This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2365801, IEEE Transactions on Knowledge and Data Engineering

ePeriodicity: Mining Event Periodicity from Incomplete Observations

Zhenhui Li, *Member, IEEE,* Jingjing Wang, and Jiawei Han, *Fellow, IEEE*

Abstract—Advanced technology in GPS and sensors enables us to track physical events, such as human movements and facility usage. Periodicity analysis from the recorded data is an important data mining task which provides useful insights into the physical events and enables us to report outliers and predict future behaviors. To mine periodicity in an event, we have to face real-world challenges of inherently complicated periodic behaviors and imperfect data collection problem. Specifically, the hidden temporal periodic behaviors could be oscillating and noisy, and the observations of the event could be incomplete.

In this paper, we propose a novel probabilistic measure for periodicity and design a practical algorithm, ePeriodicity, to detect periods. Our method has thoroughly considered the uncertainties and noises in periodic behaviors and is provably robust to incomplete observations. Comprehensive experiments on both synthetic and real datasets demonstrate the effectiveness of our method.

Index Terms—Periodicity, Incomplete Observations, Probabilistic Model

1 INTRODUCTION

Periodicity is one of the most common phenomena in the physical world. Animals often have yearly migration patterns; students usually have weekly schedules for classes; and the usage of bedroom, toilet, and kitchen could have daily periodicity, just to name a few. Nowadays, with the rapid development of GPS and mobile technologies, it becomes much easier to monitor such events. For example, cellphones enable us to track human activities [2], GPS devices attached to animals help the scientists to study the animal movement patterns [3], and sensors allow us to monitor the usage of rooms and facilities [4].

Data collected from these devices provides a valuable resource for ecological study, environmental protection, urban planning and emergency response. An observation of an event defined in this paper is a boolean value, that is, whether an event happens or not. An important aspect of analyzing such data is to *detect true periods* hidden in the observations.

Unfortunately, period detection for an event is a challenging problem, due to the *limitations of data collection methods* and the *inherent complexity of periodic behaviors*.

To illustrate these difficulties, let us first take a look at Figure 1. Suppose we have observed the occurrences of an event at timestamps 5, 18, 26, 29, 48, 50, 67, and 79. The observations of the event at other timestamps are not available. It is certainly not an easy task to infer the period directly from these *incomplete* observations. In fact, the issue with incomplete observations is a common problem in data collected from GPS and sensors. For example, a bird can only carry small sensors with one or two reported locations in three to five days. And the locations of a person may only be recorded when he uses his cellphone. Moreover, if a sensor is not functioning or a tracking facility is turned off, it could result in a large portion of missing data. Therefore, we usually have *incomplete observations*, which *are unevenly sampled* and *have large portion of missing data*. Traditional periodicity analysis methods, such as Fourier transform and auto-correlation [5], [6], [7], [3], usually require the data to be *evenly sampled*, that is, there is an observation at every timestamp. Even though some extensions of Fourier transform have been proposed to handle uneven data samples [8], [9], they are still not applicable to the case with very low sampling rate.

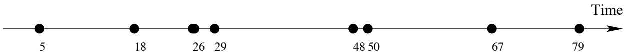


Fig. 1. Incomplete observations.

Second, the periodic behaviors could be inherently *complicated and noisy*. A periodic event does not necessarily happen at *exactly* the same timestamp in each periodic cycle. For example, the time that a person goes to work in the morning might *oscillate* between 8:00 to 10:00. *Noises* could also occur when the "in office" event is expected to be observed on a weekday but fails to happen.

In this paper, we propose a novel algorithm for event period detection, ePeriodicity, which can handle all the aforementioned difficulties occurring in data collection process and periodic behavior complexity

- *Z. Li is with the College of Information Science and Technology, The Pennsylvania State University, University Park, PA, 16802.*
E-mail: jessieli@ist.psu.edu
- *J. Wang and J. Han are with the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801.*
E-mail: {jwang112, hanj}@illinois.edu

This is a substantially extended and revised version of [1], which appears in the Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'12).

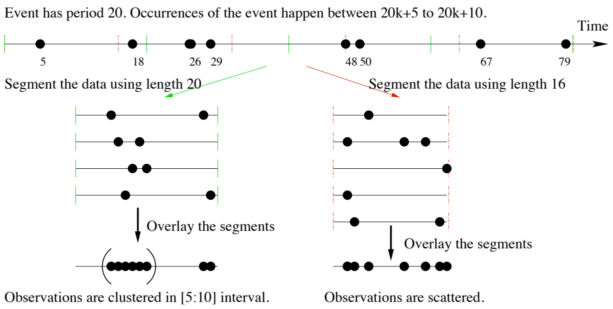

Event has period 20. Occurrences of the event happen between 20k+5 to 20k+10.
Time
5
18
26
29
48
50
67
79
Segment the data using length 20
Segment the data using length 16
Overlay the segments
Overlay the segments
Observations are clustered in [5:10] interval.
Observations are scattered.

Fig. 2. Illustration example of our method.

in a unified framework. The basic idea of ePeriodicity is illustrated in Example 1.

EXAMPLE 1. *Suppose an event has a period* $T = 20$ *and we have eight observations of the event. If we overlay the observations with the correct period* $T = 20$, *we can see in Figure 2 that most of the observations concentrate in time interval* $[5 : 10]$. *However, if we overlay the points with a wrong period, say* $T = 16$, *we cannot observe such clusters.*

As suggested by Example 1, we could segment the timeline using a potential period T and summarize the observations over all the segments. If most of the observations fall into some time intervals, such as interval $[5 : 10]$ in Example 1, T is *likely* to be the true period. In this paper, we formally characterize such likelihood by introducing a probabilistic model for periodic behaviors. The model naturally handles the oscillation and noise issues because the occurrence of an event at any timestamp is now modeled with a probability. Next, we propose a new measure for periodicity based on this model. The measure essentially examines whether the distribution of observations is highly skewed w.r.t a potential period T. As we will see later, even when the observations are incomplete, the overall distribution of observations, after overlaid with the correct T, remains skewed and is similar to the true periodic behavior model.

In summary, our major contributions are as follows. (1) We introduce a probabilistic model for periodic behaviors and a random observation model for incomplete observations. This enables us to model all the variations we encounter in practice in a unified framework. (2) We propose a novel probabilistic measure for periodicity and design a practical algorithm ePeriodicity to detect perods directly from the raw data. We further give rigorous proof of its validity under both the probabilistic periodic behavior model and the random observation model. (3) Comprehensive experiments are conducted on both real data and synthetic data. The results demonstrate the effectiveness of our method.

The rest of the paper is organized as follows. We formally define our period detection problem in Section 2 and introduce our probabilistic measure for periodicity in Section 3. Section 4 discusses the implementaion issues. We report the experimental results in Sections 5, review related work in Section 6 and conclude our study in Section 7.

2 Problem Formulation

In this section, we formally define the problem of period detection for events. We first assume that there is an observation at every timestamp. The case with incomplete observations will be discussed in Section 3.2. We use a binary sequence $\mathcal{X} = \{x(t)\}_{t=0}^{n-1}$ to denote observations. For example, if the event is "in the office", $x(t) = 1$ means this person is in the office at time t and $x(t) = 0$ means this person is *not* in the office at time t. Later we will refer $x(t) = 1$ as a *positive observation* and $x(t) = 0$ as a *negative observation*.

DEFINITION 1 (Periodic Sequence)**.** *A sequence* $\mathcal{X} = \{x(t)\}_{t=0}^{n-1}$ *is said to be periodic if there exists some* $T \in \mathbb{Z}$ *such that* $x(t + T) = x(t)$ *for all values of* t. *We call* T *a period of* $\mathcal{X}$.

A fundamental ambiguity with the above definition is that if T is a period of $\mathcal{X}$, then mT is also a period of $\mathcal{X}$ for any $m \in \mathbb{Z}$. A natural way to resolve this problem is to use the so called *prime period*.

DEFINITION 2 (Prime Period)**.** *The prime period of a periodic sequence is the smallest* $T \in \mathbb{Z}$ *such that* $x(t + T) = x(t)$ *for all values of* t.

For the rest of the paper, unless otherwise stated, we always refer the word "period" to "prime period".

As we mentioned before, in real applications the observed sequences always deviate from the perfect periodicity due to the oscillating behavior and noises. To model such deviations, we introduce a new probabilistic framework, which is based on the *periodic distribution vector* defined below.

DEFINITION 3 (Periodic Distribution Vector)**.** *We call a vector* $\mathbf{p}^T = [p_0^T, \ldots, p_{T-1}^T]$ *other than* $\mathbf{0}^T$ *or* $\mathbf{1}^T$, *where* $p_k^T \in [0, 1], \forall k$, *a periodic distribution vector of length* T.

Here, we need to exclude the trivial cases where $\mathbf{p}^T = \mathbf{0}^T$ or $\mathbf{1}^T$, since they corresponds to constant sequences and are therefore of little interest to us. In this paper, we use the periodic distribution vector to describe the periodicity of any event. The following example illustrates this concept.

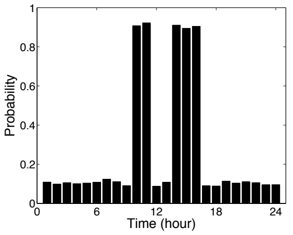


Fig. 3. (Running Example) Periodic distribution vector of an event with daily periodicity ($T_0 = 24$).

EXAMPLE 2 (Running Example)**.** *As an example, assume that a person has a daily periodicity visiting his* *office during 10am-11am and 2pm-4pm. Such behavior can be described by the periodic distribution vector shown in Figure 3, which takes large values at intervals [10:11] and [14:16] and small but nonzero values at other timestamps. We use it as a running example throughout this paper.*

2.1 A Probabilistic Model for Periodicity

In this paper, we model the observation sequence $\mathcal{X}$ as a sequence generated by some unknown stochastic process (*i.e.*, a realization of the random process). In this paper, we use $\{X(t)\}$ to denote the random process, with each $X(t)$ being a random variable.

Of course, not all random processes have periodic behaviors or admit a unique periodic distribution vector. Therefore, we may ask the following question: Under what conditions a random process may exhibit certain periodic behavior which can be captured by a unique periodic distribution vector?

To answer this question, we need to review the important concept of *ergodicity* in probability theory.

DEFINITION 4 (Ergodic Process). *Let $\theta : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ denote the shift operator. A stationary random process $\{Y(t)\}$ is said to be ergodic if either $\mathbf{P}(Y \in A) = 0$ or $\mathbf{P}(Y \in A) = 1$ whenever $A \subseteq \mathbb{R}^{\mathbb{N}}$ is a shift invariant set (i.e., $A = \theta A$).*

Note that here $\mathbf{P}(Y \in A) = 0$ means that the probability measure of the set of realizations of $\{Y(t)\}$ which fall in the set A is 0. For more detailed explanation of the definition, we refer readers to probability theory textbooks such as [10]. Most importantly, if a random process is ergodic, then its statistical properties (such as its mean and variance) can be deduced from a single, sufficient long realization of the process. For our problem, we assume that the occurrence of an event at a particular timestamp in each period (*e.g.*, a person being in his office at 10am for each day) can be modeled by an *ergodic process*. Then, the entire periodic behavior of an event (*e.g.*, a person's daily visits to his office) can be naturally modeled as the mixture of T ergodic sequences.

DEFINITION 5 (Periodically Ergodic Process). *Let $\{\{Y_k(t)\}, k = 0, 1, \ldots, T-1\}$ be the set of T subsequences obtained from $\{X(t)\}$ such that*

$$Y_k(t) = X(t \times T + k), \quad \forall k \in \{0, 1, \ldots, T-1\}. \quad (1)$$

Then, a random process $\{X(t)\}$ is said to be periodically ergodic with period T if every $\{Y_k(t)\}$ is ergodic.

An important result concerning the ergodic process is the Birkhoff's Ergodic theorem, which asserts that the time average of any ergodic process converges to a fixed value.

THEOREM 1 (Birkhoff's Ergodic Theorem). *For any ergodic process $\{Y(t)\}$, if $\mathbb{E}[|Y(0)|] < \infty$, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} Y(t) \to \mathbb{E}[Y(0)] \quad a.s.$$

Here, $|\cdot|$ denotes the absolute value, and $\mathbb{E}[\cdot]$ denotes the expectation value of a random variable. We again refer interested readers to [10] for the proof. In this paper, we focus on its implication on our problem, which leads to the following corollary.

COROLLARY 1. *Let $\{X(t)\}$ be a periodic ergodic process and $\mathcal{X} = \{x(t)\}$ be any realization of the process. Then there exists a unique vector $\mathbf{p}^T = [p_0^T, \ldots, p_{T-1}^T]$ such that the following holds with probability (w.p.) 1:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} x(t \times T + k) = p_k^T. \quad (2)$$

Clearly, when $\{X(t)\}$ is a binary non-constant process, $\mathbf{p}^T$ is the *periodic distribution vector* of $\{X(t)\}$, since it summarizes the long-term periodic behavior of the process.

As we mentioned before, the key property a periodically ergodic process is that its behavior at any fixed timestamp with respect to the period T (represented by a probability distribution) does not change over time, and can be estimated by the mean of all the samples. However, in our problem we only have seen a small portion of the samples, so one may wondering if we can still reliably estimate the periodic behavior by summarizing the observations with respect to T. Mathematically, the question becomes: For any subsequence of $\{Y_k(t)\}$ with timestamps $\{l_1, l_2, \ldots\}$, does the sample mean $\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} Y(l_t)$ still converge to p_k^T?

In probability theory, a subsequence for which this condition holds is called an *admissible* subsequence. In general, not all the subsequences are admissible for an arbitrary ergodic process $\{Y(t)\}$. However, the following lemma shows that, as long as $Y(l_1)$ is independent of $Y(l_2)$ when $|l_1 - l_2| \to \infty$, all the subsequences must be admissible.

LEMMA 1 ([11], [12]). *A stationary process $\{Y(t)\}$ is said to be* **mixing** *if for any measurable sets $A_1, A_2 \subseteq \mathbb{R}^{\mathbb{N}}$,*

$$\lim_{n\to\infty} \mathbf{P}(Y \in A_1, \theta^n Y \in A_2) = \mathbf{P}(Y \in A_1)\mathbf{P}(Y \in A_2).$$

Further, all subsequences of $\{Y(t)\}$ are admissible if and only if $\{Y(t)\}$ is mixing.

Note that mixing implies ergodicity [10]. Similarly to the definition of periodically ergodic processes, we call a process $\{X(t)\}$ *periodically mixing* if every $\{Y_k(t)\}$ is mixing. For the rest of the paper, we will focus on processes which are periodically mixing, and define our period detection problem as follows.

PROBLEM 1 (Event Period Detection). *Given a binary sequence $\mathcal{X}$ generated by some periodically mixing process $\{X(t)\}$ with (an unknown) periodic distribution vector $\mathbf{p}^{T_0}$, find T_0.*

Note that, by the definition of periodically mixing process, we assume a single, time-invariant period in the observation sequence. Also, the timestamps are assumed to be synchronized (non-drifting), although later we will show empirically that our method is insensitive to moderate oscillations in the actual sampling time.

In Section 3, we propose a novel measure for periodicity, and show that it is guaranteed to find the true period T_0 given any realization of a periodically mixing process, even with incomplete observations. Before that, we use two examples to demonstrate the practicability of our probabilistic model.

2.2 Two Examples

In this section, we give two important examples of periodically mixing processes, namely the *independent Bernoulli processes* and the *periodically inhomogeneous Markov chains*, and demonstrate how they can be used to model real periodic events.

2.2.1 Independent Bernoulli Processes

Suppose $\{X(t)\}$ is a random process with each $X(t)$ independently distributed according to $Bernoulli(p^T_{\text{mod }(t,T)})$, then $\{X(t)\}$ is a periodically ergodic process with periodic distribution vector $\mathbf{p}^T$. Note that if we restrict the value of each p_k^T to $\{0, 1\}$ only, then the resulting sequence is *strictly* periodic according to Definition 1. In addition, it is trivial to see that any i.i.d. sequence is mixing. Therefore, all independent Bernoulli processes are periodically mixing.

As an example, assuming that the probability of a person's visit to his office at each timestamp is independent of that at any other timestamps, then his periodic behavior may be modeled by an independent Bernoulli process with the periodic distribution vector shown in Figure 3.

2.2.2 Periodically Inhomogeneous Markov Chains

The independent Bernoulli process is a very simple and intuitive way to model a periodic event. However, it cannot model the dependency of consecutive observations in a sequence. For example, if we know that a person is in his office at 9am for one day, then it is very likely that he is also in his office at 10am for the same day. In order to model such dependency, we now introduce another type of random process, called *periodically inhomogeneous Markov chains*.

In general, a random process $\{X(t)\}$ is a *Markov chain* if, given the present state, the future and past states are independent. Therefore, any Markov chain with a finite state space S[1] can be characterized by a series of transition matrices $\{P(t)\}$, where the (i, j)-th entry of $P(t)$, $p_{ij}(t)$, is the probability of going from state s_i to state s_j at timestamp t:

$$p_{ij}(t) = \mathbf{P}(X(t + 1) = s_j | X(t) = s_i). \tag{3}$$

To model the periodic behavior of an event, we assume that the transition matrix $P(t)$ is changing over time (inhomogeneous) but repeats itself after every T timestamps.

DEFINITION 6 ([13]). *We call $\mathcal{X}$ a* **periodically inhomogeneous Markov chain**, *if there exists a positive integer T such that for all values of t,*

$$P(t) = P(t + T). \tag{4}$$

With this definition, our key observation is that, just like the homogeneous (time-invariant) Markov chain with transition matrix P which *often* admits a unique stationary distribution $\boldsymbol{\pi}$ (*i.e.*, $\boldsymbol{\pi}P = \boldsymbol{\pi}$), a periodically inhomogeneous Markov chain $\{X(t)\}$ can *often* be decomposed into T homogeneous Markov chains $\{\{Y_k(t)\}, k = 0, \dots, T - 1\}$, each admitting a unique stationary distribution $\boldsymbol{\pi}_k$. In such cases, the i-th entry of $\boldsymbol{\pi}_k$, $\boldsymbol{\pi}_k(i)$, can be viewed as the long run proportion of time that the Markov chain $\{Y_k(t)\}$ will stay in the i-th state. Further, let $\bar{\pi}_k = \sum_{s_i \in S} s_i \boldsymbol{\pi}_k(i)$, we can show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} Y_k(t) \to \bar{\pi}_k \quad a.s.$$

We call such a Markov chain a *periodically ergodic Markov chain* according to Definition 5.

However, unlike the case of i.i.d. sequences, a periodically inhomogeneous Markov chain is not necessarily periodically ergodic. The conditions for a Markov chain to be periodically ergodic has been previously studied, for example, in [13]. We summarize the main result below.

DEFINITION 7 (Reducibility of Markov Chain). *A homogeneous Markov chain $\{Y(t)\}$ is* **irreducible** *if for any two states s_i and s_j, if started in s_i, the chain has a non-zero probability transitioning into s_j. That is,*

$$\mathbf{P}(Y(t) = s_j | Y(0) = s_i) > 0 \quad \text{for some } t \geq 0. \tag{5}$$

Now, given a finite-state periodically inhomogeneous Markov chain $\{X(t)\}$, we construct T homogeneous Markov chains $\{\{Y_k(t)\}, k = 0, \dots, T - 1\}$ from $\{X(t)\}$ as follows:

$$Y_k(t) = X(t \times T + k). \tag{6}$$

Denote their transition matrices by $\{P^k\}_{k=0}^{T-1}$, where

$$P^k = P(k)P(k + 1)\cdots P(k + T - 1), \quad \forall k. \tag{7}$$

Then, the following lemma states that $\{X(t)\}$ is periodically ergodic if there exists a $\{Y_k(t)\}$ which is irreducible.

LEMMA 2 ([13]). *If there exists a $k \in \{0, 1, \dots, T - 1\}$ such that $\{Y_k(t)\}$ is a homogeneous irreducible Markov chain, then there exists a unique probability measure $\boldsymbol{\pi}_k$ for each k such that*

$$\boldsymbol{\pi}_k P^k = \boldsymbol{\pi}_k, \quad \text{and} \quad \boldsymbol{\pi}_k P(k) = \boldsymbol{\pi}_{k+1}.\text{[2]} \tag{8}$$

[1] In this paper, we assume that each state $s_i \in S$ takes value in $\{0, 1, \dots, m - 1\}$, where m is the size of S.

[2] Here, $\boldsymbol{\pi}_T$ is understood to be $\boldsymbol{\pi}_0$.

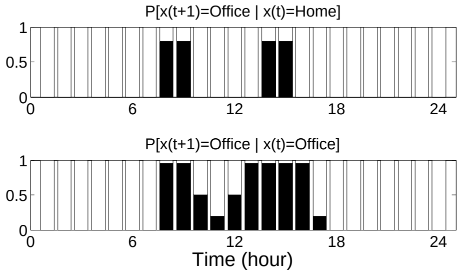


Fig. 4. A periodically inhomogeneous Markov chain ($T_0 = 24$).

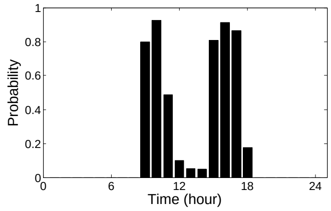


Fig. 5. The corresponding periodic distribution vector for the Markov chain shown in Figure 4.

In addition, $\{X(t)\}$ is periodically ergodic and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} X(t \times T + k) \to \bar{\pi}_k \quad a.s.$$

For our period detection problem, the Markov chain only takes two states $S = \{s_1, s_2\} = \{0, 1\}$. In such case, each $\bar{\pi}_k$ has value in $[0, 1]$ and $\Pi = [\bar{\pi}_0, \bar{\pi}_1, \ldots, \bar{\pi}_{T-1}]$ is the periodic distribution vector of $\{X(t)\}$. We illustrate the concepts and results described so far using the following example.

EXAMPLE 3. *In this example, we model a person's daily behavior as the periodically inhomogeneous Markov chain with its transition matrices shown in Figure 4. As one can see, the person has high probability going to work (from home) at 8-9am in the morning and 2-3pm in the afternoon. In addition, he tends to stay in the office between 9am-5pm, with the exception that he may go back home around noon (11am-1pm) for lunch. We further show the corresponding periodic distribution vector Π in Figure 5.*

Comparing to the Independent Bernoulli model, the Markov chain enables us to model the time dependency of the person's behavior. For example, if he is in the office at 4pm, the probability that he will stay in the office at 5pm is very high (p = 0.95), whereas if he is at home at 4pm (possibly due to illness), it is very unlikely that he will be in the office at 5pm (p = 0).

Finally, an ergodic Markov chain is not necessarily mixing. Consider the following example.

EXAMPLE 4. *Let $\{Y(t)\}$ be a homogenous Markov chain with transition matrix*

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{9}$$

Then, $\{Y(t)\}$ is an ergodic sequence with a unique stationary distribution $\boldsymbol{\pi} = [0.5, 0.5]$. However, letting $l_t = 2t$, for this subsequence we have $\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} Y(l_t)$ is equal to either 0 or 1, depending on the value of $Y(0)$. Hence, this subsequence is not admissible.

In the above example, we see that the value $Y(t)$ is solely determined by the initial value $Y(0)$, no matter how large t is. To provide the additional conditions for a (periodically) ergodic Markov chain to be (periodically) mixing, we need the following definition.

DEFINITION 8. *A Markov chain is said to be* **aperiodic** *if for any state s_i, there exists t_0 such that for all $t \geq t_0$,*

$$\mathbf{P}(Y(t) = s_i | Y(0) = s_i) > 0. \tag{10}$$

Obviously, the Markov chain in Example 4 is not aperiodic. With this definition, we have the following lemma which gives the conditions for a periodically ergodic Markov chain to be periodically mixing.

LEMMA 3 **([14]).** *A periodically ergodic Markov chain $\{X(t)\}$ is periodically mixing if and only if there exists a $k \in \{0, 1, \ldots, T-1\}$ such that $\{Y_k(t)\}$ is aperiodic.*

Table 1 summarizes the concepts and conditions w.r.t. the two examples we discussed in this section. Meanwhile, we emphasize that the periodically mixing condition is a fairly general one in probability theory. Its scope certainly goes far beyond the above examples, and so does the periodicity measure we are going to introduce next.

3 The Proposed Method

As we see in Example 2, when we overlay the binary sequence with its true period T_0, the resulting sequence correctly reveals its underlying periodic behavior. In this section, we make this observation formal using the concept of periodic distribution vector. Then, we propose a novel probabilistic measure of periodicity based on this observation and prove its validity even when the observations are incomplete.

3.1 A Probabilistic Measure of Periodicity

Given a binary sequence $\mathcal{X}$, we define $S^+ = \{t : x(t) = 1\}$ and $S^- = \{t : x(t) = 0\}$ as the collections of timestamps with 1's and 0's, respectively. For a candidate period T, let $\mathcal{I}_T$ denote the power set of $[0 : T-1]$. Then, for any set of timestamps (*possibly non-consecutive*) $I \in \mathcal{I}_T$, we can define the collections of original timestamps that fall into this set after overlay as follows:

$$S_I^+ = \{t \in S^+ : \mathcal{F}_T(t) \in I\},\ S_I^- = \{t \in S^- : \mathcal{F}_T(t) \in I\},$$

where $\mathcal{F}_T(t) = \mod(t, T)$, and further compute the ratios of 1's and 0's whose corresponding timestamps fall into I after overlay:

$$\mu_{\mathcal{X}}^+(I, T) = \frac{|S_I^+|}{|S^+|}, \quad \mu_{\mathcal{X}}^-(I, T) = \frac{|S_I^-|}{|S^-|}. \tag{11}$$

TABLE 1
Summary of conditions for different random processes to be (periodic) ergodic and (periodic) mixing.

Random process	(Periodic) ergodic	(Periodic) mixing
Independent Bernoulli	Always	Always
Homogeneous MC	$\{X(t)\}$ is irreducible	$\{X(t)\}$ is irreducible & aperiodic
Periodically inhomogeneous MC	$\exists k$ s.t. $\{Y_k(t)\}$ is irreducible	$\exists k$ s.t. $\{Y_k(t)\}$ is irreducible & aperiodic

The following lemma says that these ratios indeed reveal the true probabilistic model parameters, given that the observation sequence is sufficiently long.

LEMMA 4. *Suppose $\mathcal{X} = \{x(t)\}_{t=0}^{n-1}$ is a binary sequence generated by any periodically ergodic process with periodic distribution vector $\mathbf{p}^{T_0}$ of length T_0, write $q_i^{T_0} = 1 - p_i^{T_0}$. Then, for any T and $I \in \mathcal{I}_T$, the following holds w.p. 1:*

$$\begin{aligned}\lim_{n\to\infty}\mu^+_{\mathcal{X}}(I,T) &= \sum_{i\in I}\left(\frac{1}{T}\sum_{j=0}^{T_0-1}\frac{p^{T_0}_{\mathcal{F}_{T_0}(i+j\times T)}}{\sum_{k=0}^{T_0-1}p_k^{T_0}}\right),\\ \lim_{n\to\infty}\mu^-_{\mathcal{X}}(I,T) &= \sum_{i\in I}\left(\frac{1}{T}\sum_{j=0}^{T_0-1}\frac{q^{T_0}_{\mathcal{F}_{T_0}(i+j\times T)}}{\sum_{k=0}^{T_0-1}q_k^{T_0}}\right).\end{aligned}$$

Proof: The proof is a straightforward application of the Birkhoff's Ergodic Theorem, and we only prove the first equation. With a slight abuse of notation we write $S_i = \{t : \mathcal{F}_T(t) = i\}$ and $S_i^+ = \{t \in S^+ : \mathcal{F}_T(t) = i\}$. We further partition S_i into T_0 subsets such that

$$S_{i,j} = \{i + jT, i + (j + T_0)T, i + (j + 2T_0)T, \ldots\},$$

where $j = \{0, \ldots, T_0 - 1\}$. Since each subsequence $\{x(t) : t \in S_{i,j}\}$ is the realization of a single mixing process, we have w.p. 1 that

$$\begin{aligned}\lim_{n\to\infty}\frac{|S_i^+|}{n} &= \lim_{n\to\infty}\frac{\sum_{j=0}^{T_0-1}\sum_{t\in S_{i,j}}x(t)}{|S_i|}\cdot\frac{|S_i|}{n}\\ &= \frac{\sum_{j=0}^{T_0-1}p^{T_0}_{\mathcal{F}_{T_0}(i+j\times T)}}{T_0 T},\end{aligned}$$

where we use $\lim_{n\to\infty}\frac{|S_i|}{n} = \frac{1}{T}$ for the last equality. Also, since the random process can be decomposed into T_0 mixing processes, we have w.p. 1 that $\lim_{n\to\infty}|S^+|/n = \frac{1}{T_0}\sum_{k=0}^{T_0-1}p_k^{T_0}$. Therefore,

$$\begin{aligned}\lim_{n\to\infty}\mu^+_{\mathcal{X}}(I,T) &= \lim_{n\to\infty}\frac{|S_I^+|/n}{|S^+|/n} = \lim_{n\to\infty}\frac{\sum_{i\in I}|S_i^+|/n}{|S^+|/n}\\ &= \sum_{i\in I}\left(\frac{1}{T}\sum_{j=0}^{T_0-1}\frac{p^{T_0}_{\mathcal{F}_{T_0}(i+j\times T)}}{\sum_{k=0}^{T_0-1}p_k^{T_0}}\right).\end{aligned}$$

□

Note that, if $T = T_0$, the equations in Lemma 4 can be simplified to:

$$\lim_{n\to\infty}\mu^+_{\mathcal{X}}(I,T_0) = \frac{\sum_{i\in I}p_i^{T_0}}{\sum_{i=0}^{T_0-1}p_i^{T_0}}, \lim_{n\to\infty}\mu^-_{\mathcal{X}}(I,T_0) = \frac{\sum_{i\in I}q_i^{T_0}}{\sum_{i=0}^{T_0-1}q_i^{T_0}}.$$

This suggests that ratio of positive (negative) samples that fall into I after overlay indeed converges to what one would expect according to $\mathbf{p}^{T_0}$. Based on this result, we now introduce our measure of periodicity. For any $I \in \mathcal{I}_T$, we define its discrepancy score as:

$$\Delta_{\mathcal{X}}(I,T) = \mu^+_{\mathcal{X}}(I,T) - \mu^-_{\mathcal{X}}(I,T). \tag{12}$$

Then, the periodicity measure of $\mathcal{X}$ w.r.t. period T is:

$$\gamma_{\mathcal{X}}(T) = \max_{I\in\mathcal{I}_T}\Delta_{\mathcal{X}}(I,T). \tag{13}$$

It is obvious that $\gamma_{\mathcal{X}}(T)$ is bounded: $0 \le \gamma_{\mathcal{X}}(T) \le 1$. Moreover, $\gamma_{\mathcal{X}}(T) = 1$ if and only if $\mathcal{X}$ is strictly periodic with period T. But more importantly, we have the following lemma, which states that under our probabilistic model, $\gamma_{\mathcal{X}}(T)$ is indeed a desired measure of periodicity.

LEMMA 5. *Suppose $\mathcal{X}$ is a binary sequence generated by any periodically mixing process with periodic distribution vector $\mathbf{p}^{T_0}$ of length T_0, then the following holds w.p. 1:*

$$\lim_{n\to\infty}\gamma_{\mathcal{X}}(T) \le \lim_{n\to\infty}\gamma_{\mathcal{X}}(T_0), \quad \forall T \in \mathbb{Z}.$$

Proof: Define

$$c_i = \frac{p_i^{T_0}}{\sum_{k=0}^{T_0-1}p_k^{T_0}} - \frac{q_i^{T_0}}{\sum_{k=0}^{T_0-1}q_k^{T_0}},$$

it is easy to see that the value $\lim_{n\to\infty}\gamma_{\mathcal{X}}(T_0)$ is achieved by $I^* = \{i \in \{0, \ldots, T_0 - 1\} : c_i > 0\}$. So it suffices to show that for any $T \in \mathbb{Z}$ and $I \in \mathcal{I}_T$,

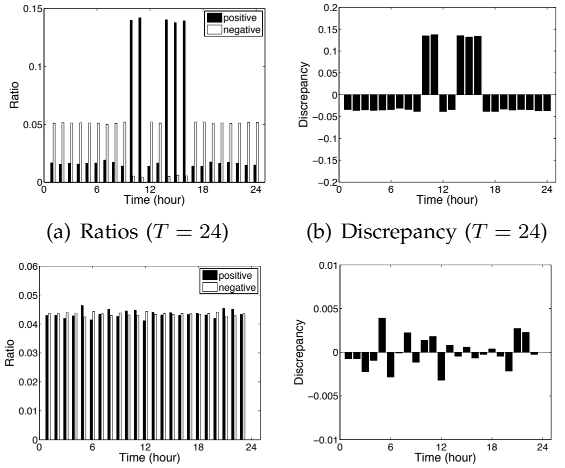
$$\lim_{n\to\infty}\Delta_{\mathcal{X}}(I,T) \le \lim_{n\to\infty}\Delta_{\mathcal{X}}(I^*,T_0) = \sum_{i\in I^*}c_i.$$

Meanwhile, from Lemma 4, we have

$$\begin{aligned}\lim_{n\to\infty}\Delta_{\mathcal{X}}(I,T) &= \frac{1}{T}\sum_{i\in I}\sum_{j=0}^{T_0-1}\left(\frac{p^{T_0}_{\mathcal{F}_{T_0}(i+j\times T)}}{\sum_{k=0}^{T_0-1}p_k^{T_0}} - \frac{q^{T_0}_{\mathcal{F}_{T_0}(i+j\times T)}}{\sum_{k=0}^{T_0-1}q_k^{T_0}}\right)\\ &= \frac{1}{T}\sum_{i\in I}\sum_{j=0}^{T_0-1}c_{\mathcal{F}_{T_0}(i+j\times T)}\\ &\le \frac{1}{T}\sum_{i\in I}\sum_{j=0}^{T_0-1}\max(c_{\mathcal{F}_{T_0}(i+j\times T)}, 0)\\ &\le \frac{1}{T}\sum_{j=0}^{T_0 T-1}\max(c_{\mathcal{F}_{T_0}(j)}, 0)\\ &= \frac{1}{T}\times T\sum_{i\in I^*}c_i = \sum_{i\in I^*}c_i,\end{aligned}$$

where the third equality uses the definition of I^*. So the proof is complete. □

Note that, similar to the deterministic case, the ambiguity of multiple periods still exists as we can easily see that $\lim_{n\to\infty}\gamma_{\mathcal{X}}(mT_0) = \lim_{n\to\infty}\gamma_{\mathcal{X}}(T_0)$ for all $m \in \mathbb{Z}$. But in this paper we are only interested in finding the smallest one.



(a) Ratios ($T = 24$)

(b) Discrepancy ($T = 24$)

(c) Ratios ($T = 23$)

(d) Discrepancy ($T = 23$)

Fig. 6. (a) and (c): Ratios of 1's and 0's at a single timestamp (*i.e.*, $\mu^+_{\mathcal{X}}(\cdot, T)$ and $\mu^-_{\mathcal{X}}(\cdot, T)$) when $T = 24$ and $T = 23$, respectively. (b) and (d): Discrepancy scores at a single timestamp (*i.e.* $\Delta_{\mathcal{X}}(\cdot, T)$) when $T = 24$ and $T = 23$.

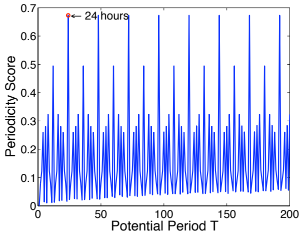


Fig. 7. Periodicity scores of potential periods.

EXAMPLE 5 (Running Example (cont.)). *Suppose our observation sequence $\mathcal{X}$ is generated by an independent Bernoulli process (see Section 2.2.1) with the periodic distribution vector shown in Figure 3. When we overlay the sequence using potential period $T = 24$, Figure 6(a) shows that positive observations have high probability to fall into the set of timestamps: $\{10, 11, 14, 15, 16\}$. However, when using the wrong period $T = 23$, the distribution is almost uniform over time, as shown in Figure 6(c). Consequently, we see large discrepancy scores for T=24 (Figure 6(b)) whereas the discrepancy scores are very small for T=23 (Figure 6(d)). Therefore, we will have $\gamma_{\mathcal{X}}(24) > \gamma_{\mathcal{X}}(23)$. Figure 7 shows the periodicity scores for all potential periods in $[1 : 200]$. We can see that the score is maximized at $T = 24$, which is the true period of the sequence.*

3.2 Random Observation Model

Next, we extend our analysis on the proposed periodicity measure to the case of incomplete observations with a random observation model. To this end, we introduce a new label "-1" to the sequence $\mathcal{X}$ to indicate that the observation is unavailable at a specific timestamp. In the random observation model, each observation $x(t)$ is associated with a probability $d_t \in [0, 1]$ and we write $\mathbf{d} = \{d_t\}_{t=0}^{n-1}$.

DEFINITION 9. *Let $\mathcal{X}_0$ be a binary sequence generated by any periodically mixing process. A sequence $\mathcal{X}$ is said to be generated according to $(\mathcal{X}_0, \mathbf{d})$ if*

$$x(t) = \begin{cases} x_0(t) & \text{w.p. } d_t \\ -1 & \text{w.p. } 1 - d_t \end{cases} \tag{14}$$

In general, we may assume that each d_t is independently drawn from some fixed but unknown distribution f over the interval $[0, 1]$. To avoid the trivial case where $d_t \equiv 0$ for all t, we further assume that it has nonzero mean: $\rho_f > 0$. Although this model seems to be very flexible, in the section we prove that our periodicity measure is still valid. In order to do so, we need the following lemma, which states that $\mu^+_{\mathcal{X}}(I, T)$ and $\mu^-_{\mathcal{X}}(I, T)$ remain the same as before, assuming infinite length observation sequence.

LEMMA 6. *Let $\mathcal{X}_0$ be a binary sequence generated by any periodically mixing process with periodic distribution vector $\mathbf{p}^{T_0}$. Suppose $\mathbf{d} = \{d_t\}_{t=0}^{n-1}$ are i.i.d. random variables in $[0, 1]$ with nonzero mean, and a sequence $\mathcal{X}$ is generated according to $(\mathcal{X}_0, \mathbf{d})$, write $q_i^{T_0} = 1 - p_i^{T_0}$. Then, for any T and $I \in \mathcal{I}_T$, the following holds w.p. 1:*

$$\begin{aligned} \lim_{n\to\infty} \mu^+_{\mathcal{X}}(I, T) &= \sum_{i \in I} \left(\frac{1}{T} \sum_{j=0}^{T_0-1} \frac{p^{T_0}_{\mathcal{F}_{T_0}(i+j\times T)}}{\sum_{k=0}^{T_0-1} p_k^{T_0}} \right), \\ \lim_{n\to\infty} \mu^-_{\mathcal{X}}(I, T) &= \sum_{i \in I} \left(\frac{1}{T} \sum_{j=0}^{T_0-1} \frac{q^{T_0}_{\mathcal{F}_{T_0}(i+j\times T)}}{\sum_{k=0}^{T_0-1} q_k^{T_0}} \right). \end{aligned}$$

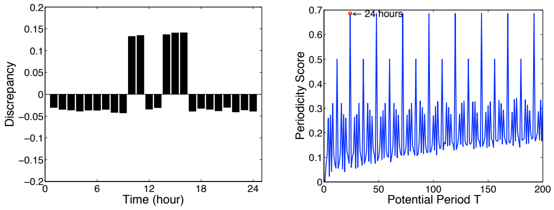
The proof is similar to that of Lemma 4 and is given in Appendix A. Since our periodicity measure only depends on $\mu^+_{\mathcal{X}}(I, T)$ and $\mu^-_{\mathcal{X}}(I, T)$, it is now straightforward to prove its validity under the random observation model. We summarize our main result below.

THEOREM 2. *Let $\mathcal{X}_0$ be a binary sequence generated by any periodically mixing process with periodic distribution vector $\mathbf{p}^{T_0}$. Suppose $\mathbf{d} = \{d_t\}_{t=0}^{n-1}$ are i.i.d. random variables in $[0, 1]$ with nonzero mean, and a sequence $\mathcal{X}$ is generated according to $(\mathcal{X}_0, \mathbf{d})$, then the following holds w.p. 1:*

$$\lim_{n\to\infty} \gamma_{\mathcal{X}}(T) \leq \lim_{n\to\infty} \gamma_{\mathcal{X}}(T_0), \quad \forall T \in \mathbb{Z}.$$

The proof is exactly the same as that of Lemma 5 given the result of Lemma 6, hence is omitted here.

Here we make two useful comments on this result. First, the assumption that d_t's are independent of each other plays an important role in the proof. In fact, if this does not hold, the observation sequence could exhibit very different periodic behavior from its underlying periodic distribution vector. But a thorough discussion on this issue is beyond the scope of this paper. Second, this result only holds exactly with infinite length sequences. However, it provides a good estimate on the situation with finite length sequences, assuming that the sequences are long enough. Note that this length requirement is particularly important when a majority of samples are missing (*i.e.*, ρ_f is close to 0). We will discuss this issue in more detail in Section 4.



(a) Discrepancy ($T = 24$) (b) Periodicity scores

Fig. 8. Period detection with unknown observations.

EXAMPLE 6 (Running Example (cont.)). *To introduce random observations, we sample the original sequence with sampling rate* 0.2. *The generated sequence will have* 80% *of its entries marked as unknown. Comparing Figure 8(a) with Figure 6(b), we can see very similar discrepancy scores over time. Random sampling has little effect on our period detection method. As shown in Figure 8(b), we can still detect the correct period at* 24.

3.3 Handling Sequences without Negative Samples

In many real world applications, negative samples may be completely unavailable to us. For example, if we have collected data from a local cellphone tower, we will know that a person is in town when he makes phone call through the local tower. However, we are not sure whether this person is in town or not for the rest of time, because he could either be out of town or simply not making any call. In this case, the observation sequence $\mathcal{X}$ takes value in $\{1, -1\}$ only, with -1 indicating the missing entries. In this section, we modify our measure of periodicity to handle this case.

Note that due to the lack of negative samples, $\mu^-_{\mathcal{X}}(I, T)$ can no longer be computed from $\mathcal{X}$. Thus, we need find another quantity to compare $\mu^+_{\mathcal{X}}(I, T)$ with. To this end, consider a binary sequence $\mathcal{U} = \{u(t)\}_{t=0}^{n-1}$ which is generated by an i.i.d. Bermoulli(p) random process for some fixed $p > 0$. It is easy to see that for any T and $I \in \mathcal{I}_T$, we have

$$\lim_{n\to\infty} \mu^+_{\mathcal{U}}(I, T) = \frac{|I|}{T}. \tag{15}$$

This corresponds to the case where the positive samples are evenly distributed over all entries after overlay. So we propose the following new discrepancy score for I:

$$\Delta^+_{\mathcal{X}}(I, T) = \mu^+_{\mathcal{X}}(I, T) - \frac{|I|}{T}, \tag{16}$$

and define the periodicity measure as:

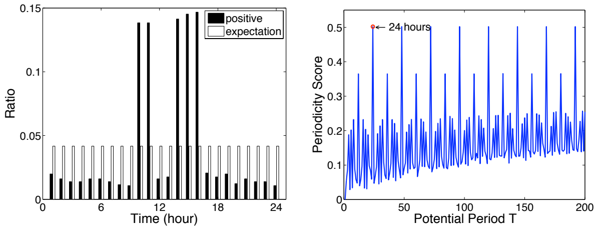
$$\gamma^+_{\mathcal{X}}(T) = \max_{I\in\mathcal{I}_T} \Delta^+_{\mathcal{X}}(I, T). \tag{17}$$

In fact, with some slight modification to the proof of Lemma 5, we can show that it is indeed a desired measure for periodicity under our probabilistic model.

THEOREM 3. *Let $\mathcal{X}_0$ be a binary sequence generated by any periodically mixing process with periodic distribution vector $\mathbf{p}^{T_0}$. Suppose $\mathbf{d} = \{d_t\}_{t=0}^{n-1}$ are i.i.d. random variables in $[0, 1]$ with nonzero mean, and a sequence $\mathcal{X}$ is generated according to $(\mathcal{X}_0, \mathbf{d})$, then the following holds w.p. 1:*

$$\lim_{n\to\infty} \gamma^+_{\mathcal{X}}(T) \leq \lim_{n\to\infty} \gamma^+_{\mathcal{X}}(T_0), \quad \forall T \in \mathbb{Z}.$$

The proof is given in Appendix B. Here, we note that this new measure $\gamma^+_{\mathcal{X}}(T)$ can also be applied to the cases where negative samples are available. Given the same validity result, readers may wonder if it can replace $\gamma_{\mathcal{X}}(T)$. This is certainly not the case in practice, as our results only hold exactly when the sequence has infinite length. As we will see from the experiment results, negative samples indeed provide additional information for period detection in finite length observation sequences.



(a) Distribution ($T = 24$) (b) Periodicity scores

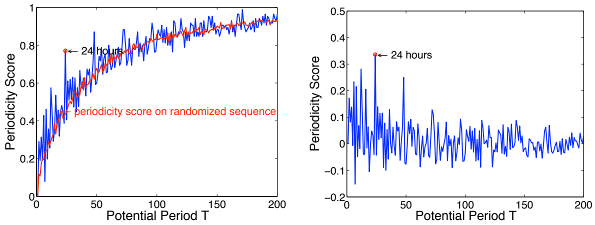
Fig. 9. (Running Example) Period detection on sequences without negative samples.

EXAMPLE 7 (Running Example (cont.)). *In this example we further marked all the negative samples in the sequence we used in Example 6 as unknown. When there is no negative samples, the portion of positive samples at a single timestamp i is expected to be $\frac{1}{T}$, as shown in Figure 9(a). The discrepancy scores when $T = 24$ still have large values at $\{10, 11, 14, 15, 16\}$. Thus the correct period can be successfully detected as shown in Figure 9(b).*

4 ALGORITHM

In Section 3, we have introduced our periodicity measure for any potential period T. Our period detection algorithm ePeriodicity simply computes the periodicity scores for every T and report the one with the highest score. In this section, we first address a practical issue when applying it to finite length sequence and then discuss the time complexity of the algorithm.

Normalization. As one may already notice in our running example, we usually see a general increasing trend of periodicity scores $\gamma_{\mathcal{X}}(T)$ and $\gamma^+_{\mathcal{X}}(T)$ for a larger potential period T. This trend becomes more dominating as the number of observations decreases. For example, the original running example has observations for 1000 days. If the observations are only for 20 days, our method may obtain incorrect period detection result, as the case shown in Figure 10(a). In fact, this phenomenon is expected and can be understood in the following way. Let us take $\gamma^+_{\mathcal{X}}(T)$ as an example. Given a sequence $\mathcal{X}$ with *finite number* of positive observations, it is easy to see that the size of I that maximizes $\gamma^+_{\mathcal{X}}(T)$ for any T is bounded above by the number of positive observations. Therefore the value $\frac{|I^*|}{T}$ always decreases as T increases, no matter whether or not T is a true period of $\mathcal{X}$.



(a) Periodicity scores (b) Normalized scores

Fig. 10. Normalization of periodicity scores.

To remedy this issue, we use the periodicity scores of *randomized* sequences to normalize the original periodicity scores. Specifically, we randomly permute the positions of observations along the timeline and compute the periodicity score for each potential period T. This procedure is repeated N times and the average periodicity scores over N trials are output as the base scores. The redline in Figure 10(a) shows the base scores generated from randomized sequences by setting $N = 10$, which agree well with the trend.

For every potential period T, we subtract the base score from the original periodicity score, resulting in the normalized periodicity score. Note that the normalized score also slightly favors shorter period, which helps us avoid detecting duplicated periods (*i.e.*, multiples of the prime period).

Time/Space Complexity Analysis. For every potential period T, it takes $O(n)$ time to compute discrepancy score for a single timestamp (*i.e.*, $\frac{|S^+_i|}{|S^+|} - \frac{|S^-_i|}{|S^-|}$) and then $O(T)$ time to compute the periodicity score $\gamma_{\mathcal{X}}(T)$ (*i.e.*, find all the timestamps whose discrepancy scores are above 0 and obtain the aggregate score). Since potential period should be in range $[1 : n]$, the time complexity of our algorithm is $O(n^2)$. In practice, it is usually unnecessary to try all the potential periods. For example, we may have common sense that the periods will be no larger than certain values. So we only need to try potential periods up to n_0, where $n_0 \ll n$. This will make our algorithm efficient in practice with time complexity as $O(n \times n_0)$. In this paper, we fix $n_0 = 200$ for all the experiments. Of course, in practice, the choice of n_0 heavily relies on prior knowledge about the range in which the true period lies and is problem-dependent. Also, the computational cost would be higher if a large n_0 is used (*e.g.*, to detect large period).

In Figure 11, we show the computation time of ePeriodicity as a function of n and n_0 on a synthetic dataset (see Section 5.1 for details about our synthetic dataset generation). The experiment is performed in MATLAB on a desktop PC with 3.40GHz CPU and 12GB memory. The default parameters are $n = T \times TN = 24 \times 1000 = 24000$ and $n_0 = 200$. It clearly shows that the time grows linearly in both n and n_0. In addition, by limiting n_0 to 200, our algorithm takes less than 1 second to process a sequence of length 120000. For a sequence of length 24000, it takes about 0.8 second to detect any period up to 1000.

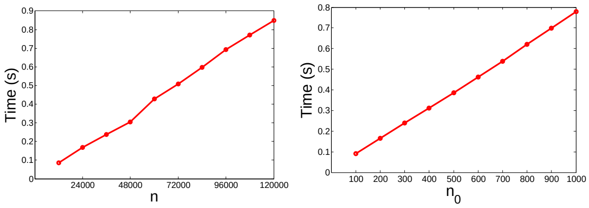


Fig. 11. Time of our algorithm.

Finally, it is easy to see that the space complexity of our algorithm is $O(n)$, as we simply store the input sequence in an array.

5 Experiment

In this section, we systematically evaluate the proposed techniques on both synthetic and real datasets.

5.1 Synthetic Dataset Generation

To test the effectiveness of our method under various scenarios, we take the following steps to generate a synthetic test sequence SEQ according to the *periodically inhomogeneous Markov chain* model. Meanwhile, we refer readers to [1] for additional experiment results on independent Bernoulli sequences.

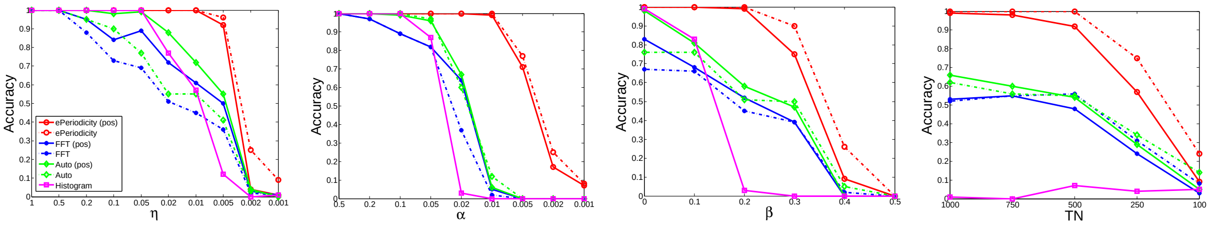
Step 1. We first fix a period T (*e.g.*, $T = 24$) and the transition matrices $\{P(t)\}$ (*e.g.*, see Figure 4) for the Markov chain model. Then, given the number of repetitions TN, the complete observation sequence SEQ_{std} is generated according to the model. Note that SEQ_{std} is a boolean sequence of length $T \times TN$, with values -1 and 1 indicating negative and positive observations, respectively.

Step 3 (Random sampling η). We sample the standard sequence with sampling rate η. For any element in SEQ_{std}, we set its value to 0 (*i.e.*, unknown) with probability $(1 - \eta)$.

Step 4 (Missing segments α). For any segment in sequence SEQ_{std}, we set all the elements in that segment to 0 (*i.e.*, unknown) with probability $(1 - \alpha)$.

Step 5 (Random noise β). For any remaining observation in SEQ_{std}, we reverse its original values (making -1 as 1 and 1 as -1) with probability β.

The input sequence SEQ has values -1, 0, and 1 indicating negative, unknown, and positive observations. In the case when negative samples are unavailable, all the -1 values are set to 0. Note that here we set negative observations as -1 and unknown ones as 0, which is different from the description in previous sections. The reason is that if the unknown entries are set as -1, in the presence of many missing



(a) Sampling rate η (b) Observed segements α (c) Noise ratio β (d) Repetitions TN

Fig. 12. Comparison results on synthetic data with various parameter settings.

entries, traditional methods such as Fourier transform will be dominated by missing entries instead of actual observations. The purpose of such adjustment is to facilitate traditional methods and it has no effect on our method.

5.2 Methods for Comparison

We will compare ePeriodicity with the following methods, which are frequently used to detect periods in boolean sequence [15]. The Matlab implementations of all the algorithms, as well as the synthetic data generation, are publicly available online.[3]

1. Fourier Transform (FFT): The frequency with the highest spectral power from Fourier transform via FFT is converted into time domain and output as the result.

2. Auto-correlation and Fourier Transform (Auto): We first compute the auto-correlation of the input sequence. Since the output of auto-correlation will have peaks at all the multiples of the true period, we further apply Fourier transform to it and report the period with the highest power.

3. Histogram and Fourier Transform (Histogram): We calculate the distances between any two positive observations and build a histogram of the distances over all the pairs. Then we apply Fourier transform to the histogram and report the period with the highest power.

We will use FFT(pos) and Auto(pos) to denote the methods FFT and Auto-correlation for cases without any negative observations. For Histogram, since it only considers the distances between positive observations, the results for cases with or without negative observations are exactly the same.

5.3 Performance on Synthetic Dataset

In this section, we test all the methods on synthetic data under various settings. We adopt the Markov chain model with period $T = 24$ and transition matrices shown in Figure 4 in this experiment. The default parameter setting is the following: $TN = 1000$, $\eta = 0.1$, $\alpha = 0.5$, and $\beta = 0.2$. In Figure 12, we report the performance of all the methods with one of these parameters varying while the others are fixed. For each parameter setting, we repeat the experiment for 100 times and report the accuracy, which is the percentage of correct period detections over 100 trials.

Performance w.r.t sampling rate η. To better study the effect of sampling rate, we set $\alpha = 1$ in this experiment. Figure 12(a) shows that ePeriodicity is significantly better than other methods in terms of handling data with low sampling rate. The accuracy of ePeriodicity remains 100% even when the sampling rate is as low as 0.01. The accuracies of other methods start to decrease when sampling rate is lower than 0.5. Also note that Auto is slightly better than FFT because auto-correlation essentially generates a smoothed version of the categorical data for Fourier transform. In addition, it is interesting to see that FFT and Auto performs better in the case without negative observations.

Performance w.r.t ratio of observed segments α. In this set of experiments, sampling rate η is set to 1 to better study the effect of α. Figure 12(b) depicts the performance of the methods. ePeriodicity again performs much better than other methods. ePeriodicity is almost perfect even when $\alpha = 0.01$. And when all other methods fail at $\alpha = 0.005$, ePeriodicity still achieves above 70% accuracy.

Performance w.r.t noise ratio β. In Figure 12(c), we show the performance of the methods w.r.t different noise ratios. Histogram is very sensitive to random noises since it considers the distances between any two positive observations. ePeriodicity is still the most robust one among all. For example, with $\beta = 0.3$, ePeriodicity achieves accuracy as high as 90%.

Performance w.r.t number of repetitions TN. Figure 12(d) shows the accuracies as a function of TN. As expected, the accuracies decrease as TN becomes smaller for all the methods, but ePeriodicity again significantly outperforms the others.

Comparison with Lomb-Scargle method. Lomb-Scargle periodogram (Lomb) [8], [9] was introduced as a variation of Fourier transform to detect periods in *unevenly* sampled data. The method takes the timestamps with observations and their corresponding values as input. It does not work for the positive-sample-only case, because all the input values will be the same hence no period can be detected. The reason we do not compare with this method systematically is that the method performs poorly on the binary data and it is very slow. Here, we run it on a smaller dataset by setting $TN = 100$. We can see from Table 2 that, when $\eta = 0.5$ or $\alpha = 0.5$, ePeriodicity and FFT perform well

[3] http://faculty.ist.psu.edu/jessieli/Site/research.html

whereas the accuracy of Lomb is already approaching 0. As pointed out in [16], Lomb does not work well on bi-modal periodic signals and sinusoidal signals with non-Gaussian noises, hence not suitable for our purpose.

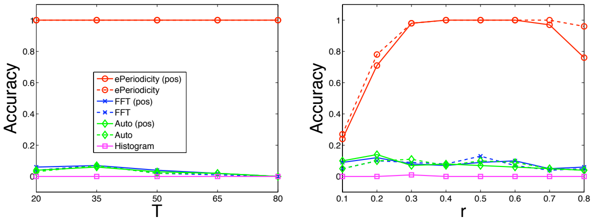
TABLE 2
Comparison with Lomb-Scargle method.

Parameter	Accuracy		
	ePeriodicity	FFT	Lomb
$\eta = 0.5$	1	0.77	0.12
$\eta = 0.1$	1	0.53	0.12
$\alpha = 0.5$	1	0.95	0.01
$\alpha = 0.1$	0.97	0.25	0

5.4 Robustness to Arbitrary Periodic Behaviors

We further study the performance of all the methods on arbitrary periodic behaviors. In this experiment, instead of using any specific random process (*e.g.*, Markov chain) to generate the observation sequence, we directly specify the periodic behavior using a randomly generated boolean sequence SEG of length T. In particular, given the period T and the ratio of 1's in SEG as r, we generate SEG by setting each element to 1 with probability r. Then, the complete observation sequence SEQ_{std} is obtained by repeating SEG for TN times.

Sequences generated in this way will have positive observations scattered within a period, which will cause big problems for all the methods using Fourier transform, as evidenced in Figure 13. *This is because Fourier transform is very likely to have high spectral power at short periods if the input values alternate between 1 and -1 frequently.* We refer interested readers to Appendix C for more discussion about this issue. In Figure 13(a) we set $r = 0.4$ and show the results w.r.t period length T. In Figure 13(b), we fix $T = 24$ and show the results with varying r. As we can see, all the other methods fail miserably when the periodic behavior is randomly generated. In addition, when the ratio of positive observations is low, *i.e.* fewer observations, it is more difficult to detect the correct period in general.



(a) True period T (b) ratio of 1's in SEG

Fig. 13. Comparison results on randomly generated periodic behaviors.

5.5 A Case Study on Real Human Movements

In this section, we use the real GPS locations of a person who has tracking record for 492 days. We first pick one of his frequently visited locations and generate a boolean observation sequence by treating all the visits to this location as positive observations and visits to other locations as negative observations. We study the performance of the methods on this symbolized movement data at different sampling rates. In Figure 14, we compare the methods at two sampling rates, 1 hour and 20 minutes. As one can see in Figure 14(a), when overlaying this person's activity using the period of one day, most of the visits occur in time interval $[40 : 60]$ for sampling rate of 20 minutes, or equivalently, in interval $[15 : 20]$ when the time unit is 1 hour. On one hand, when sampling rate is 20 minutes, all the methods except FFT(pos) and Histogram successfully detect the period of 24 hours, as they all have the strongest peaks at 24 hours (so we take 24 hours as the true period). On the other hand, when the data is sampled at each hour only, all the other methods fail to report 24 hours as the strongest peak whereas ePeriodicity still succeeds. In fact, the success of ePeriodicity can be easily inferred from Figure 14(a), as one can see that lowering the sampling rate has little effect on the distribution graph of the overlaid sequence. We further show the periods reported by all the methods at various sampling rates in Table 3. ePeriodicity obviously outperforms the others in terms of tolerating low sampling rates.

TABLE 3
Periods reported by different methods.

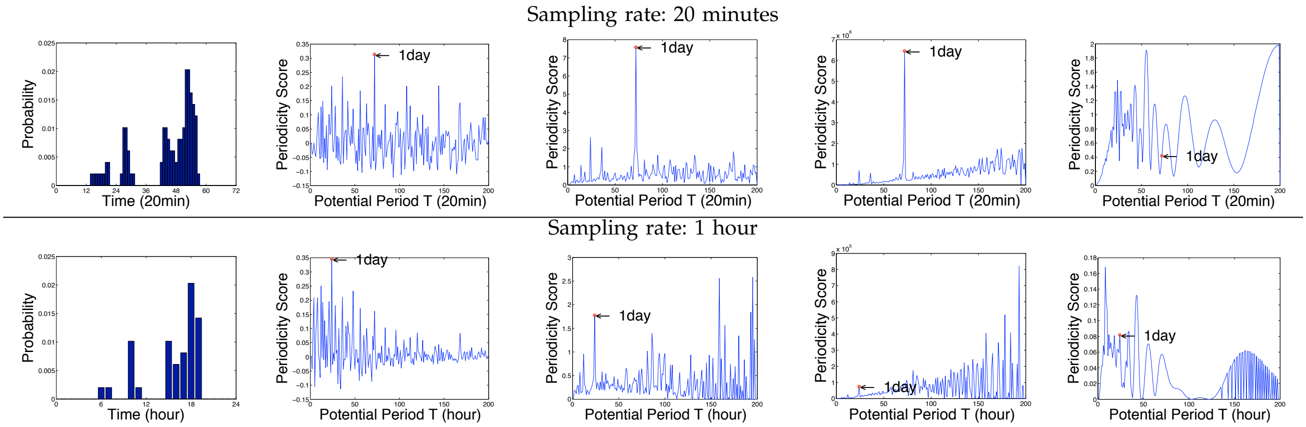
Method	Sampling rate			
	20min	1hour	2hour	4hour
ePeriodicity(pos)	24	24	24	8
ePeriodicity	24	24	24	8
FFT(pos)	9.3	9	8	8
FFT	24	195	372	372
Auto(pos)	24	9	42	8
Auto	24	193	372	780
Histogram	66.33	8	42	48

Next, in Figure 15, we use the binary sequence of the same person w.r.t. a different location and demonstrate the ability of our method in detecting multiple potential periods, especially those long ones. As we can see in Figure 15(a), this person clearly has weekly periodicity w.r.t this location. It is very likely that this location is his office which he only visits during weekdays. ePeriodicity correctly detects 7-day with the highest periodicity score and 1-day has second highest score. But all other methods are dominated by the short period of 1-day. Please note that, in the figures of other methods, 1-week point is not even on the peak. This shows the strength of our method at detecting both long and short periods.

5.6 Performance on Real Sensor Dataset

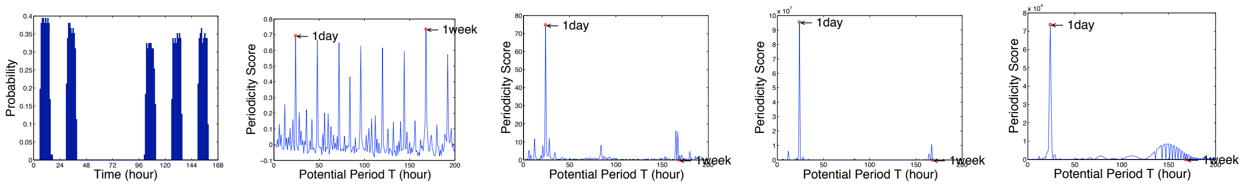
We now conduct systematic performance study using the public sensor event dataset provided by the CASAS smart home project.[4] The dataset consists of large-scale sensor data collected from a number of *smart apartments* located on the Washington State

[4] http://ailab.wsu.edu/casas/datasets/



(a) Event (b) ePeriodicity (c) FFT (d) Auto (e) Histogram

Fig. 14. Comparison of period detection methods on a person's movement data.



(a) Event (b) ePeriodicity (c) FFT (d) Auto (e) Histogram

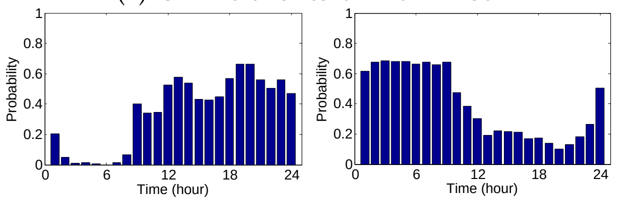
Fig. 15. Comparison of methods on detecting long period, *i.e.* one week (168 hours).

University campus. Each apartment is instrumented with various types of sensors, such as motion sensors on the ceiling, door sensors on cabinets and doors, and temperature sensors in each room. Sensor data are generated and stored while volunteer participants perform their daily activities in these apartments over the course of several months.

In this experiment, we use the motion sensor data generated from four different apartments, referred to as Tulum, Milan, Paris and Aruba, respectively. In Figure 16(a), we show some characteristics of each apartment. Note that the residents in this dataset exhibit a great deal of diversity in terms of age, marriage status, health condition (healthy or having dementia) and pets [17], [18].

Apartment	#Residents	#Sensors	Time Range
Tulum	2	31	9/25/09 – 3/28/10
Milan	1 + dog	28	10/16/09 – 1/6/10
Paris	2 + cat	29	3/31/10 – 9/6/10
Aruba	1	31	6/12/11 – 7/13/12

(a) Characteristics of the dataset



(b) Two representative periodic behaviors in Tulum

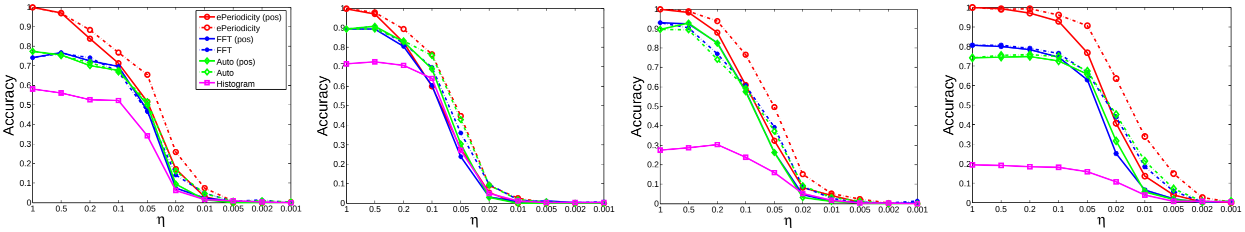
Fig. 16. CASAS smart home dataset.

In the dataset, each raw sensor event e_k is represented by the pair (sid_k, t_k), where sid_k is the sensor ID and t_k is the timestamp. From the raw data, we generate the complete observation sequence $\mathcal{X}_i, 1 \leq i \leq N$, for each of the N sensors by sampling all of its events at the 1-hour rate. That is, we set $x_i(t) = 1$ if at least one event of the i-th sensor occurred in the t-th hour, and $x_i(t) = -1$ otherwise. Note that since these sensors are densely positioned in the rooms (typically 1 meter apart), their events are often highly correlated. In Figure 16(b), we show the two most representative periodic behaviors in Tulum when overlaying the events of each sensor using the period of one day (*i.e.*, $T = 24$). As we can see, the sensor activities clearly exhibit a daily periodicity, with the first behavior corresponding to sensors that are more active during the daytime (*e.g.*, sensors in the living room), and the second behavior corresponding to sensors that are more active at night (*e.g.*, sensors in the bedroom). Sensors in other apartments also show similar daily behaviors, although the exact pattern may vary depending on the residents' life style. Therefore, we use $T = 24$ as the ground truth period in this experiment.

In Figure 17, we report the average period detection accuracy for all sensors in each apartment with varying sampling rate η. Recall that for each element in the original sequence, we set its value to 0 with probability $(1 - \eta)$. Here, for each η, we repeat the experiment for 10 times. As one can see, ePeriodicity outperforms the other methods on all four testbeds at all sampling rates. Most notably, ePeriodicity is the only method which achieves 100% accuracy when the sequences are complete, indicating that it is not sensitive to different periodic behaviors.

6 Related Work

Fourier transform and auto-correlation are the two most popular methods to detect periods [5]. However,



(a) Tulum (b) Milan (c) Paris (d) Aruba

Fig. 17. Comparison results on CASAS dataset w.r.t. sample rate η.

Fourier transform has known problem in detecting the periods from sparse data [15]. It also performs poorly on data with multiple non-consecutive occurrence in a period, as it tends to prefer short periods [6]. Auto-correlation offers accurate estimation for both short and long periods, but is more difficult to find the unique period due to the fact that the multiples of the true period will have the same score as the true period itself. In addition, both Fourier transform and auto-ccorelation require evenly sampled input data. Lomb-Scargle periodogram [8], [9] is proposed as a variation of Fourier transform to handle unevenly spaced data using least-squares fitting of sinusoidal curves. But it suffers the same problems as Fourier transform. In bioinformatics, several methods have been proposed to address the issue of unevenly spaced gene data [19], [20]. However, this issue is only one aspect of our problem whereas the low sampling rate and missing data problem have not been studied in these papers. An interesting previous work [15] has studied the problem of periodic pattern detection in sparse boolean sequences for gene data, where the ratio of the number of 1′s to 0′s is small. However, sparsity in our problem is a result of low sampling rate and missing data, and we do not make any assumption on the sparsity of original periodic patterns.

Studies on periodicity analysis in data mining and database area usually assume the input to be a sequence of symbols instead of real value time series, and most of them have been focused on the *efficiency* of the algorithms. Han *et al.* [21], [22] first developed algorithms for mining frequent partial periodic patterns, a special type of frequent patterns. Following this pioneering work, Yang *et al.* presented a series of work dealing with different variations of the periodic patterns, such as asynchronous periodic patterns [23], surprising periodic patterns [24], patterns with gap penalties [25], and high level patterns [26]. Meanwhile, methods have been proposed to mine partial periodic pattern with unknown periods [27], from incremental datasets [28], or in local segments [29]. Recently, using suffix tree as the underlying data structure, [30] proposed a unified framework to mine partial periodic patterns from subsection of the time-series despite various types of noise. However, all these works are based on the definition of *frequent* pattern with a strict min_sup threshold. They tend to output a large set of patterns, most of which are only slightly different from each other. Also, it is unclear how to extend these methods to deal with incomplete observations.

There are also papers addressing the automatic period detection problem in time-series [31], [32], [33], [7], [34]. Indyk *et al.* [33] develops an $O(n \log^2 n)$ time complexity algorithm using sketch approaches to find representative trend, where n is the length of the sequence. Berberidis *et al.* [31] detects the period candidates for each symbol using autocorrelation. Improved from [31], [33], Elfeky *et al.* [7] proposes a more efficient convolution method which considers multiple symbols together while detecting the period. In addition, in [34], a method based on dynamic time warping is proposed to handle insertion and deletion noises at the expense of higher time complexity. However, none of these methods is able to handle incomplete observation sequences. Our recent work [3] has studied probabilistic periodic behavior mining for moving objects. But it has been focused on processing spatiotemporal data, while period detection is still based on Fourier transform and auto-correlation.

In summary, none of previous studies can handle all the practical issues we mentioned in this paper, *i.e.*, the observations are incomplete, and the periodic behavior is oscillating and noisy.

7 Conclusion

In this paper, we address the important and challenging problem of period detection from incomplete observations. We first propose a probabilistic model for periodic behaviors. Then, we design a novel measure for periodicity and a practical algorithm ePeriodicity to detect periods in real scenarios. We give a rigorous proof of its validity for our probabilistic framework. Empirical studies show that our method is robust to imperfectly collected data and complicated periodic behaviors. A case study on real human movement data further demonstrates the effectiveness of our method.

While our approach is designed for binary sequences, one important extension is to handle symbolic or real-valued sequences. For example, GPS locations are often associated with the land types; sensors may not only detect the usage of a room but also report the temperature and humidity. Such data could also be sparse, noisy and unevenly sampled due to the limitations of devices. We consider this as interesting future work.

ACKNOWLEDGMENTS

The work was supported in part by the startup funding from Penn State University, Boeing company, NASA NRA-NNH10ZDA001N, NSF IIS-0905215 and IIS-1017362, and the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA). The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

REFERENCES

[1] Z. Li, J. Wang, and J. Han, "Mining event periodicity from incomplete observations," in *KDD*, 2012, pp. 444–452.

[2] M. C. González, C. A. Hidalgo, and A.-L. Barabás, "Understanding individual human mobility patterns," in *Nature*, 2008.

[3] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye, "Mining periodic behaviors for moving objects," in *KDD*, 2010.

[4] T. van Kasteren, A. K. Noulas, G. Englebienne, and B. J. A. Kröse, "Accurate activity recognition in a home setting," in *UbiComp*, 2008.

[5] M. B. Priestley, *Spectral Analysis and Time Series*. London: Academic Press, 1981.

[6] M. Vlachos, P. S. Yu, and V. Castelli, "On periodicity detection and structural periodic similarity," in *SDM*, 2005.

[7] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid, "Periodicity detection in time series databases," *IEEE Trans. Knowl. Data Eng.*, 2005.

[8] N. R. Lomb, "Least-squares frequency analysis of unequally spaced data," in *Astrophysics and Space Science*, 1976.

[9] J. D. Scargle, "Studies in astronomical time series analysis. ii - statistical aspects of spectral analysis of unevenly spaced data," in *Astrophysical Journal*, 1982.

[10] R. Durrett, *Probability: Theory and Examples*. Brooks/Cole Publishing Company, 1991.

[11] J. R. Blum and D. Hanson, "On the mean ergodic theorem for subsequences," *Bull. Amer. Math. Soc.*, vol. 66, pp. 308–311, 1960.

[12] J. Blum and B. Eisenbery, "The law of large numbers for subsequences of a stationary process," *The Annals of Probability*, vol. 3, no. 2, pp. 281–288, 1975.

[13] H. Ge, D.-Q. Jiang, and M. Qian, "A simple discrete model of brownian motors: Time-periodic markov chains," *Journal of Statistical Physics*, vol. 123, no. 4, pp. 831–859, 2006.

[14] R. C. Bradley, "Basic properties of strong mixing conditions. a survey and some open questions," *Probability Surveys*, vol. 2, pp. 107–144, 2005.

[15] I. Junier, J. Herisson, and F. Kepes, "Periodic pattern detection in sparse boolean sequences," in *Algorithms for Molecular Biology*, 2010.

[16] M. Schimmel, "Emphasizing difficulties in the detection of rhythms with lomb-scargle periodograms," in *Biological Rhythm Research*, 2001.

[17] D. J. Cook and M. Schmitter-Edgecombe, "Assessing the quality of activities in a smart environment," *Methods of Information in Medicine*, vol. 48, no. 5, pp. 480–485, 2009.

[18] D. J. Cook, "Learning setting-generalized activity models for smart spaces," *IEEE Intelligent Systems*, vol. 27, no. 1, pp. 32–38, 2012.

[19] E. F. Glynn, J. Chen, and A. R. Mushegian, "Detecting periodic patterns in unevenly spaced gene expression time series using lomb-scargle periodograms," in *Bioinformatics*, 2005.

[20] K.-C. Liang, X. Wang, and T.-H. Li, "Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data," in *BMC Bioinformatics*, 2009.

[21] J. Han, W. Gong, and Y. Yin, "Mining segment-wise periodic patterns in time-related databases," in *KDD*, 1998.

[22] J. Han, G. Dong, and Y. Yin, "Efficient mining of partial periodic patterns in time series database," in *ICDE*, 1999.

[23] J. Yang, W. Wang, and P. S. Yu, "Mining asynchronous periodic patterns in time series data," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 3, pp. 613–628, 2003.

[24] ——, "Infominer: mining surprising periodic patterns," in *KDD*, 2001.

[25] ——, "Infominer+: Mining partial periodic patterns with gap penalties," in *ICDM*, 2002.

[26] W. Wang, J. Yang, and P. S. Yu, "Meta-patterns: Revealing hidden periodic patterns," in *ICDM*, 2001.

[27] S. Ma and J. L. Hellerstein, "Mining partially periodic event patterns with unknown periods," in *ICDE*, 2001.

[28] W. G. Aref, M. G. Elfeky, and A. K. Elmagarmid, "Incremental, online, and merge mining of partial periodic patterns in time-series databases," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 3, pp. 332–342, 2004.

[29] C. Sheng, W. Hsu, and M.-L. Lee, "Mining dense periodic patterns in time series data," in *ICDE*, 2006, p. 115.

[30] F. Rasheed, M. Al-Shalalfa, and R. Alhajj, "Efficient periodicity mining in time series databases using suffix trees," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 79–94, 2011.

[31] C. Berberidis, W. G. Aref, M. J. Atallah, I. P. Vlahavas, and A. K. Elmagarmid, "Multiple and partial periodicity mining in time series databases," in *ECAI*, 2002.

[32] H. Cao, D. W. Cheung, and N. Mamoulis, "Discovering partial periodic patterns in discrete data sequences," in *PAKDD*, 2004, pp. 653–658.

[33] P. Indyk, N. Koudas, and S. Muthukrishnan, "Identifying representative trends in massive time series data sets using sketches," in *VLDB*, 2000.

[34] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid, "Warp: Time warping for periodicity detection," in *ICDM*, 2005.

Zhenhui Li is an assistant professor in the College of Information Sciences and Technology at the Pennsylvania State University. Her research interest lies in data mining and database systems. She has been working on designing effective and scalable methods for mining various kinds of complex patterns from spatial and temporal data. She received the B.Eng degree in Computer Science from Shanghai Jiao Tong University in 2007, and Ph.D degree in Computer Science from University of Illinois at Urbana-Champaign in 2012. She is a menber of IEEE.

Jingjing Wang received her B.E. in Electrical Engineering and a minor in Economics from Tsinghua Univ. China. She is currently a PhD candidate in the Department of Computer Science at the University of Illinois. Her research interests mainly includes data mining and information network analysis.

Jiawei Han is Abel Bliss Professor in the Department of Computer Science at the University of Illinois. He has been researching into data mining, information network analysis, and database systems, with over 600 publications. He served as the founding Editor-in-Chief of ACM Transactions on Knowledge Discovery from Data (TKDD). Jiawei has received ACM SIGKDD Innovation Award (2004), IEEE Computer Society Technical Achievement Award (2005), IEEE Computer Society W. Wallace McDowell Award (2009), and Daniel C. Drucker Eminent Faculty Award at UIUC (2011). He is a Fellow of ACM and a Fellow of IEEE. He is currently the Director of Information Network Academic Research Center (INARC) supported by the Network Science-Collaborative Technology Alliance (NS-CTA) program of U.S. Army Research Lab. His co-authored textbook "Data Mining: Concepts and Techniques" (Morgan Kaufmann) has been adopted worldwide.