework to assess QoE," Proc. 2014 IEEE International Conference on Communications (ICC 2014), pp.1705–1710, June 2014.
[24] R. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," IEEE Commun. Mag., vol.49, no.11, pp.32–39, Nov. 2011.
[25] A. Faggiani, E. Gregori, L. Lenzini, V. Luconi, and A. Vecchio, "Smartphone-based crowdsourcing for network monitoring: Opportunities, challenges, and a case study," IEEE Commun. Mag., vol.52, no.1, pp.106–113, Jan. 2014.
[26] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM J. Imaging Sci., vol.2, no.1, pp.183–202, April 2009.
[27] S.K. Narang, A. Gadde, and A. Ortega, "Signal processing techniques for interpolation in graph structured data," Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp.5445–5449, May 2013.



**Hideaki Kinsho** received B.E., M.E., degrees from Osaka University in Japan in 2016 and 2018, respectively. He was engaged in research on QoS estimation technologies in Osaka University. He is a Researcher in Communication Quality Group, Communication Traffic & Service Quality Project, NTT Network Technology Laboratories, Tokyo, Japan. He joined NTT in 2018 and has been engaged in research on traffic control and analysis. He is a member of IEICE.

**Rie Tagyo** received her B.S. and M.S. degrees in physics from Ochanomizu University in 2012 and 2014, respectively. Since joining NTT Laboratories in 2014, she has been engaged in the research on performance analysis for mobile communication networks at NTT Network Technology Laboratories. She received the Young Engineer Award of the IEICE in 2017.

**Daisuke Ikegami** received B.E., M.E. and Ph.D. degrees from Waseda University, Tokyo in 2002, 2004, and 2007, respectively. Since joining NTT Laboratories, he has been engaged in research on IP service network engineering including network planning of NGN, analysis of TCP performance and QoS estimation technologies. He received the Young Engineer Award of the IEICE in 2010.

**Takahiro Matsuda** is B.E. with honors, M.E., and Ph.D. in communications engineering from Osaka University in 1996, 1997, 1999, respectively. He joined the Department of Communications Engineering at the Graduate School of Engineering, Osaka University in 1999. In the same department, he was an Assistant Professor from 1999 to 2005, a Lecturer from 2005 to 2009, and an Associate Professor from 2009 to 2018. He is currently a Professor in the Department of Computer Science, Graduate School of Systems Design, Tokyo Metropolitan University. His research interests include performance analysis and the design of communication networks and wireless communications. He received Best Tutorial Paper Award and Best Magazine Paper Award from IEICE ComSoc in 2012, and Best Paper Award from IEICE in 2014. He is a member of IPSJ and IEEE.

**Jun Okamoto** received his B.E. and M.E. degrees in electrical engineering from the Science University of Tokyo in Japan in 1994 and 1996. He joined NTT Laboratories in 1996 and has been engaged in the quality assessment of multimedia telecommunication and network performance measurement methods. Currently, he is the Manager of the Communication Quality Group in NTT Laboratories. He received the telecommunications Advancement Foundation Award in 2009 and the International Telecommunication Union Encouragement Award in Japan in 2010.

**Tetsuya Takine** is currently a Professor in the Department of Information and Communications Technology, Graduate School of Engineering, Osaka University, Suita, Japan. He was born in Kyoto, Japan, on November 28, 1961, and received B.Eng., M.Eng., and Dr.Eng. degrees in applied mathematics and physics from Kyoto University, Kyoto, Japan, in 1984, 1986, and 1989, respectively. In April 1989, he joined the Department of Applied Mathematics and Physics, Faculty of Engineering, Kyoto University, as an Assistant Professor. Beginning in November 1991, he spent one year at the Department of Information and Computer Science, University of California, Irvine, on leave of absence from Kyoto University. In April 1994, he joined the Department of Information Systems Engineering, Faculty of Engineering, Osaka University as a Lecturer, and from December 1994 to March 1998, he was an Associate Professor in the same department. From April 1998 to May 2004, he was an Associate Professor in the Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University. His research interests include queueing theory, emphasizing numerical computation, and its application to performance analysis of computer and communication networks. He is now serving as an associate editor of Stochastic Models and International Transactions in Operational Research. He received Telecom System Technology Award from TAF in 2003 and 2010, Best Paper Awards from ORSJ in 1997, from IEICE in 2004, 2009, and 2014, and from ISCIE in 2006, and Best Tutorial Paper Award and Best Magazine Paper Award from IEICE ComSoc in 2012. Dr. Takine is a fellow of ORSJ and a member of IPSJ, ISCIE, and IEEE.