Detection of Time-varying Directional Flows in Wireless Networks

Jinsub Kim and Lang Tong
School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853
Email: {jk752, lt35}@cornell.edu

***Abstract*—The problem of detecting packet flows between two nodes in a wireless network is considered. Especially, the transmission timings of two nodes are recorded, and their transmission rates can be time-varying (piecewise constant). Based on the timing measurements, our objective is to detect the presence of packet flows between them.**

Two different scenarios are considered; the first is that a flow may exist in only one specific direction, and the other is that a flow may exist in any direction. For each case, a detection algorithm is provided, and for the latter scenario, an additional algorithm aimed at estimating the direction of the underlying flow is proposed. When the transmission processes are nonhomogeneous Poisson processes, under certain conditions, our algorithms are proved to be consistent. The algorithms are tested using the MSN Voice over IP (VoIP) traffic and the synthetic Poisson traffic.

I. Introduction

This paper considers detection of flows between two nodes having time-varying transmission rates. Fig. 1 illustrates the problem. In the wireless network, R_1 and R_2 may have time-varying transmission rates, and their transmission timings are recorded. We say that a packet flow exists from R_1 to R_2, if R_1 is sending packets to R_2, and R_2 is forwarding them to its neighbor. Based on the transmission timing measurements, our objective is to detect the presence of packet flows between R_1 and R_2. The timing measurements may correspond to different scenarios: They may represent independent transmissions of R_1 and R_2 with no packet flow. They may have epochs that belong to a packet flow from R_1 to R_2 or vice versa. Unsurprisingly, packet flows may exist in both directions.

Flow detection can find its application in various problems. As illustrated in Fig. 1, using simple monitors, one may infer about network routes and configuration. Another application is in detection of interactive stepping stone attack [1], in which a chain of compromised nodes (stepping stones) carry packets between an attacker and a victim. Flow detection can be employed to trace back the stepping stone chain, and eventually the attacker. Fig. 2 describes a specific application scenario where wireless transmission epochs of a wireless device (P_1) and an access point (A_2) are recorded. By detecting a packet flow from P_1 to A_2, one can check whether P_1 is communicating with any device in the area covered by A_2.

Work in this paper was sponsored by National Science Foundation under Contract CCF-0728872 and Army Research Office MURI Program under award W911NF-08-1-0238. The first author was partially supported by Samsung Scholarship.

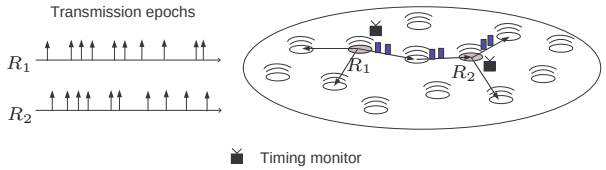

Fig. 1. Transmission timings of R_1 and R_2 are measured. In this example, a packet flow exists from R_1 to R_2. However, detecting its presence based on the timing measurements is nontrivial.

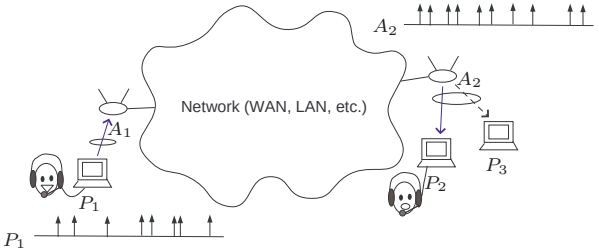

Fig. 2. A_1 and A_2 are access points connecting wireless devices to the network. If P_1 sends packets to P_2, a packet flow should exist from P_1 to A_2.

Transmission timings can be easily measured by simple monitoring devices. However, timing-based flow detection is certainly a nontrivial problem, partly because we do not assume any information from packet headers[1]: only the transmission timings are used. Another source of difficulty is the presence of noise-like epochs. Even when a packet flow exists from R_1 to R_2, R_1 and R_2 may have many transmissions that do not belong to the packet flow. They may multiplex transmissions of intersecting packet flows involving other nodes, or possibly superpose dummy transmissions to confuse detection systems. We refer to the epochs of such transmissions as *chaff epochs*.

Since we are entirely relying on timing measurements, we need to impose certain constraint on packet flows, so that transmission epochs of packet flows are distinguishable from independent transmissions. Hence, we assume a maximum end-to-end delay constraint Δ on flow packets. Such constraint can be found in latency-sensitive applications such as VoIP, video conference, etc.

We study flow detection under two different scenarios. The first scenario is that a packet flow can exist in only one

[1]In practice, header information may be available in many cases, and in such cases it should be exploited to enhance the detection performance. However, it is beyond the scope of this paper, and we assume that header information is not available due to encryption or some other technical issues.

specific direction, and the direction is priorly known. The second scenario is that packet flows may exist in either or both directions between two nodes. Without prior knowledge of flow direction, most problems fall into the second scenario. However, depending on applications, some problems may fit into the first scenario; the problem described in Fig. 2 is a good example. We refer to flow detection under the first scenario as detection of *unidirectional flow*, and flow detection under the second scenario as detection of *general flow*.

A. Related Work

Detection of unidirectional flow has been actively studied in the intrusion detection literature, especially in the field of stepping-stone detection [1]. To deal with encrypted traffic, many researchers assumed the absence of header information and entirely relied on timing measurements. Donoho *et al.* [2] were the first to employ the flow model with a maximum delay constraint. Their wavelet analysis was shown to be able to detect a flow if the chaff part is independent of the flow part and the sample size is large enough. Following their seminal work, numerous practical detectors were developed to detect flows with a maximum delay constraint (see references in [3]). The counting-based method of Blum *et al.* [4] was shown to be able to detect a flow in arbitrary chaff if the fraction of chaff is lower than certain level. Also, the matching-based detector of He and Tong [3] can deal with arbitrary chaff insertion. Under the Poisson traffic assumption, it was shown that there exists a threshold τ such that if fraction of chaff is less than τ, a flow is detectable; otherwise, a flow can be hidden by proper chaff insertion. Furthermore, in [5], their matching-based detector was proved to be able to detect a flow with any positive rate if the chaff parts are independent Poisson processes. Motivated by [3], Kim and Tong [5] proposed a matching-based algorithm for detection of a general flow.

To the best of our knowledge, most previous studies did not give enough attention to the traffic with time-varying rates; their detectors may fail if the traffic has time-varying rates and simultaneously contains a large amount of chaff transmissions. Even though Blum *et al.* [4] studied the nonhomogeneous Poisson traffic case, the algorithm was analyzed only for the non-chaff case, and even the insertion of independent chaff may cause the algorithm to fail.

Recently, Kim and Tong [6] proposed a detection scheme that is especially designed to deal with traffic with time-varying rates. In this paper, we further develop the idea in [6] to study a wider range of problems.

B. Summary of Contributions and Organization

In [6], Kim and Tong proposed Adaptive Flow Detector (AFD), an algorithm for detecting a general flow in traffic with time-varying rates. In this paper, we extended their idea to propose a more general detection scheme. Fig. 3 describes our detection scheme for time-varying traffic. The scheme consists of three parts: detection of unidirectional flow, detection of general flow, and estimation of flow direction. We present algorithms for detection of unidirectional flow and estimation of flow direction. For detection of general flow, we employ AFD in [6].

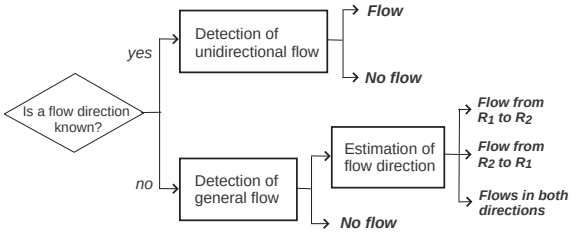

Fig. 3. If we know that a flow can exist in only one priorly known direction, a unidirectional flow detector is employed to detect a flow in that specific direction. Otherwise, a general flow detector first detects the presence of flows; and if a flow exists, we estimate its direction.

Under the nonhomogeneous Poisson traffic assumption and some additional conditions, our algorithms are proved to be consistent. We tested the algorithms using the MSN VoIP traffic and the synthetic nonhomogeneous Poisson traffic; the results are promising. Even though the algorithms are analyzed under the nonhomogeneous Poisson traffic assumption, the intuition behind the algorithms suggests that they may perform well on more general network traffic; the test results using the MSN VoIP traffic are supportive to this claim.

Due to the space limit, all the results are stated without proof. The reader may refer to the below for the details:
http://acsp.ece.cornell.edu/members/jinsub/allerton10.html

The rest of the paper is organized as follows. Section II first introduces mathematical formulation of flow detection problems. Section III considers detection of unidirectional flow, and Section IV considers detection of general flow and estimation of flow direction. In each section, algorithms are presented with consistency results. After that, Section V follows with supporting numerical results, and Section VI finally concludes the paper with remarks.

II. Mathematical Formulation

A. Notations and Definitions

We model the transmission epochs of each node by a point process on $[0, \infty)$. An upper case bold letter $\mathbf{S}$ denotes a point process, and $S(i)$ is the ith point of $\mathbf{S}$. A lower case bold letter $\mathbf{s}$ denotes a realization of a point process $\mathbf{S}$, and $s(i)$ is a realization of $S(i)$. In addition, $\mathcal{S}$ denotes a set of elements in a realization $\mathbf{s}$: $\mathcal{S} = \{s(i),\, i \geq 1\}$. We define a *superposition* operator $\bigoplus$ as follows: for a pair of increasing sequences $(a_i)_{i=1}^{\infty}$ and $(b_i)_{i=1}^{\infty}$, $(a_i)_{i=1}^{\infty} \oplus (b_i)_{i=1}^{\infty} \triangleq (c_i)_{i=1}^{\infty}$, where c_i is the ith smallest element among all the elements of two sequences[2]

We can formally define a *unidirectional flow* as below.

Definition 2.1: An ordered pair of processes $(\mathbf{F}_1, \mathbf{F}_2)$ is *a unidirectional flow*, if for every realization $(\mathbf{f}_1, \mathbf{f}_2)$, there exists a bijection $g : \mathcal{F}_1 \rightarrow \mathcal{F}_2$ satisfying $g(s) - s \in [0, \Delta], \;\; \forall s \in \mathcal{F}_1$.

[2]Order all the elements of $(a_i)_{i=1}^{\infty}$ and $(b_i)_{i=1}^{\infty}$ in an increasing order without removing any of them; some numbers might appear multiple times if they appeared multiple times in two sequences. Then, c_i is the ith element in the ordered sequence.

The bijection condition means packet conservation, and $g(s) - s \in [0, \Delta]$ means that each transmission satisfies causality and the maximum delay constraint. Based on the above definition, we define a *general flow* as below.

Definition 2.2: A pair of processes $(\mathbf{F}_1, \mathbf{F}_2)$ is *a general flow*, if $\mathbf{F}_i$ can be decomposed into $\mathbf{F}_i^{12}$ and $\mathbf{F}_i^{21}$ (*i.e.*, $\mathbf{F}_i = \mathbf{F}_i^{12} \oplus \mathbf{F}_i^{21}$, $i = 1, 2$), such that $(\mathbf{F}_1^{12}, \mathbf{F}_2^{12})$ and $(\mathbf{F}_2^{21}, \mathbf{F}_1^{21})$ are unidirectional flows.

A general flow, as the name stands, is more general in terms of directionality, compared to a unidirectional flow; it may consist of two unidirectional flows with opposite directions, or it may be just a unidirectional flow[3] in certain direction.

B. Problem Statement

Let $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the transmission processes of nodes R_1 and R_2, respectively. In each problem, we assume the same marginal distributions of $\mathbf{S}_1$ and $\mathbf{S}_2$ for all hypotheses; in other words, one cannot make a decision relying on the marginal distribution of $\mathbf{S}_i$.

1) Detection of unidirectional flow: Suppose it is priorly known that a packet flow may exist only from R_1 to R_2. Given the measurements $(\mathbf{s}_i)_{i=1}^2$ in $[0,\, t]$, we test the following hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mathbf{S}_1 \text{ and } \mathbf{S}_2 \text{ are independent} \\ \mathcal{H}_1: &\quad \mathbf{S}_i = \mathbf{F}_i \oplus \mathbf{W}_i,\ i = 1,\, 2. \\ &\quad (\mathbf{F}_1,\, \mathbf{F}_2) \text{ is a unidirectional flow} \end{aligned} \tag{1}$$

where under $\mathcal{H}_1$

- $\mathbf{F}_1$ and $\mathbf{F}_2$ have non-zero rates[4].
- $\mathbf{F}_1$ and $\mathbf{F}_2$ are not independent.
- $(\mathbf{F}_1, \mathbf{F}_2)$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are independent.

$\mathcal{H}_1$ represents the case that the traffic contains a unidirectional flow, and $\mathbf{W}_i$ denotes the chaff part of $\mathbf{S}_i$.

The listed assumptions are needed to make the problem well-posed; they are needed to guarantee that $\mathcal{H}_0$ and $\mathcal{H}_1$ are disjoint. For instance, without the third assumption, it can be shown that independent homogeneous Poisson processes $\mathbf{S}_1$ and $\mathbf{S}_2$ satisfy both $\mathcal{H}_0$ and $\mathcal{H}_1$ statements.

2) Detection of general flow: Suppose that a packet flow may exist in either or both directions between R_1 and R_2. Given the measurements $(\mathbf{s}_i)_{i=1}^2$ in $[0,\, t]$, we test the following hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mathbf{S}_1 \text{ and } \mathbf{S}_2 \text{ are independent} \\ \mathcal{H}_1: &\quad \mathbf{S}_i = \mathbf{F}_i \oplus \mathbf{W}_i,\ i = 1,\, 2. \\ &\quad (\mathbf{F}_1,\, \mathbf{F}_2) \text{ is a general flow} \end{aligned} \tag{2}$$

where under $\mathcal{H}_1$

- $\mathbf{F}_1$ and $\mathbf{F}_2$ have non-zero rates.
- $\mathbf{F}_1$ and $\mathbf{F}_2$ are not independent.
- $(\mathbf{F}_1, \mathbf{F}_2)$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are independent.

[3]Definition 2.2 implies that a unidirectional flow $(\mathbf{F}_1, \mathbf{F}_2)$ is also a general flow, because we can set $\mathbf{F}_1^{21}$ and $\mathbf{F}_2^{21}$ to be empty sequences

[4]A point process $\mathbf{F}$ is said to have non-zero rate if $\exists \delta$ s.t. $\liminf_{t\to\infty} \frac{N_{\mathbf{F}}([0,\, t])}{t} > \delta$, a.s., where $N_{\mathbf{F}}([0,\, t])$ is the number of points of $\mathbf{F}$ in $[0,\, t]$.

3) Estimation of flow direction: Suppose that $\mathbf{S}_1$ and $\mathbf{S}_2$ satisfy the $\mathcal{H}_1$ statement and the assumptions in the hypothesis testing (2). Given the measurements $(\mathbf{s}_i)_{i=1}^2$ in $[0,\, t]$, we test the following hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad \mathbf{S}_i = \mathbf{F}_i^{12} \oplus \mathbf{W}_i,\ i = 1,\, 2. \\ &\quad (\mathbf{F}_1^{12},\, \mathbf{F}_2^{12}) \text{ is a unidirectional flow} \\ \mathcal{H}_1: &\quad \mathbf{S}_i = \mathbf{F}_i^{21} \oplus \mathbf{W}_i,\ i = 1,\, 2. \\ &\quad (\mathbf{F}_2^{21},\, \mathbf{F}_1^{21}) \text{ is a unidirectional flow} \\ \mathcal{H}_2: &\quad \mathbf{S}_i = (\mathbf{F}_i^{12} \oplus \mathbf{F}_i^{21}) \oplus \mathbf{W}_i,\ i = 1,\, 2. \\ &\quad (\mathbf{F}_i^{12})_{i=1}^2,\ (\mathbf{F}_i^{21})_{i=2}^1 \text{ : unidirectional flows} \end{aligned} \tag{3}$$

where

- $\mathbf{F}_1^{12}$ and $\mathbf{F}_2^{21}$ have non-zero rates.
- $\mathbf{F}_1^{12}$ and $\mathbf{F}_2^{12}$ are not independent; $\mathbf{F}_2^{21}$ and $\mathbf{F}_1^{21}$ are not independent.
- $(\mathbf{F}_i^{12})_{i=1}^2$, $(\mathbf{F}_i^{21})_{i=2}^1$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are independent.

$\mathcal{H}_0$ and $\mathcal{H}_1$ represent the case that the traffic contains a unidirectional flow in only one direction. $\mathcal{H}_2$ represents the case that the traffic contains unidirectional flows in both directions.

III. Detection of Unidirectional Flow

In this section, we study detection of unidirectional flow, the binary hypothesis test (1) in Section II-B1. We first introduce the matching-based detector in [3] as a solution for the homogeneous Poisson traffic case. Then, we present our detection algorithm for the traffic with time-varying rates.

A. Relative Flow Rate

In timing-based detection, chaff epochs are analogous to noise in signal detection problems. Under $\mathcal{H}_1$, similar to signal-to-noise ratio, we use a metric called *relative flow rate* to measure the relative strength of the flow with respect to the whole traffic.

Definition 3.1: Suppose that a unidirectional flow $(\mathbf{F}_1,\, \mathbf{F}_2)$ is contained in $(\mathbf{S}_i)_{i=1}^2$. Let $(\mathbf{f}_i)_{i=1}^2$ and $(\mathbf{s}_i)_{i=1}^2$ denote the realizations of $(\mathbf{F}_i)_{i=1}^2$ and $(\mathbf{S}_i)_{i=1}^2$ respectively. Then, *relative flow rate* of $(\mathbf{f}_1,\, \mathbf{f}_2)$ is defined as

$$\begin{aligned} \mathrm{R_f}(t) &\triangleq \frac{\sum_{i=1}^{2} |\mathcal{F}_i \cap [0,\, t]|}{\sum_{i=1}^{2} |\mathcal{S}_i \cap [0,\, t]|}, \\ \mathrm{R_f} &\triangleq \liminf_{t\to\infty} \mathrm{R_f}(t) \end{aligned} \tag{4}$$

In other words, $\mathrm{R_f}(t)$ is the fraction of the unidirectional flow epochs in the measurements up to time t, and $\mathrm{R_f}$ is its limiting value as t increases to infinity.

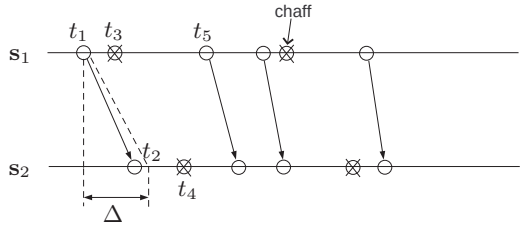

Fig. 4. Bounded-Greedy-Match [4]

TABLE I
BOUNDED-GREEDY-MATCH [4]

BGM($\mathbf{s}_1$, $\mathbf{s}_2$, Δ):

1: $m = n = 1$;
2: while $m \le |\mathcal{S}_1|$ and $n \le |\mathcal{S}_2|$
3: if $s_2(n) < s_1(m)$
4: $s_2(n)$ is chaff; $n \leftarrow n + 1$;
5: else if $s_2(n) > s_1(m) + \Delta$
6: $s_1(m)$ is chaff; $m \leftarrow m + 1$;
7: else
8: match $s_1(m)$ with $s_2(n)$;
9: $m \leftarrow m + 1$; $n \leftarrow n + 1$;
10: end
11: end
12: mark $s_1(i)$, $s_2(j)$ with $m \le i$, $n \le j$ as chaff;
13: $\bar{\mathrm{R}}_\mathrm{f} \leftarrow \frac{|\{\text{Matched epochs}\}|}{|\mathcal{S}_1| + |\mathcal{S}_2|}$;
14: return $\bar{\mathrm{R}}_\mathrm{f}$

B. Homogeneous Poisson Traffic: Detect-Bounded-Delay

This section introduces Detect-Bounded-Delay (DBD), a matching-based algorithm in [3].

DBD calculates an upper bound $\bar{\mathrm{R}}_\mathrm{f}(t)$ of $\mathrm{R}_\mathrm{f}(t)$, and compares it to a threshold τ to make a decision. Specifically, DBD takes the following form:

$$\begin{cases} \text{declare } \mathcal{H}_0 & \text{if } \bar{\mathrm{R}}_\mathrm{f}(t) < \tau \\ \text{declare } \mathcal{H}_1 & \text{otherwise} \end{cases} \tag{5}$$

where $\bar{\mathrm{R}}_\mathrm{f}(t)$ is defined as

$$\max_{\substack{\mathbf{f}_i, \mathbf{w}_i : \\ \mathbf{s}_i = \mathbf{f}_i \oplus \mathbf{w}_i \sim \mathcal{H}_1}} \frac{\sum_{i=1}^{2} |\mathcal{F}_i \cap [0,\, t]|}{\sum_{i=1}^{2} |(\mathcal{F}_i \cup \mathcal{W}_i) \cap [0,\, t]|}$$

where $\mathbf{s}_i = \mathbf{f}_i \oplus \mathbf{w}_i \sim \mathcal{H}_1$ denotes the constraint that $(\mathbf{f}_1, \mathbf{f}_2)$ is a realization[5] of a unidirectional flow.

To evaluate $\bar{\mathrm{R}}_\mathrm{f}(t)$, DBD employs Bounded-Greedy-Match (BGM), a matching algorithm by Blum *et al.* in [4]. BGM was shown to find a maximum number of matches between $\mathcal{S}_1$ and $\mathcal{S}_2$, that satisfy causality and the delay constraint. Given the measurements $(\mathbf{s}_i)_{i=1}^2$, BGM with Δ operates as follows.

1) Let l_1 be the earliest epoch in $\mathcal{S}_1$. Match l_1 with the earliest unmatched epoch in $[l_1,\, l_1 + \Delta]$ in $\mathcal{S}_2$.
2) Move to the second epoch l_2 in $\mathcal{S}_1$. Match l_2 with the earliest unmatched epoch in $[l_2,\, l_2 + \Delta]$ in $\mathcal{S}_2$. Repeat this step to find matches for all the epochs in $\mathcal{S}_1$.
3) After the trial to match the last epoch in $\mathcal{S}_1$, label all the unmatched epochs in $\mathcal{S}_1 \cup \mathcal{S}_2$ as chaff, and terminate.

Fig. 4 illustrates the operation of BGM. BGM first tries to find a match for t_1. Since t_2 is the earliest unmatched epoch in $[t_1,\, t_1 + \Delta] \cap \mathcal{S}_2$, t_1 is matched to t_2. Then, BGM searches for a match for t_3. However, t_2 is the only epoch in $[t_3,\, t_3 + \Delta] \cap \mathcal{S}_2$, and it is already matched with t_1. Hence, BGM leaves t_3 unmatched, and moves to t_5.

The implementation of BGM is given in Table I. Its computational complexity is linear with respect to the sample size.

The intuition behind DBD is that $\bar{\mathrm{R}}_\mathrm{f}(t)$ tends to be bigger when $(\mathbf{S}_i)_{i=1}^2$ contains a unidirectional flow[6], compared to when $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent. If under $\mathcal{H}_0$, as t grows, $\bar{\mathrm{R}}_\mathrm{f}(t)$ converges to or stay close to τ_0 with high probability, we can set the threshold τ to be slightly larger than τ_0 (*i.e.*, $\tau = \tau_0 + \epsilon$) and make the false alarm probability reasonably small for a large t; ϵ should be chosen carefully to balance the false alarm probability and the miss detection probability.

Finding such τ_0 is nontrivial, because it requires us to infer the distribution of $\mathbf{S}_i$. In [3], a closed-form expression for τ_0 is provided under the homogeneous Poisson traffic assumption. If $\mathbf{S}_1$ and $\mathbf{S}_2$ are homogeneous Poisson processes with rates λ_1 and λ_2, $\bar{\mathrm{R}}_\mathrm{f}(t)$ under $\mathcal{H}_0$ converges almost surely to a constant $\tau_0(\lambda_1, \lambda_2)$, which is a function of λ_1 and λ_2. Based on this result, Kim and Tong [5] proved that if under $\mathcal{H}_1$, $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent homogeneous Poisson processes, for any $\rho \in (0,\, 1)$ we can find a small positive ϵ such that DBD with a threshold $\tau_0(\lambda_1, \lambda_2) + \epsilon$ can consistently detect any unidirectional flow with $\mathrm{R}_\mathrm{f} \ge \rho$.

However, if the traffic is allowed to have time-varying rates, we face a great difficulty in setting a threshold τ. The analyses in [3] and [5] are entirely based on the homogeneous Poisson traffic assumption, in which one can infer the distribution of $\mathbf{S}_i$ based on its average rate $\frac{|\mathcal{S}_i \cap [0,\, t]|}{t}$. However, even a small deviation from the homogeneous Poisson traffic assumption would make the distribution inference extremely difficult.

As a simple example of traffic with time-varying rates, suppose that $\mathbf{S}_1$ and $\mathbf{S}_2$ are nonhomogeneous Poisson processes. Unlike estimating rates in homogeneous case, estimating local intensities of nonhomogeneous Poisson processes is nontrivial. Hence, it is difficult to infer the distribution of $\mathbf{S}_i$. This leads to the difficulty in estimating the behavior of $\bar{\mathrm{R}}_\mathrm{f}(t)$ under $\mathcal{H}_0$, and eventually the difficulty in setting a threshold τ.

To overcome such limitations of DBD, in the following section we introduce an algorithm to deal with the traffic with time-varying rates.

C. Unsupervised Nonparametric Flow Detector: Unidirectional Flow

This section presents a unidirectional flow detector called Unsupervised Nonparametric Flow Detector (UNFD). The name represents the characteristics of our problem. The problem is unsupervised in the sense that without any training data,

[5]In other words, $\exists g : \mathcal{F}_1 \to \mathcal{F}_2,$ s.t. $g(s) - s \in [0,\, \Delta],\ \forall s \in \mathcal{F}_1$.

[6]Note that under $\mathcal{H}_1$, the optimality of BGM guarantees that $\bar{\mathrm{R}}_\mathrm{f}(t)$ is at least greater than $\mathrm{R}_\mathrm{f}(t)$.

we should classify the type of relation between a pair of transmission processes. In addition, it is nonparametric because we do not assume any specific distribution on transmission processes.

UNFD is similar to DBD in that it compares $\bar{\mathrm{R}}_{\mathrm{f}}(t)$ to certain threshold for making a decision. However, in contrast to DBD, UNFD gives a specific method to set a threshold: it generates '$\mathcal{H}_0$-like' traffic, and run BGM on it to calculate a threshold. Specifically, UNFD has the following form:

$$\begin{cases} \text{declare } \mathcal{H}_0 & \text{if } \bar{\mathrm{R}}_{\mathrm{f}}(t) < \bar{\tau}(t) + \epsilon \\ \text{declare } \mathcal{H}_1 & \text{otherwise} \end{cases} \tag{6}$$

where $\bar{\tau}(t)$ is obtained by running BGM on the $\mathcal{H}_0$-like traffic, which will be explained soon.

Recall that if under $\mathcal{H}_0$, as t increases, $\bar{\mathrm{R}}_{\mathrm{f}}(t)$ converges to or stay near τ_0 with high probability, then we can set τ of DBD to be slightly bigger than τ_0. In UNFD, $\bar{\tau}(t)$ plays a role of an estimate of τ_0, and a small positive ϵ gives a slight gap between the threshold and $\bar{\tau}(t)$.

To obtain $\bar{\tau}(t)$, we first employ Independent Traffic Approximation (ITA) in [6] to generate the $\mathcal{H}_0$-like traffic. ITA is a heuristic to generate the $\mathcal{H}_0$-like traffic based on the measurements. It has two parameters: the synthesis window size W_S and the gap α ($\alpha \geq \Delta$) between subsequent synthesis windows[7]. Fig. 5 is describing the procedure. ITA relies on the intuition that if α is large enough, then $\mathbf{S}_1$ epochs in $A1$ and $\mathbf{S}_2$ epochs in $B1$ will tend to be uncorrelated, even when a unidirectional flow exists. Given the measurements $(\mathbf{s}_i)_{i=1}^n$ in $[0,\, t]$, ITA with $(W_S,\, \alpha)$ operates as follows [6]:

1) $(\bar{\mathbf{s}}_i)_{i=1}^2$ denotes the resulting data. Initially, $\bar{\mathbf{s}}_1$ and $\bar{\mathbf{s}}_2$ contain no epoch.
2) Take the epochs of $\mathbf{s}_1$ in $[0,\, W_S]$, and add them to $\bar{\mathbf{s}}_1$.
3) Take the epochs of $\mathbf{s}_2$ in $[W_S + \alpha,\, 2W_S + \alpha]$, subtract $W_S + \alpha$ from the epochs, and add them to $\bar{\mathbf{s}}_2$.
4) For $i = 1,\, 2,\, \ldots,\, \lfloor \frac{t}{2(W_S+\alpha)} \rfloor - 1$:
   a) Take the epochs of $\mathbf{s}_1$ in $[2i(W_S + \alpha),\, 2i(W_S + \alpha) + W_S]$, subtract $i(W_S + 2\alpha)$ from the epochs, and add them to $\bar{\mathbf{s}}_1$.
   b) Take the epochs of $\mathbf{s}_2$ in $[(2i+1)(W_S+\alpha),\, (2i+1)(W_S+\alpha)+W_S]$, subtract $i(W_S+2\alpha)+(W_S+\alpha)$ from the epochs, and add them to $\bar{\mathbf{s}}_2$.

The sample size of $(\bar{\mathbf{s}}_i)_{i=1}^2$ can be doubled by a heuristic referred to as ITAh in [6]. Fig. 6 illustrates its operation.

Given the measurements $(\mathbf{s}_i)_{i=1}^2$, UNFD with ϵ incorporates ITA and works as follows:

1) Run BGM on $(\mathbf{s}_i)_{i=1}^2$: $\bar{\mathrm{R}}_{\mathrm{f}}(t)$ denotes the return value.
2) Run ITA with (W_S, α) $(\alpha \geq \Delta)$ on $(\mathbf{s}_i)_{i=1}^2$ to generate $(\bar{\mathbf{s}}_i)_{i=1}^2$, and run BGM on $(\bar{\mathbf{s}}_i)_{i=1}^2$: $\bar{\tau}(t)$ denotes the return value.
3) If $\bar{\mathrm{R}}_{\mathrm{f}}(t) \geq \bar{\tau}(t) + \epsilon$, declare $\mathcal{H}_1$; otherwise, declare $\mathcal{H}_0$.

[7] W_S and α need to be properly adjusted for different traffic characteristics. Unless some information is available for parameter setting, we recommend to set $\alpha = \Delta$ and $W_S \in [10\alpha,\, 20\alpha]$.

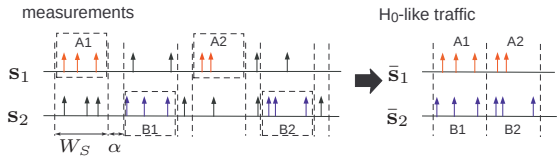

Fig. 5. Independent Traffic Approximation [6]: The W_S-second intervals $A1$, $A2$, $B1$, and $B2$ are cut from the measurements and assembled to generate the $\mathcal{H}_0$-like traffic.

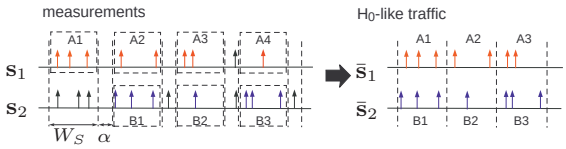

Fig. 6. ITAh [6]: Unlike ITA, ITAh does not throw away $A2$, $A4, \ldots$ and $B2$, $B4, \ldots$. Here, $A1$, $A2$, $A3, \ldots$ and $B1$, $B2$, $B3, \ldots$ are cut from the measurements and assembled to generate $\mathcal{H}_0$-like traffic.

The computational complexity of UNFD is linear with respect to the sample size, because BGM and ITA, its main components, have linear complexity.

Since $\bar{\mathbf{S}}_i$ is obtained by sequentially sampling subintervals of $\mathbf{S}_i$ and assembling them together, it is expected to retain several traffic characteristics of $\mathbf{S}_i$ (*e.g.*, the trend in rate changes, interarrival distribution, etc.) in some degree. Furthermore, $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are approximately uncorrelated. Therefore, depending on the distribution of $(\mathbf{S}_i)_{i=1}^2$, the return value of BGM on $(\bar{\mathbf{s}}_i)_{i=1}^2$, which is $\bar{\tau}(t)$, may well approxmiate $\bar{\mathrm{R}}_{\mathrm{f}}(t)$ under $\mathcal{H}_0$. Hence, we suggest that UNFD would perform well on quite general network traffic, not restricted to nonhomogeneous Poisson traffic assumed in our consistency result.

Under the nonhomogeneous Poisson traffic assumption and some additional conditions, the below theorem states that a unidirectional flow with any positive rate can be consistently detected by UNFD with proper ϵ.

Theorem 3.1: Suppose that $\mathbf{S}_1$ and $\mathbf{S}_2$ are nonhomogeneous Poisson processes. For any $\omega \in (0, 1)$, there exists an ϵ such that UNFD with ϵ can consistently detect[8] any unidirectional flow with $\mathrm{R}_{\mathrm{f}} \geq \omega$, if the following assumptions hold:

- Under $\mathcal{H}_1$, $\mathbf{F}_1$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are nonhomogeneous Poisson processes, and $\mathbf{F}_2 = \mathrm{sort}\{\mathbf{F}_1(i) + \alpha_i,\, i \geq 1\}$ where $\alpha_i \in [0, \Delta]$ a.s. and $\{\alpha_i\} \perp\!\!\!\perp \mathbf{F}_1,\, \mathbf{W}_1$.
- Let $\vec{\lambda}(t) = (\lambda_1(t),\, \lambda_2(t),\, \lambda_f(t))$ denote the local intensities[9] of $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{F}_1$. λ_1, λ_2, and λ_f are piecewise constant, and $\vec{\lambda}(t)$ can take values in the finite set $\Lambda \triangleq \{\vec{\lambda}^{(k)} = (\lambda_1^{(k)},\, \lambda_2^{(k)},\, \lambda_f^{(k)}),\ 1 \leq k \leq M\}$. $\vec{\lambda}(t)$ and Λ are deterministic and unknown.
- Let $\rho_k(t)$ $(1 \leq k \leq M)$ denote the fraction of time in $[0,\, t]$ that $\vec{\lambda}(t) = \vec{\lambda}^{(k)}$. As t increases, $\rho_k(t)$ converges to a constant ρ_k, $1 \leq k \leq M$.
- Let $c(t)$ denote the number of times that $\vec{\lambda}(t)$ changes in $[0,\, t]$. As t increases, $\frac{c(t)}{t}$ converges to 0.

[8] A detector is said to be *consistent*, if its false alarm probability and miss detection probability decreases to 0 as the sample size grows.

[9] If the measurements come from $\mathcal{H}_0$, then $\lambda_f(t) = 0,\, \forall t$.

IV. Detection of General Flow

In this section, we study detection of general flow and estimation of flow direction, which are mathematically formulated in Section II-B2 and Section II-B3.

A. Unsupervised Nonparametric Flow Detector: General Flow

This section presents a modification of UNFD for general flow detection[10]. We denote this detector by UNFD-G, where 'G' stands for general flow.

In this section, Definition III-A of *relative flow rate* will be extended to represent the relative strength of a general flow; $\mathrm{R_f}(t)$ denotes the fraction of the general flow epochs in the measurements up to time t, and $\mathrm{R_f}$ is $\liminf_{t\to\infty} \mathrm{R_f}(t)$.

For UNFD-G, we replace every BGM in the steps of UNFD with Bidirectional-Bounded-Greedy-Match (BiBGM) [5], the bidirectional version of BGM. Similar to BGM, under $\mathcal{H}_1$, BiBGM gives the below upper bound $\bar{\mathrm{R}}_\mathrm{f}(t)$ of $\mathrm{R_f}(t)$,

$$\max_{\substack{\mathbf{f}_i,\,\mathbf{w}_i:\\ \mathbf{s}_i=\mathbf{f}_i\oplus\mathbf{w}_i\sim\mathcal{H}_1}} \frac{\sum_{i=1}^{2}|\mathcal{F}_i\cap[0,\,t]|}{\sum_{i=1}^{2}|(\mathcal{F}_i\cup\mathcal{W}_i)\cap[0,\,t]|}$$

where $\mathbf{s}_i=\mathbf{f}_i\oplus\mathbf{w}_i\sim\mathcal{H}_1$ denotes the constraint that $(\mathbf{f}_1,\,\mathbf{f}_2)$ is a realization[11] of a general flow.

Given the measurements $(\mathbf{s}_i)_{i=1}^2$, BiBGM with Δ works as follows [5]:

1) Let s be the earliest epoch in $\mathcal{S}_1\cup\mathcal{S}_2$. Match s with the first unmatched epoch in $[s,\,s+\Delta]$ in the other node.
2) Move to the next unmatched epoch t in $\mathcal{S}_1\cup\mathcal{S}_2$. Match t with the first unmatched epoch in $[t,\,t+\Delta]$ in the other node. Keep moving to the next unmatched epoch in $\mathcal{S}_1\cup\mathcal{S}_2$ and finding its match based on the same rule.
3) After the trial to match the last unmatched epoch, label all the unmatched epochs as chaff and terminate.

For the detailed description of BiBGM, see [5]. Given the measurements $(\mathbf{s}_i)_{i=1}^2$, UNFD-G with ϵ works as follow [6]:

1) Run BiBGM with Δ on $(\mathbf{s}_i)_{i=1}^2$: $\bar{\mathrm{R}}_\mathrm{f}(t)$ denotes the return value.
2) Run ITA with (W_S,α) $(\alpha\geq\Delta)$ on $(\mathbf{s}_i)_{i=1}^2$ to generate $(\bar{\mathbf{s}}_i)_{i=1}^2$, and run BiBGM on $(\bar{\mathbf{s}}_i)_{i=1}^2$: $\bar{\tau}(t)$ denotes the return value.
3) If $\bar{\mathrm{R}}_\mathrm{f}(t)\geq\bar{\tau}(t)+\epsilon$, declare $\mathcal{H}_1$; otherwise, declare $\mathcal{H}_0$.

Under the nonhomogeneous Poisson traffic assumption and some additional conditions, the following theorem implies that a general flow with any positive rate can be consistently detected by UNFD-G with proper ϵ. This is a stronger result than the consistency result given in Theorem 3.1 in [6].

Theorem 4.1: Suppose that $\mathbf{S}_1$ and $\mathbf{S}_2$ are nonhomogeneous Poisson processes. For any $\omega\in(0,1)$, there exists an ϵ such that UNFD-G with ϵ can consistently detect any general flow with $\mathrm{R_f}\geq\omega$, if the following assumptions hold:

- Under $\mathcal{H}_1$, $\mathbf{S}_i=(\mathbf{F}_i^{12}\oplus\mathbf{F}_i^{21})\oplus\mathbf{W}_i$. $\mathbf{F}_1^{12}$, $\mathbf{F}_2^{21}$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are independent nonhomogeneous Poisson processes[12], $\mathbf{F}_2^{12}=\mathrm{sort}\{\mathbf{F}_1^{12}(i)+\alpha_i,\,i\geq1\}$ where $\alpha_i\in[0,\Delta]$ a.s., and $\mathbf{F}_1^{21}=\mathrm{sort}\{\mathbf{F}_2^{21}(i)+\beta_i,\,i\geq1\}$ where $\beta_i\in[0,\Delta]$ a.s..
  $\{\alpha_i\}\perp\!\!\!\perp\mathbf{W}_1$, $\{\beta_i\}\perp\!\!\!\perp\mathbf{W}_2$, and $\perp\!\!\!\perp\{\alpha_i\}$, $\{\beta_i\}$, $\mathbf{F}_1^{12}$, $\mathbf{F}_2^{21}$.
- Let $\vec{\lambda}(t)=(\lambda_1(t),\,\lambda_2(t),\,\lambda_{f1}(t),\,\lambda_{f2}(t))$ denote the local intensities[13] of $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{F}_1^{12}$, and $\mathbf{F}_2^{21}$. λ_1, λ_2, λ_{f1}, and λ_{f2} are piecewise constant, and $\vec{\lambda}(t)$ can take values in the finite set $\Lambda\triangleq\{\vec{\lambda}^{(k)}=(\lambda_1^{(k)},\,\lambda_2^{(k)},\,\lambda_{f1}^{(k)},\,\lambda_{f2}^{(k)}),\ 1\leq k\leq M\}$. $\vec{\lambda}(t)$ and Λ are deterministic and unknown.
- Let $\rho_k(t)$ $(1\leq k\leq M)$ denote the fraction of time in $[0,\,t]$ that $\vec{\lambda}(t)=\vec{\lambda}^{(k)}$. As t increases, $\rho_k(t)$ converges to a constant ρ_k, $1\leq k\leq M$.
- Let $c(t)$ denote the number of times that $\vec{\lambda}(t)$ changes in $[0,\,t]$. As t increases, $\frac{c(t)}{t}$ converges to 0.

B. Estimation of Flow Direction

This section considers estimation of flow direction, which is mathematically formulated in Section II-B3. The measurements are assumed to contain a general flow, and our objective is to estimate the direction of the underlying flow.

We propose an algorithm called Unsupervised Nonparametric Direction Estimator (UNDE). UNDE tests whether a unidirectional flow exists in a specific direction. In other words, UNDE tests $\{\mathcal{H}_0,\,\mathcal{H}_2\}$ versus $\mathcal{H}_1$ for a unidirectional flow from R_1 to R_2, and $\{\mathcal{H}_1,\,\mathcal{H}_2\}$ versus $\mathcal{H}_0$ for a unidirectional flow from R_2 to R_1. The specific steps to estimate directions are as follows:

1) Run UNDE to test $\{\mathcal{H}_0,\,\mathcal{H}_2\}$ versus $\mathcal{H}_1$: if the decision is $\mathcal{H}_1$, declare $\mathcal{H}_1$ and terminate; otherwise, go to 2.
2) Run UNDE to test $\{\mathcal{H}_1,\,\mathcal{H}_2\}$ versus $\mathcal{H}_0$: if the decision is $\mathcal{H}_0$, declare $\mathcal{H}_0$; otherwise, declare $\mathcal{H}_2$.

Suppose we want to test $\{\mathcal{H}_0,\,\mathcal{H}_2\}$ versus $\mathcal{H}_1$. Given the measurements $(\mathbf{s}_i)_{i=1}^2$ in $[0,\,t]$, UNDE with ϵ executes the following steps:

1) Increase all the epochs in $\mathbf{s}_1$ by $\frac{\Delta}{2}$.
2) Divide the observation interval into $\frac{\Delta}{2}$-second subintervals. Let A_i denote the ith subinterval, $(\frac{\Delta}{2}(i-1),\,\frac{\Delta}{2}i]$.
3) Assemble A_is to generate four data: for $k=1:1:4$, assemble $\{A_{4n+k},\,n\geq0\}$ to generate $(\mathbf{s}_i^{(k)})_{i=1}^2$. (Refer to Fig. 7)
4) For $k=1:1:4$, run ITA with (W_S,α) $(\alpha\geq\frac{\Delta}{2})$ on $(\mathbf{s}_i^{(k)})_{i=1}^2$ to generate $(\bar{\mathbf{s}}_i^{(k)})_{i=1}^2$.

[10] We already presented this algorithm in [6] as Adaptive Flow Detector. To better describe its connection with UNFD, we will illustrate this algorithm as a modification of UNFD.

[11] In other words, $\mathbf{f}_i$ can be partitioned into $\mathbf{f}_i^{12}$ and $\mathbf{f}_i^{21}$ such that there exist bijections $g_1:\mathcal{F}_1^{12}\to\mathcal{F}_2^{12}$ and $g_2:\mathcal{F}_2^{21}\to\mathcal{F}_1^{21}$ satisfying $g_1(s)-s\in[0,\,\Delta],\ \forall s\in\mathcal{F}_1^{12}$ and $g_2(s)-s\in[0,\,\Delta],\ \forall s\in\mathcal{F}_2^{21}$.

[12] Either $\mathbf{F}_1^{12}$ or $\mathbf{F}_2^{21}$ may have zero rate if the underlying general flow is a unidirectional flow.

[13] If the measurements come from $\mathcal{H}_0$, then $\lambda_{f1}(t)=\lambda_{f2}(t)=0,\ \forall t$.

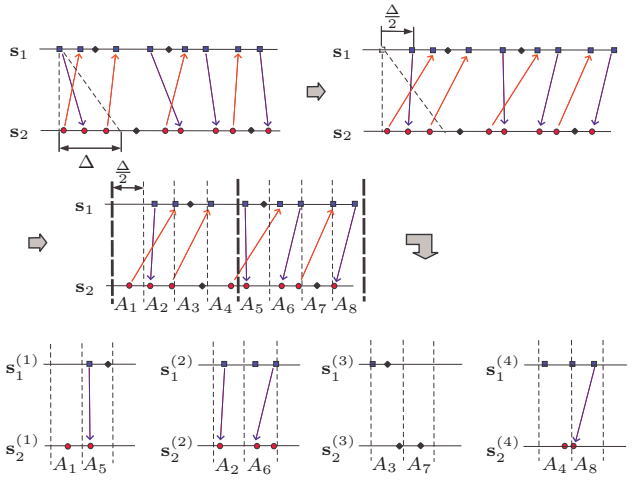

Fig. 7. Steps of Unsupervised Nonparametric Direction Estimator: Some of the matches (blue arrows pointing downward) corresponding to the packet flow from R_1 to R_2 survives in $(\mathbf{s}_i^{(k)})_{i=1}^2$, $k = 1, \ldots, 4$, but the matches (red arrows pointing upward) corresponding to the packet flow from R_2 to R_1 are all removed.

5) For $k = 1:1:4$, run BiBGM with $\frac{\Delta}{2}$ on $(\mathbf{s}_i^{(k)})_{i=1}^2$ and $(\bar{\mathbf{s}}_i^{(k)})_{i=1}^2$; $\bar{\mathrm{R}}_{\mathrm{f}}^{(k)}(t)$ and $\bar{\tau}^{(k)}(t)$ denote the return values.
6) If $\frac{\sum_{k=1}^4 \bar{\mathrm{R}}_{\mathrm{f}}^{(k)}(t)}{4} < \frac{\sum_{k=1}^4 \bar{\tau}^{(k)}(t)}{4} + \epsilon$, declare $\mathcal{H}_1$; otherwise, declare $\{\mathcal{H}_0, \mathcal{H}_2\}$.

Fig. 7 illustrates the first three steps of UNDE, which are aimed at removing the packet flow from R_2 to R_1 if it exists. Suppose that $s_2(i_2) - s_1(i_1) \in [0, \Delta]$ and $s_1(j_1) - s_2(j_2) \in [0, \Delta]$. Then, once the epochs in $\mathbf{s}_1$ are all increased by $\frac{\Delta}{2}$, $s_2(i_2) - s_1(i_1) \in [-\frac{\Delta}{2}, \frac{\Delta}{2}]$ and $s_1(j_1) - s_2(j_2) \in [\frac{\Delta}{2}, \frac{3\Delta}{2}]$. Hence, after the chopping, $s_1(i_1)$ and $s_2(i_2)$ may belong to the same A_i, but $s_1(j_1)$ and $s_2(j_2)$ will belong to different intervals A_i and A_j with $|i-j| \in \{1, 2, 3\}$. This implies that, after assembling, the matches corresponding to the packet flow from R_1 to R_2 may survive in $(\mathbf{s}_i^{(k)})_{i=1}^2$, $k = 1, \ldots, 4$, with some probability, but the matches corresponding to the packet flow from R_2 to R_1 are completely removed. In addition, if a match (a, b) corresponding to the packet flow from R_1 to R_2 survives in $(\mathbf{s}_i^{(k)})_{i=1}^2$ for some k, it satisfies $|a - b| < \frac{\Delta}{2}$. Therefore, we can detect the unidirectional flow from R_1 to R_2 by detecting the general flow with delay constraint $\frac{\Delta}{2}$ in $(\mathbf{s}_i^{(k)})_{i=1}^2$, $k = 1, \ldots, 4$; this is the objective of the step 4, 5, and 6.

Under the nonhomogeneous Poisson traffic assumption and some additional conditions, the following consistency result implies that if a unidirectional flow with any positive rate exists in a specific direction, then UNDE with proper ϵ on that direction can consistently detect its presence.

Theorem 4.2: Suppose that $\mathbf{S}_1$ and $\mathbf{S}_2$ are nonhomogeneous Poisson processes. For any $\omega \in (0, 1)$, there exists an ϵ such that UNDE with ϵ on a specific direction can consistently detect the presence of a unidirectional flow with $\mathrm{R_f} \geq \omega$ in that direction, if the following assumptions hold:

- All the assumptions listed in Theorem 4.1 hold.
- $\{\alpha_i\}$ is an i.i.d. sequence; $\{\beta_i\}$ is an i.i.d. sequence.
- $\Pr(\alpha_1 \in [\frac{\Delta}{2} - x, \Delta - x])$ and $\Pr(\beta_1 \in [\frac{\Delta}{2} - x, \Delta - x])$ are constant[14] for $x \in [0, \frac{\Delta}{2}]$.

V. Numerical Results

A. Simulation Results: Synthetic Poisson Traffic

We first tested the performance of UNFD using the synthetic nonhomogeneous Poisson traffic. $\mathbf{S}_1$ and $\mathbf{S}_2$ are set to be nonhomogeneous Poisson processes with piecewise constant rates $\lambda_1(t)$ and $\lambda_2(t)$. Under $\mathcal{H}_1$, we generated a unidirectional flow $(\mathbf{F}_1, \mathbf{F}_2)$ from R_1 to R_2. $\mathbf{F}_1$ is a nonhomogeneous Poisson process with a piecewise constant rate $\lambda_f(t)$. $\mathbf{F}_2$ is generated by adding i.i.d. random delays to the epochs of $\mathbf{F}_1$; delays are uniformly distributed in $[0, \Delta]$. $\mathbf{W}_1$ and $\mathbf{W}_2$ are also generated as independent nonhomogeneous Poisson processes, and they are independent of the flow part. In each measurements, $(\lambda_1(t), \lambda_2(t), \lambda_f(t))$ is piecewise constant, and it changes twice; it takes different values for the first third, the second third, and the last third of the measurements. In each measurements, $(\lambda_1(t), \lambda_2(t), \lambda_f(t))$ takes one of the four different rate changes[15] with equal probability. For $\mathcal{H}_0$, $\mathbf{S}_1$ and $\mathbf{S}_2$ are generated as independent nonhomogeneous Poisson processes whose rates change in the same manner as in $\mathcal{H}_1$.

Fig. 8 shows the ROC curves of UNFD. ROC curves are obtained by increasing ϵ of UNFD from 0 to 1 by 0.01, and plotting the false alarm probability (x-axis) and the detection probability (y-axis) of each case. The curves with circles are ROCs for the case that UNFD employs ITAh, and the curves with rectangles are ROCs when UNFD employs ITA. It is evident from the plot that ITAh, a heuristic to double the sample size of $(\bar{\mathbf{s}}_i)_{i=1}^2$, gives better ROCs. For comparison, UNFD-G was also tested on the same data assuming no prior information about flow direction is available. The curves with plus signs are ROCs of UNFD-G with ITAh. It requires much larger sample size to obtain a similar ROC curve with UNFD, due to the lack of direction information. In all cases, as the sample size grows, ROC curves move closer to the upper left corner implying better detection performance. In all other numerical results to be presented, the algorithms employed ITAh instead of ITA, unless otherwise specified.

For testing UNDE, $\mathbf{S}_1$ and $\mathbf{S}_2$ are generated to satisfy one of the three hypotheses ($\mathcal{H}_0$, $\mathcal{H}_1$, and $\mathcal{H}_2$ in (3)) with equal probability. In every hypothesis, $(\lambda_1(t), \lambda_2(t))$ is $(8, 8)$ for the first half of the measurements, and $(10, 10)$ for the second half of the measurements. In each hypothesis, a unidirectional flow $((\mathbf{F}_1^{12}, \mathbf{F}_2^{12})$ or $(\mathbf{F}_2^{21}, \mathbf{F}_2^{21}))$ with the constant rate λ_f is generated in the same way as in the simulation for UNFD. We ran UNDE on both directions to make a decision. Fig. 9 shows the plots of the error probability versus the sample size[16] for different λ_fs. As expected, the bigger λ_f results in a smaller

[14] A uniform distribution on $[0, \Delta]$ satisfies this condition.
[15] 1. $(10, 10, 6) \to (10, 20, 6) \to (20, 20, 6)$.
2. $(6, 10, 5) \to (10, 10, 5) \to (12, 12, 7)$.
3. $(25, 20, 15) \to (20, 20, 15) \to (20, 16, 15)$.
4. $(15, 10, 5) \to (25, 20, 15) \to (25, 25, 15)$.
[16] An error occurs if our decision is different from the true hypothesis.

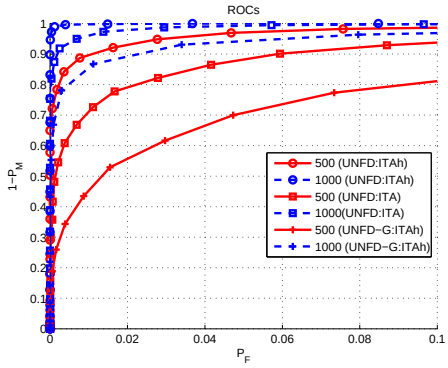

Fig. 8. ROC curves of UNFD and UNFD-G for different sample sizes. $W_S = 2$, $\alpha = \Delta = 0.1$, 10000 Monte Carlo runs.

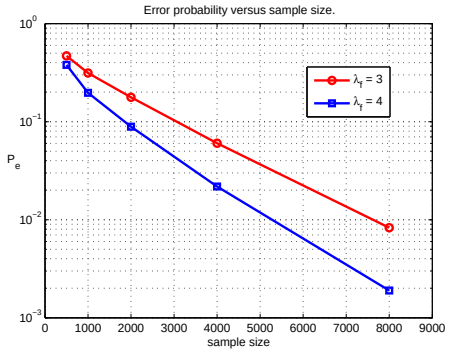

Fig. 9. Error probability of UNDE. $W_S = 2$, $\alpha = \Delta = 0.1$, 10000 Monte Carlo runs.

error probability. The error probability displays exponential decay as the sample size grows.

B. Experimental Results: MSN VoIP Traffic

UNFD was tested using the real-world MSN VoIP traffic, which is a representative example of traffic with an end-to-end delay constraint. As illustrated in Fig. 2, we located one laptop (P_1) in a room covered by the access point A_1, and two laptops (P_2 and P_3) in a different room covered by A_2. We measure[17] the transmission epochs of P_1 ($\mathbf{s}_1$) and those of A_2 ($\mathbf{s}_2$), and our objective is to detect the VoIP call between P_1 and any laptop in the area covered by A_2. Under $\mathcal{H}_1$, P_1 has an MSN VoIP call with P_2, and P_3 downloads a file from an FTP server with 20kB/s rate limit. Under $\mathcal{H}_0$, P_1 and P_2 have independent MSN VoIP calls, and P_3 does the same job as in $\mathcal{H}_1$. Hence, $\mathbf{s}_1$ consists of transmission epochs of VoIP packets and control/management packets, and $\mathbf{s}_2$ consists of transmission epochs of VoIP packets for P_2, FTP packets for P_3, and control/management packets (except beacon packets). $\mathbf{s}_1$ and $\mathbf{s}_2$ displayed time-varying rates. We evaluated the average rates of $\mathbf{s}_1 \oplus \mathbf{s}_2$ in consecutive 1000-second intervals, and the average rate dynamically fluctuated between 46 and 71 epochs/sec.

Table II shows the false alarm probability and the miss detection probability of UNFD for different sample sizes.

[17]Window Live Messenger 2009 (14.0.8089.726) was used for MSN VoIP calls, and Wireshark network protocol analyzer (ver 1.2.6.) with the AirPcap classic adaptor was used to collect the timings of wireless transmissions.

TABLE II
UNFD ON MSN VOIP TRAFFIC: $W_S = 2$, $\alpha = \Delta = 0.15$, $\epsilon = 0.05$. NUMBER OF EXPERIMENTS: 160, 80, AND 40 FOR SAMPLE SIZE 5000, 10000, AND 20000, RESPECTIVELY. TOTAL TRAFFIC RATES: $\lambda_1 = 26.80$, $\lambda_2 = 34.93$. FTP DATA RATE: 11.11.

	UNFD		DBD	
sample size	P_F	P_M	P_F	P_M
5000	0.0750	0.0750	0	0.9875
10000	0	0.0625	0	1
20000	0	0.0250	0	1

UNFD resulted in reasonably small error probabilities. For Δ in UNFD, we used 150ms, which is the upper bound of acceptable end-to-end delays of VoIP packets, recommended by ITU-T recommendation G.114 [7]. For comparison, we also tested Detect-Bounded-Delay (DBD) [3] while using the Poisson threshold[18]. The results clearly show that UNFD outperforms DBD. DBD with the Poisson threshold does not work for our experimental data. The results support our claim that UNFD may perform well on traffic with more general distribution than Poisson process.

VI. CONCLUSION

In this paper, we considered timing-based detection of packet flows in traffic with time-varying rates. Assuming time-varying traffic, we studied three different problems: detection of unidirectional flow, detection of general flow, and estimation of flow direction. For each problem, a timing-based algorithm with linear complexity was presented with a consistency result under the nonhomogeneous Poisson traffic assumption. Furthermore, the algorithms were tested using the MSN VoIP traffic and the synthetic nonhomogeneous Poisson traffic. The test results are promising.

Even though the algorithms were analyzed under the non-homogeneous Poisson traffic assumption, the intuition behind the algorithms suggests that they would perform well on traffic with more general distribution; our experimental results using the MSN VoIP traffic support this claim.

REFERENCES

[1] S. Staniford-Chen and L. Heberlein, "Holding intruders accountable on the internet," in *Proc. the 1995 IEEE Symposium on Security and Privacy*, Oakland, CA, May 1995, pp. 39–49.
[2] D. Donoho, A. Flesia, U. Shankar, V. Paxson, J. Coit, and S. Staniford, "Multiscale stepping-stone detection: Detecting pairs of jittered interactive streams by exploiting maximum tolerable delay," in *5th International Symposium on Recent Advances in Intrusion Detection, Lecture Notes in Computer Science 2516*, 2002.
[3] T. He and L. Tong, "Detection of Information Flows," *IEEE Trans. Inf. Theory*, vol. 54, pp. 4925–4945, Nov. 2008.
[4] A. Blum, D. Song, and S. Venkataraman, "Detection of Interactive Stepping Stones: Algorithms and Confidence Bounds," in *Conference of Recent Advance in Intrusion Detection (RAID)*, Sophia Antipolis, French Riviera, France, September 2004.
[5] J. Kim and L. Tong, "Timing-based Detection of Packet Forwarding in MANETs," in *11th International Workshop on Signal Processing Advances in Wireless Communications*, Marrakech, Morocco, June 2010.
[6] ——, "Detection of Time-varying Flows in Wireless Networks," in *IEEE Military Communications Conference*, San Jose, CA, November 2010.
[7] ITU-T Recommendation G.114, "One way transmission time."

[18]For DBD, we set τ to be $\tau_0 + \epsilon$, where τ_0 is the closed-form expression of $\lim_{t\to\infty} \bar{\mathrm{R}}_{\mathbf{f}}(t)$ under the independent homogeneous Poisson traffic assumption, evaluated by plugging $\frac{|S_i \cap [0,t]|}{t}$ into λ_i. We set $\epsilon = 0.05$.