Mining Event Periodicity from Incomplete Observations

Zhenhui Li, Jingjing Wang and Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA
zli28@uiuc.edu, jwang112@uiuc.edu, hanj@uiuc.edu

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 method 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.

Categories and Subject Descriptors

H.2.8 [**Data Management**]: Database Applications - Data Mining; H.4.0 [**Information Systems**]: General

General Terms

Algorithms

Keywords

Periodicity, incomplete observations

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 [4], GPS devices attached to animals help the scientists to study the animal movement patterns [7], and sensors allow us to monitor the usage of rooms and facilities [14].

Data collected from such tracking and sensor 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.

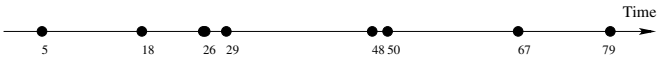


Figure 1: Incomplete observations.

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

To illustrate the 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 on 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 [11, 15, 1, 7], 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 [9, 12], they are still not applicable to the case with very low sampling rate.

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*

KDD'12, August 12–16, 2012, Beijing, China.

could also occur when the "in office" event is expected to be observed on a weekday but fails to happen.

In this paper, we take a completely different approach to the period detection problem and handle all the aforementioned difficulties occurring in data collection process and periodic behavior complexity in a unified framework. The basic idea of our method is illustrated in Example 1.

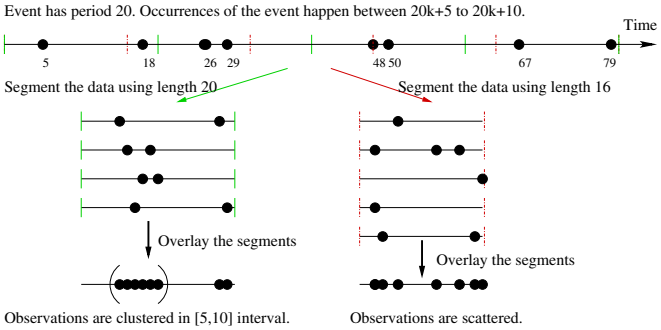


Figure 2: Illustration example of our method.

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 that most of the observations concentrate in time interval [5,10]. On the contrary, 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 to detect periods 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 and outlines the algorithm. We report our experimental results in Section 5, discuss 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 vectors* as defined below.

DEFINITION 3 (Periodic Distribution Vector). *We call any vector $\mathbf{p}^T = [p_0^T, \ldots, p_{T-1}^T] \in [0, 1]^T$ other than $\mathbf{0}^T$ and $\mathbf{1}^T$ a periodic distribution vector of length T. A binary sequence $\mathcal{X}$ is said to be generated according to $\mathbf{p}^T$ if $x(t)$ is independently distributed according to Bernoulli$(p^T{}_{\bmod (t,T)})$.*

Here we need to exclude the trivial cases where $\mathbf{p}^T = \mathbf{0}^T$ or $\mathbf{1}^T$. Also note that if we restrict the value of each p_i^T to $\{0, 1\}$ only, then the resulting $\mathcal{X}$ is *strictly* periodic according to Definition 1. We are now able to formulate our period detection problem as follows.

PROBLEM 1 (Event Period Detection). *Given a binary sequence $\mathcal{X}$ generated according to any periodic distribution vector $\mathbf{p}^{T_0}$, find T_0.*

EXAMPLE 2 (Running Example). *We will use a running example throughout the paper to illustrate our method. Assume that a person has a daily periodicity visiting his office during 10am-11am and 2pm-4pm. His observation sequence is generated from the periodic distribution vector with high probabilities at time interval [10:11] and [14:16] and low but nonzero probabilities at other timestamps, as shown in Figure 3.*

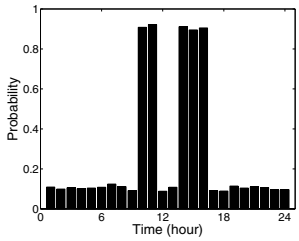


Figure 3: (Running Example) Periodic distribution vector of a event with daily periodicity $T_0 = 24$.

3. A PROBABILISTIC MODEL FOR PERIOD DETECTION

As we see in Example 1, 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 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:

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

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:

$$\mu_{\mathcal{X}}^+(I,T) = \frac{|S_I^+|}{|S^+|}, \quad \mu_{\mathcal{X}}^-(I,T) = \frac{|S_I^-|}{|S^-|}. \tag{1}$$

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

LEMMA 1. *Suppose a binary sequence $\mathcal{X} = \{x(t)\}_{t=0}^{n-1}$ is generated according to some periodic distribution vector $\mathbf{p}^T$ of length T, write $q_i^T = 1 - p_i^T$. Then $\forall I \in \mathcal{I}_T$,*

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

PROOF. The proof is a straightforward application of the Law of Large Numbers (LLN), and we only prove the first equation. With a slight abuse of notation we write $S_i = \{t : \mathcal{F}_T(t) = i\}$ and $S_i^+ = \{t \in S^+ : \mathcal{F}_T(t) = i\}$. Since $\{x(t) : t \in S_i\}$ are i.i.d. Bernoulli(p_i^T) random variables, by LLN we have

$$\lim_{n\to\infty} \frac{|S_i^+|}{n} = \lim_{n\to\infty} \frac{\sum_{t\in S_i} x(t)}{n} = \frac{p_i^T}{T},$$

where we use $\lim_{n\to\infty} \frac{|S_i|}{n} = \frac{1}{T}$ for the last equality. So,

$$\begin{aligned}\lim_{n\to\infty} \mu_{\mathcal{X}}^+(I,T) &= \lim_{n\to\infty} \frac{|S_I^+|/n}{|S^+|/n} = \lim_{n\to\infty} \frac{\sum_{i\in I} |S_i^+|/n}{\sum_{i=0}^{T-1} |S_i^+|/n} \\ &= \frac{\sum_{i\in I} p_i^T/T}{\sum_{i=0}^{T-1} p_i^T/T} = \frac{\sum_{i\in I} p_i^T}{\sum_{i=0}^{T-1} p_i^T}.\end{aligned}$$

□

Now we introduce our measure of periodicity based on Lemma 1. For any $I \in \mathcal{I}_T$, its discrepancy score is defined as:

$$\Delta_{\mathcal{X}}(I,T) = \mu_{\mathcal{X}}^+(I,T) - \mu_{\mathcal{X}}^-(I,T). \tag{2}$$

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

$$\gamma_{\mathcal{X}}(T) = \max_{I\in\mathcal{I}_T} \Delta(I,T). \tag{3}$$

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 periodic behavior model, $\gamma_{\mathcal{X}}(T)$ is indeed a desired measure of periodicity.

LEMMA 2. *If a binary sequence $\mathcal{X}$ is generated according to any periodic distribution vector $\mathbf{p}^{T_0}$ for some T_0, then*

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

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 $I^* = \{i \in [0, T_0 - 1] : c_i > 0\}$. So it suffices to show that for any $T \in \mathbb{Z}$ and $I \in \mathcal{I}_T$,

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

Observe now that for any (I,T),

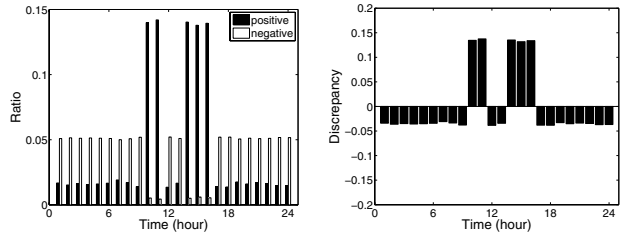
$$\begin{aligned}\lim_{n\to\infty} \mu_{\mathcal{X}}^+(I,T) &= \sum_{i\in I} \left(\frac{1}{T} \sum_{j=0}^{T_0-1} \frac{p_{\mathcal{F}_{T_0}(i+j\times T)}^{T_0}}{\sum_{k=0}^{T_0-1} p_k^{T_0}} \right), \\ \lim_{n\to\infty} \mu_{\mathcal{X}}^-(I,T) &= \sum_{i\in I} \left(\frac{1}{T} \sum_{j=0}^{T_0-1}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

Therefore we have

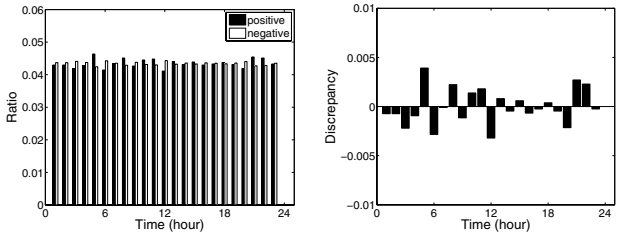
$$\begin{aligned}\lim_{n\to\infty} \Delta_{\mathcal{X}}(I,T) &= \frac{1}{T} \sum_{i\in I} \sum_{j=0}^{T_0-1} \left(\frac{p_{\mathcal{F}_{T_0}(i+j\times T)}^{T_0}}{\sum_{k=0}^{T_0-1} p_k^{T_0}} - \frac{q_{\mathcal{F}_{T_0}(i+j\times T)}^{T_0}}{\sum_{k=0}^{T_0-1} q_k^{T_0}} \right) \\ &= \frac{1}{T} \sum_{i\in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

where the third equality uses the definition of I^*. □

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 ($T = 23$)

Figure 4: (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 $T = 23$.

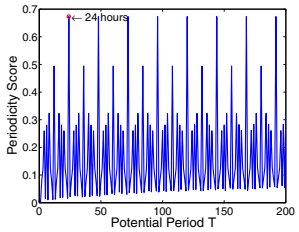


Figure 5: Periodicity scores of potential periods.

EXAMPLE 3 (RUNNING EXAMPLE (CONT.)). *When we overlay the sequence using potential period $T = 24$, Figure 4(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 4(c). Similarly, we see large discrepancy scores for T=24 (Figure 4(b)) whereas the discrepancy scores are very small for T=23 (Figure 4(d)). Therefore, we will have $\gamma_{\mathcal{X}}(24) > \gamma_{\mathcal{X}}(23)$. Figure 5 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 binary sequence $\mathcal{X}$ which indicates 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 4. *A sequence $\mathcal{X}$ is said to be generated according to $(\mathbf{p}^T, \mathbf{d})$ if*

$$x(t) = \begin{cases} \text{Bernoulli}(p^T_{\mathcal{F}_T(t)}) & \text{w.p. } d_t \\ -1 & \text{w.p. } 1 - d_t \end{cases} \quad (4)$$

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 3. *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 $(\mathbf{p}^T, \mathbf{d})$, write $q_i^T = 1 - p_i^T$. Then $\forall I \in \mathcal{I}_T$,*

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

PROOF. We only prove the first equation. Let $y(t)$ be a random variable distributed according to Bernoulli(d_t) and $z(t) = x(t)y(t)$. Then $\{z(t)\}_{t=0}^{n-1}$ are independent random variables which take value in $\{0, 1\}$, with mean $\mathbb{E}[z(t)]$ computed as follows:

$$\begin{aligned} \mathbb{E}[z(t)] &= \mathbf{P}(z(t) = 1) = \mathbf{P}(x(t) = 1, y(t) = 1) \\ &= \mathbf{P}(x(t) = 1 | y(t) = 1)\mathbf{P}(y(t) = 1) \\ &= p^T_{\mathcal{F}_T(t)}\mathbf{P}(y(t) = 1) = p^T_{\mathcal{F}_T(t)}\mathbb{E}[d_t] = p^T_{\mathcal{F}_T(t)}\rho_f. \end{aligned}$$

Define $S_i = \{t : \mathcal{F}_T(t) = i\}$ and $S_i^+ = \{t \in S^+ : \mathcal{F}_T(t) = i\}$, it is easy to see that $|S_i^+| = \sum_{t\in S_i} z(t)$. Using LLN we get

$$\lim_{n\to\infty}\frac{|S_i^+|}{n} = \lim_{n\to\infty}\frac{\sum_{t\in S_i} z(t)}{n} = \frac{p_i^T\rho_f}{T},$$

where we use $\lim_{n\to\infty}\frac{|S_i|}{n} = 1/T$ for the last equality. Therefore,

$$\begin{aligned} \lim_{n\to\infty}\mu^+_{\mathcal{X}}(I, T) &= \lim_{n\to\infty}\frac{|S_I^+|/n}{|S^+|/n} = \lim_{n\to\infty}\frac{\sum_{i\in I}|S_i^+|/n}{\sum_{i=0}^{T-1}|S_i^+|/n} \\ &= \frac{\sum_{i\in I}\frac{p_i^T\rho_f}{T}}{\sum_{i=0}^{T-1}\frac{p_i^T\rho_f}{T}} = \frac{\sum_{i\in I} p_i^T}{\sum_{i=0}^{T-1} p_i^T}. \end{aligned}$$

□

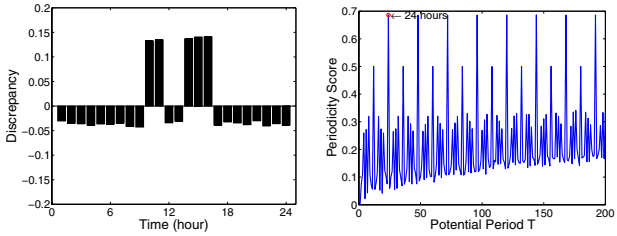
Since our periodicity measure only depends on $\mu^+_{\mathcal{X}}(I, T)$ and $\mu^-_{\mathcal{X}}(I, T)$, it is now straightforward to prove its validity under the random observation model. We summarize our main result as the following theorem.

THEOREM 1. *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 any $(\mathbf{p}^{T_0}, \mathbf{d})$ for some T_0, then*

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

The proof is exactly the same as that of Lemma 2 given the result of Lemma 3, 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



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

Figure 6: Period detection with unknown observations.

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.

EXAMPLE 4 (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 6(a) with Figure 4(b), we can see very similar discrepancy scores over time. Random sampling has little effect on our period detection method. As shown in Figure 6(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 $\mathcal{U} = \{u(t)\}_{t=0}^{n-1}$ where each $u(t)$ is an i.i.d. Bermoulli(p) random variable for some fixed $p > 0$. It is easy to see that for any T and $I \in \mathcal{I}_T$, we have

$$\lim_{n\to\infty} \mu^+_{\mathcal{U}}(I,T) = \frac{|I|}{T}. \tag{5}$$

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

$$\Delta^+_{\mathcal{X}}(I,T) = \mu^+_{\mathcal{X}}(I,T) - \frac{|I|}{T}, \tag{6}$$

and define the periodicity measure as:

$$\gamma^+_{\mathcal{X}}(T) = \max_{I\in\mathcal{I}_T} \Delta^+_{\mathcal{X}}(I,T). \tag{7}$$

In fact, with some slight modification to the proof of Lemma 2, we can show that it is a desired measure under our probabilistic model, resulting in the following theorem.

THEOREM 2. *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 any* $(\mathbf{p}^{T_0}, \mathbf{d})$ *for some* T_0, *then*

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

PROOF. Define $c_i^+ = \frac{p_i^{T_0}}{\sum_{k=0}^{T_0-1} p_k^{T_0}} - \frac{1}{T_0}$, it is easy to see that the value $\lim_{n\to\infty} \gamma^+_{\mathcal{X}}(T_0)$ is achieved by $I^* = \{i \in [0, T_0-1] : c_i^+ > 0\}$. So it suffices to show that for any $T \in \mathbb{Z}$ and $I \in \mathcal{I}_T$,

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

Observe now that for any (I,T),

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

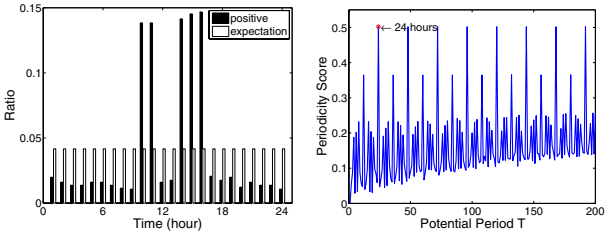
Therefore we have

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

where the fourth equality uses the definition of I^*. □

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 in experiment results, negative samples indeed provide additional information for period detection in finite length observation sequences.

EXAMPLE 5 (RUNNING EXAMPLE (CONT.)). *In this example we further marked all the negative samples in the sequence we used in Example 4 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 7(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 7(b).*



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

Figure 7: (Running Example) Period detection on sequences without negative samples.

4. ALGORITHM

In Section 3, we have introduced our periodicity measure for any potential period $T \in \mathbb{Z}$. Our period detection method simply computes the periodicity scores for every T and report the one with the highest score.

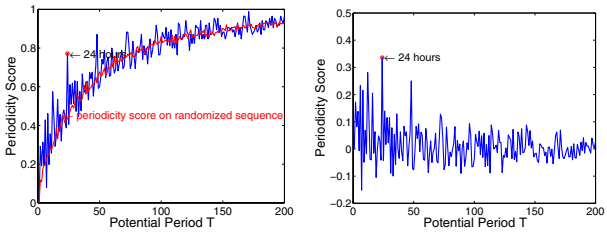
In this section, we first describe how to compute the periodicity score for a potential period and then discuss a practical issue when applying our method to finite length sequence. We will focus on the case with both positive and negative observations. The case without negative observations can be solved in the same way.

As we have seen in Section 3.1, the set of timestamps I^* that maximizes $\gamma_{\mathcal{X}}(T)$ can be expressed as

$$I^* = \{i \in [0, T_0 - 1] : c_i > 0\}, \tag{8}$$

where $c_i = \frac{p_i^{T_0}}{\sum_{k=0}^{T_0-1} p_k^{T_0}} - \frac{q_i^{T_0}}{\sum_{k=0}^{T_0-1} q_k^{T_0}}$. Therefore, to find I^*, it suffices to compute c_i for each $i \in [0, T_0 - 1]$ and select those ones with $c_i > 0$.

Time Complexity Analysis. For every potential period T, it takes $O(n)$ time to compute discrepancy score for a single timestamp (*i.e.*, c_i) and then $O(T)$ time to compute periodicity $\gamma_{\mathcal{X}}(T)$. Since potential period should be in range $[1, n]$, the time complexity of our method 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 method efficient in practice with time complexity as $O(n \times n_0)$.



(a) Periodicity scores (b) Normalized scores

Figure 8: Normalization of periodicity scores.

Now we want to point out a practical issue when applying our method on finite length sequence. 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 result in incorrect period detection result, as the case shown in Figure 8(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 $\mathcal{X}$.

To remedy this issue for finite length sequence, we use periodicity scores on *randomized* sequence 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 8(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 to avoid detecting duplicated periods (*i.e.*, multiples of the prime period).

5. EXPERIMENT

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

5.1 Synthetic Dataset Generation

In order to test the effectiveness of our method under various scenarios, we first use synthetic datasets generated according to a set of parameter. We take the following steps to generate a synthetic test sequence SEQ.

Step 1. We first fix a period T, for example, $T = 24$. The periodic segment SEG is a boolean sequence of length T, with values -1 and 1 indicating negative and positive observations, respectively. For simplicity of presentation, we write $SEG = [s_1 : t_1, s_2 : t_2, \ldots]$ where $[s_i, t_i]$ denote the i-th interval of SEG whose entries are all set to 1.

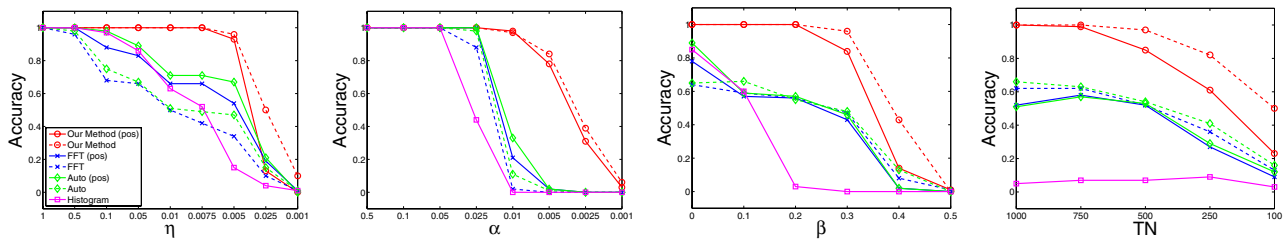
Step 2. Periodic segment SEG are repeated for TN times to generate the complete observation sequence, denoted as standard sequence SEQ_{std}. SEQ_{std} has length $T \times TN$.

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 standard segment SEQ_{std}, we set all the elements in that segment as 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 will be set to 0. Note that here we set negative observations as -1 and unknown ones as 0, which is different from the description in Section 2. The reason is that the unknown entries are set as -1, in the presence of many missing entries, traditional methods such as Fourier transform will be dominated by missing entries instead of actual observations.



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

Figure 9: Comparison results on synthetic data with various parameter settings.

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 our method with the following methods, which are frequently used to detect periods in boolean sequence [6].

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 Studies

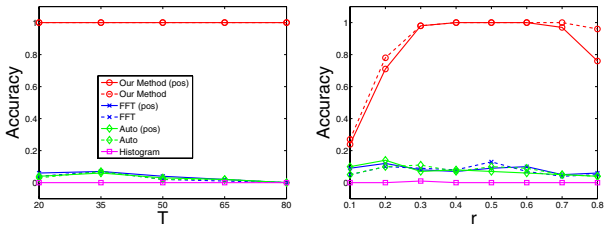
In this section, we test all the methods on synthetic data under various settings. The default parameter setting is the following: $T = 24$, $SEG = [9 : 10, 14 : 16]$. $TN = 1000$, $\eta = 0.1$, $\alpha = 0.5$, and $\beta = 0.2$. For each experiment, we report the performance of all the methods with 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 number of correct period detections over 100 trials. Results are shown in Figure 9.

Performance w.r.t sampling rate η. To better study the effect of sampling rate, we set $\alpha = 1$ in this experiment. Figure 9(a) shows that our method is significantly better than other methods in terms of handling data with low sampling rate. The accuracy of our method remains 100% even when the sampling rate is as low as 0.0075. 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 as 1 to better study the effect of α. Figure 9(b) depicts the performance of the methods. Our method again performs much better than other methods. Our method is almost perfect even when $\alpha = 0.025$. And when all other methods fail at $\alpha = 0.005$, our method still achieves 80% accuracy.

Performance w.r.t noise ratio β. In Figure 9(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. Our method is still the most robust one among all. For example, with $\beta = 0.3$, our method achieves accuracy as high as 80%.

Performance w.r.t number of repetitions TN. Figure 9(d) shows the accuracies as a function of TN. As expected, the accuracies decrease as TN becomes smaller for all the methods, but our method again significantly outperforms the other ones.

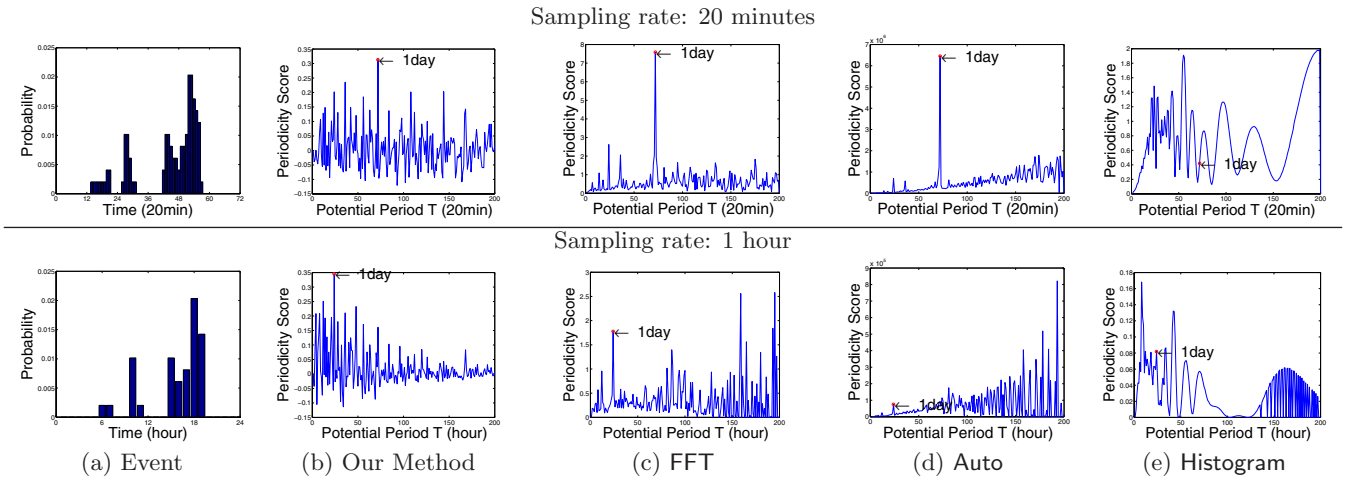


(a) True period T (b) Periodic segment SEG

Figure 10: Comparison results on randomly generated periodic behaviors.

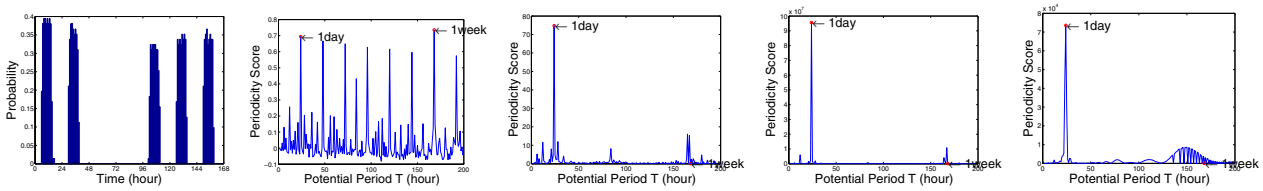
Performance w.r.t periodic behavior. We also study the performance of all the methods on randomly generated periodic behaviors. Given a period T and fix the ratio of 1's in a SEG as r, we generate SEG by setting each element to 1 with probability r. Sequences generated in this way will have positive observations scattered within a period, which will cause big problems for all the methods using Fourier transform, as evidenced in Figure 10. *This is because Fourier transform is very likely to have high spectral power at short periods if the input values alternate between 1 and 0 frequently.* In Figure 10(a) we set $r = 0.4$ and show the results w.r.t period length T. In Figure 10(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.

Comparison with Lomb-Scargle method. Lomb-Scargle periodogram (Lomb) [9, 12] was introduced as a variation of Fourier transform to detect periods in *unevenly* sampled data. The method takes the timestamps with observations



(a) Event (b) Our Method (c) FFT (d) Auto (e) Histogram

Table 1: Comparison of period detection methods on a person's movement data.



(a) Event (b) Our Method (c) FFT (d) Auto (e) Histogram

Figure 11: Comparion of methods on detecting long period, *i.e.* one week (168 hours).

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 $\alpha = 0.5$, our method and FFT perform well whereas the accuracy of Lomb is already approaching 0. As pointed out in [13], Lomb does not work well in bi-modal periodic signals and sinusoidal signals with non-Gaussian noises, hence not suitable for our purpose.

5.4 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 Table 1, we compare the methods at two sampling rates, 1 hour and 20 minutes. As one can see in Table 1(a), when overlaying this person's activity onto an 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 our method still succeeds. In fact, the success of our method can be easily inferred from Table 1(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. Our method obviously outperforms the others in terms of tolerating low sampling rates.

Parameter	Accuracy		
	Our Method	FFT	Lomb
$\eta = 0.5$	1	0.7	0.09
$\eta = 0.1$	1	0.52	0.10
$\alpha = 0.5$	1	1	0.01
$\alpha = 0.1$	0.99	0.35	0

Table 2: Comparison with Lomb-Scargle method.

Method	Sampling rate			
	20min	1hour	2hour	4hour
Our Method (pos)	24	24	24	8
Our Method	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

Table 3: Periods reported by different methods at various sampling rates.

Next, in Figure 11, we use the symbolized sequence of the same person at 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 11(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. Our method 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.

6. RELATED WORK

Fourier transform and auto-correlation are the two most popular methods to detect periods [11]. However, Fourier transform has known problem in detecting the periods from sparse data [6]. It also performs poorly on data with multiple non-consecutive occurrence in a period, as it tends to prefer short periods [15]. 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 [9, 12] 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 [3, 8]. 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 [6] 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 period detection 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 period detection algorithms [5, 1]. The presence of noises in the data has been considered in [10, 16, 2]. Our recent work [7] has studied probabilistic periodic behavior mining for moving objects. But it has been focused on dealing with 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 complicated 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 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 observation sequences with real values. For example, sensors may not only detect the usage of a room but also report the temperature and humidity, and such data could also be sparse, incomplete and unevenly sampled due to the limitations of sensors. We consider this as interesting future work.

8. ACKNOWLEDGMENTS

The work was supported in part by Boeing company, NASA NRA-NNH10ZDA001N, NSF IIS-0905215 and IIS-1017362, 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.

9. REFERENCES

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

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

[3] 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.

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

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

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

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

[8] 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.

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

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

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

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

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

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

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

[16] J. Yang, W. Wang, and P. S. Yu. Mining asynchronous periodic patterns in time series data. In *KDD*, 2000.