Spatio-Temporal Correlations in Cyber-Physical Systems: A Defense Against Data Availability Attacks *

Biplab Sikdar
National University of Singapore
Singapore 117583
bsikdar@nus.edu.sg

ABSTRACT

Many cyber-physical systems (CPS) use geographically distributed instrumentation to monitor and control the operation of the underlying system in real time. The availability of real-time measurements from deployed instrumentation is critical for the operation of CPS, which in turn makes them vulnerable to attacks that limit access to this stream of information. However, the impact of such attacks may be mitigated by exploiting the spatio-temporal correlation in the data streams that exist in many CPS, as shown in this paper. In addition to establishing the extent of spatio-temporal correlations in CPS data stream in the context of natural gas distribution systems, we propose and validate a methodology that exploits these correlations to accurately recreate data that might be lost or unavailable due to cyber attacks.

1. INTRODUCTION

Cyber-physical systems integrate computing technologies with physical components in order to enhance their ability to adapt, scale, and operate efficiently. CPS rely on networking technologies to provide connectivity between the various entities of the underlying physical system, and facilitating the collection, processing, and storage of data that allows computing algorithms to control and manage the CPS in real-time. CPS have applications in a number of sectors such as energy, transportation, healthcare, manufacturing, and building automation, to name a few.

Due to their importance in the economic and social aspects of any country, CPS associated with the basic services and infrastructure are an attractive target for adversaries that include nation states, corporations, organized crime, as well as independent actors. A broad class of attacks that may be launched on these systems is "availability" attacks that limits the real-time access to information generated in the system. Such attacks may be realized by a variety of mechanisms including denial of service, data modification, and hijacking of routers (and dropping packets). Under such attacks, it is critical for the CPS to be able to reconstruct any lost data stream with accuracy, in order to maintain its normal operation. This paper addresses this problem and develops a methodology for reconstructing lost CPS data by exploiting spatio-temporal correlations in data streams.

Security for CPS has received considerable attention in the recent past. Existing literature is primarily focused on detection of attacks on such systems with a large emphasis on smart grids. The detection of data modification or data injection attacks is one of the most commonly studied problems, in particular due to the possibility of economic or physical damage to the underlying cyber-physical system using this data. In order to maintain normal operation during such attacks, it is necessary to use data that reflects the true state of the system. Thus, it is important to develop data recovery algorithms that can be invoked during an attack to synthetically generate data streams for use by the system.

This *work-in-progress* paper addresses the problem of data recovery and regeneration in cyber-physical systems. This paper specifically considers the scenario of natural gas distribution systems where smart meters are installed at homes to monitor the gas consumption at fine grained time scales (compared to traditional manual meter readings at monthly intervals). We consider the attack scenario where the meters and/or network infrastructure have been compromised by the adversary, leading to a "loss" of the smart meter data. We use the term "loss" to include a wide range of scenario such as those where the adversary causes routers in the network to drop data packets, or modifies the data so that true values are no longer available (in these scenarios we assume that the attack can be detected). For scenarios where the true data values from the smart meters are not available, this paper develops a methodology that exploits the spatio-temporal correlations in the smart meter data from different users to recreate the lost data.

The rest of the paper is organized as follows. Section 2 presents the related work and Section 3 presents the system model. Section 4 presents an analysis of the spatio-temporal correlations in the meter data. Section 5 presents the methodology for recovering lost data and its evaluation. Finally, Section 6 presents the concluding remarks.

2. RELATED WORK

Existing work on security for CPS and industrial control

*This work supported in part by the National Research Foundation, Singapore and Singtel Corporation under grant R-252-004-628-281.

CPSS'17, April 02 2017, Abu Dhabi, United Arab Emirates

DOI: http://dx.doi.org/10.1145/3055186.3055200

systems has primarily focused on detecting various types of attacks. While a wide range of attacks has been considered, each existing individual piece of literature tends to focus on a specific attack in a given system. Existing literature on security for CPS shows a particular emphasis on electricity grids [1, 2] though other systems such as water networks have also been considered [3, 4, 5, 6]. Security related literature specific to gas distribution systems is limited [7].

Attacks on CPS that have been considered in literature include deception (i.e. data modification and data injection attacks, including the possibility modifying data packets as well as sensors), denial of service attacks, replay attacks, and packet drop attacks, to name a few. Detection of data modification attacks on Supervisory Control and Data Acquisition (SCADA) systems is considered in [8, 9, 10]. Methodologies for detecting replay attacks in control systems have been proposed in [11, 12]. The methodology is based on injecting a signal unknown to the attacker into the system. Denial of service attacks on networked control systems is considered in [13] which also proposes a countermeasure based on semi-definite programming. Finally, [14] considers the detection of packet drop attacks on smart grid data.

Resilience against attacks on CPS have been considered in the context of control systems where the objective is to ensure the stability of the system despite the attack [15]. In the context of smart grids, defense mechanisms against data modification attacks on smart grids have been proposed in [16, 17, 19] and are based on making a subset of the measurements safe against modification (which may be impractical in real life). The problem of estimating the state of a linear system in the presence of corrupted measurements has been considered in [18].

Techniques for recovery of missing data have been considered in both statistics as well as applied topics such as sensor networks. Statistical techniques for estimating missing data include maximum likelihood, expectation maximization, and multiple imputation [20]. These data imputation techniques typically have high space and/or time complexities. The most common data recovery technique that exploits spatial correlation between data sources is inverse-distance weighted averaging (IDWA) [23, 24]. IDWA assumes uniform correlation among adjacent sources of data, and estimates the missing values as a linear combination of the values at neighboring data sources, with weights based on the physical distance. An algorithm that exploits spatio-temporal correlations between sensor nodes for data imputation is proposed in [25]. The methodology is based on substituting the most common value from the neighboring sensors as the missing value. A genetic algorithm based technique for data recovery is proposed in [26]. Finally, compressive sensing based techniques are capable of recovering datasets based on a small number of data samples [21, 22]. However, compressive sensing based methods require the underlying data to have sparsity (i.e. low rank) and redundancy.

3. SYSTEM AND ATTACK MODEL

We consider the scenario of a natural gas distribution system as shown in Figure 1. A pressurized piping infrastructure is normally used to deliver gas to domestic consumers. The distribution network for natural gas begins at the national transmission system (NTS). The NTS usually also has reception terminals where producers supply gas into the system. Local transmission systems (LTS) connect to the NTS take-off stations or high pressure (HP) storage tanks and distribute the gas to consumers. Before reaching the end consumer, the gas may be stored in low pressure (LP) storage tanks and pass through a series of pipelines where the pressure is reduced to a level suitable for distribution to consumers. The pressure tiers are LTS (7-38 bar), intermediate pressure system (IPS) (2-7 bar), medium pressure system (MPS) (0.075-2 bar), and low pressure system (LPS) (below 0.075 bar). Finally, the gas is distributed to the consumers through a district governor. The gas intake in at the consumer premises passes through a meter (which for this paper is a smart gas meter). The smart meters are connected to the utility's data collection and processing facility through the Internet. The meters relay their readings to a gateway inside the consumer's home which forwards the messages to the utility using the Internet.

3.1 Threat Model

The threat model assumed in this paper is that the adversary is capable to hindering or disrupting the flow of smart gas meter data to the utility's data management and storage system. This may be achieved by compromising the smart meters, compromising routers or links connecting the meter to the utility's data collection facility, and denial-of-service attacks. Once the adversary successfully launches an attack, the real-time flow of measurement data from the meters is interrupted.

We assume that the adversary may use any arbitrary strategy for denying information availability. For example, in the simplest case, the adversary may drop packets with meter reading at random. In other variations, the adversary may drop "blocks" of packets (i.e. a number of successive packets). The adversary may also choose any combination of dropping strategies in order to make detection of an attack more difficult. As an alternative to dropping data packets, the adversary may modify the values reported in the data packets. The objective of this paper is not the detection of such attacks, but to *reconstruct the lost data*.

3.2 Meter Data

The smart gas meter data used for this paper was obtained from the Pecan Street project [28]. The source of the data are homes in the Mueller neighborhood of Austin, Texas, USA. The homes in this neighborhood are primarily newly constructed, and include single-family homes, apartments, and town homes. Itron Centron SR smart gas meters are deployed in these homes and these meters send their information to a gateway inside the home. The gateway uses the home's Internet connection to send the data to the meter data management system (MDMS) or the processing center. The gas meters have a reading frequency of 15 seconds, and measure the cumulative gas consumption. The meters register a reading (in terms of the cumulative consumption) when the last marginal 2 cubic foot (or higher) of natural gas passes through the meter.

The data used in this paper consists of smart gas meter data from 131 homes in Austin, Texas's Mueller neighborhood. A six month interval (October 1, 2015 to March 31, 2016) of data is considered. The data from each meter for this 183 day period was considered at intervals of one hour, leading to 4392 meter readings for each user. We use the marginal consumption in each hour as our data (ex-

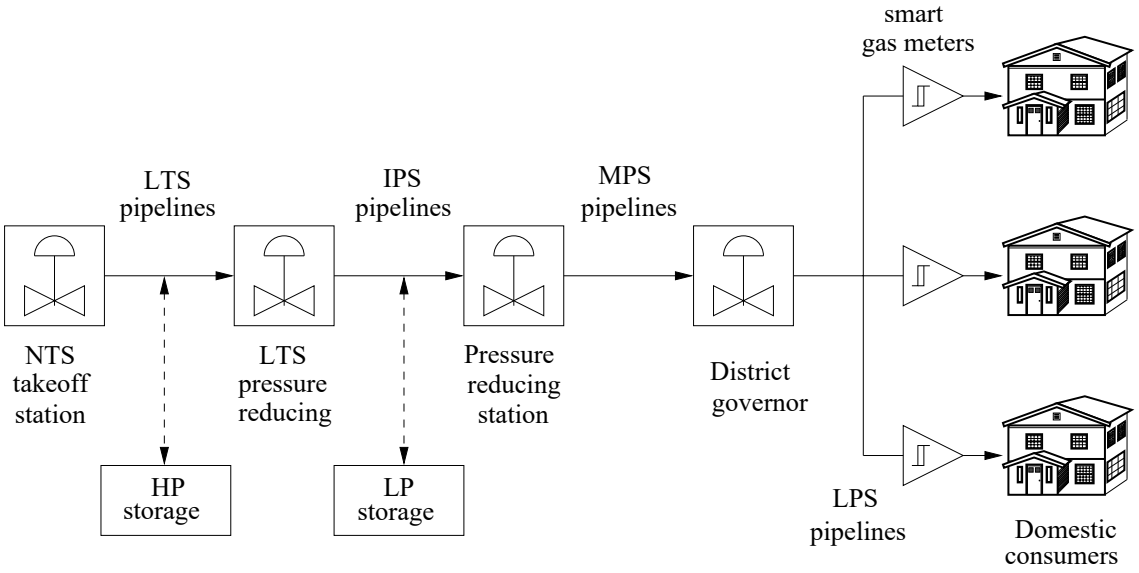

Figure 1: Natural gas transmission and distribution system network [27].

tracted from the cumulative consumption reported by the meters). We assume that the data obtained from the Pecan street project corresponds to scenarios where there is no attack, i.e., the data has not been modified by an adversary. We synthetically generate traces corresponding to attacks by deleting values from these traces, using the procedure outlined in Section 5.

4. SPATIO-TEMPORAL CORRELATION IN METER DATA

In this section we analyze the gas meter data to evaluate the spatial and temporal correlations in the data.

4.1 Spatial Correlations

The evaluation of spatial correlation seeks to identify the existence of similar behavior in the gas consumption patterns of different homes in the same geographical neighborhood. Let the data stream corresponding to the i-th user (we use the term home and user interchangeably in the rest of this paper) be denoted by $X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,n}\}$, where $x_{i,k}$ is the meter reading at the k-th time interval. The metric for evaluating the correlation between users i and j is

$$\rho_{ij} = \frac{\sum_{k=1}^{n}(x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{i,k} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{j,k} - \bar{x}_j)^2}} \quad (1)$$

where

$$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{i,k}. \quad (2)$$

We have $-1 \leq \rho_{ij} \leq 1$, and the pair-wise correlation coefficients for all users in the data set form the correlation matrix $\mathbf{R}$. Figure 2 shows the correlation coefficient (as a heat map) for all pairs of users in the data set. Figure 3 shows the correlation values for each user (each line is one row of the correlation matrix). As can be seen, there exist high levels of correlation between many users. Around 11% of the homes show a correlation value of greater than 0.5, while about 60% of the homes show a correlation value of greater than 0.25. Also, most users ($> 92\%$) show positive correlation values. A correlation coefficient of more than 0.15 is obtained in 70% of the user pairs in our dataset, and these correlation values have confidence levels of 99.9%.

This behavior in the pairwise correlation coefficients is intuitive. The primary use of gas in homes is for cooking and heating. Thus it is expected that a majority of the homes will demonstrate increased gas consumption during evening hours, and at night during winter. The spatial correlation is also affected by the fact the users that are in geographically close to each other are subjected to the same weather conditions and thus will react similarly to environmental factors.

4.2 Temporal Correlations

Next, we consider the temporal patterns and periodicity in each user's behavior. To evaluate the temporal behavior in the gas usage patterns of each user, we consider the auto-correlation function for each user at different lags. The auto-correlation function for user i at lag k is defined as

$$r_i(k) = \frac{\sum_{j=1}^{n}(x_{i,j} - \bar{x}_i)(x_{i,j+k} - \bar{x}_i)}{\sum_{j=1}^{n}(x_{i,j} - \bar{x}_i)^2}. \quad (3)$$

While the auto-correlation function may be used to ascertain the presence of randomness in data, it may also show evidence of periodicity in the data. Periodicity in the data is shown by peaks in the auto-correlation function at regular intervals in the lag.

Figure 4 shows the auto-correlation function for three users for $0 \leq k \leq 120$ (i.e. for lags of 0 to five days). As can be seen, users exhibit periodicity in their behavior, some more strongly than others. Also, while some users have a single dominant period, many others have multiple periods. As with the spatial correlation, the temporal correlation in an user's gas usage is not unexpected. Activities related to gas consumption such as cooking are usually done at specific times of the day and thus correlations are expected in the temporal patterns of gas consumption.

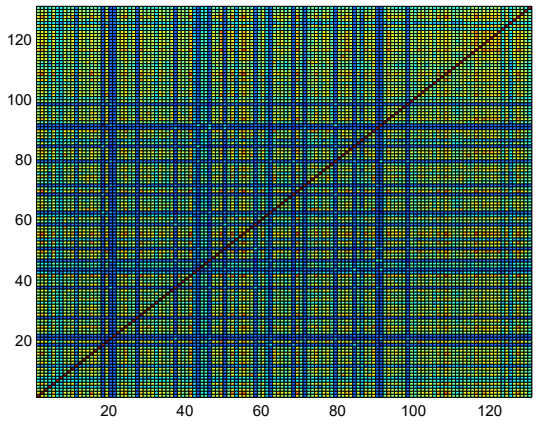

Figure 2: Heatmap of the correlation matrix.

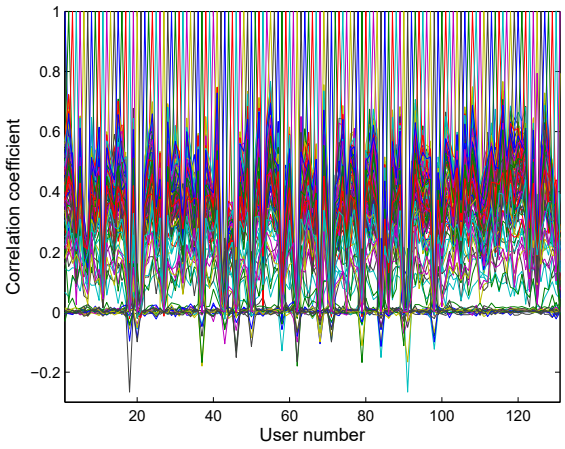

Figure 3: Pairwise correlation values for each user.

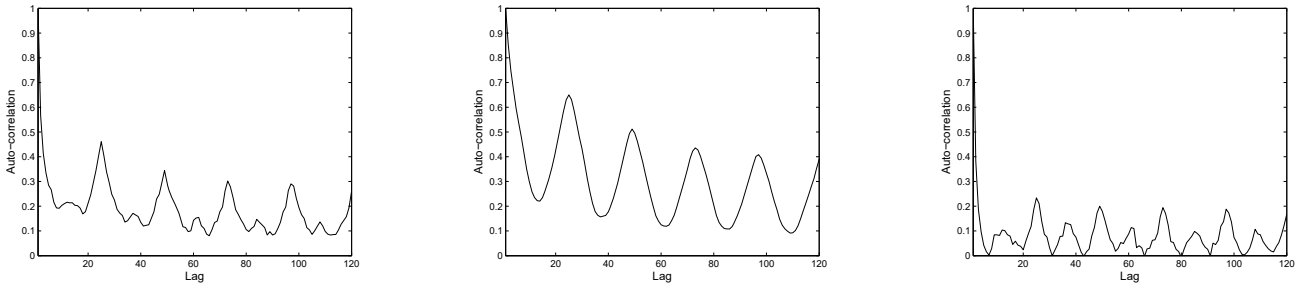

Figure 4: Auto-correlation values for three users for lags of 0 to 120 hours (0 to 5 days).

5. CORRELATION BASED DATA RECOVERY MECHANISM

As demonstrated in the previous section, the data from smart gas meters in individual homes show correlations across users as well as periodicity in its auto-correlation function. In this section we use these observations to develop simple but effective mechanisms for recovering missing data.

5.1 Spatial Correlation Based Data Recovery

Given the evidence in Figures 2 and 3 supporting the existence of significant correlations in the gas consumption of homes in the same geographical area, an intuitive approach to recovering lost data at a user is to use data from other users with which it shows high levels of correlation. The proposed data recovery algorithm is as follows. For each user, the algorithm maintains a window of the past w values and the window slides on the arrival of a new meter reading. If the data at any user (say user i) is found to be missing (either due to loss or the detection of data modification attack), the loss recovery mechanism is triggered. The past w data values from all users are used to compute the pairwise correlation coefficients (as defined in (1) with user i. The value of the missing data is then computed as a weighted average of the values of the other users, with weights decided based on the correlation coefficients. The missing data at user i at hour n is then calculated as

$$\hat{x}_{i,n} = \frac{1}{\sum_{k \in \mathcal{M}} \rho_{i,k}} \sum_{j \in \mathcal{M}} \rho_{i,j} x_{j,n} \tag{4}$$

where $\mathcal{M}$ is the set of other users whose data is used for the data recovery. The users in $\mathcal{M}$ may be chosen, for example, by including only those users whose correlation coefficients is greater than some threshold value (say 0.5). Alternatively, $\mathcal{M}$ may consist of the k (e.g. $k = 5$) users with the highest correlation coefficient with user i. Our test results show that the threshold based selection of users in the set $\mathcal{M}$ performs better in our scenario. However, such a process for constructing $\mathcal{M}$ fails for users who do not exhibit large enough correlation levels with other users.

5.2 Temporal Correlation Based Data Recovery

The temporal correlations in a user's data, as shown in Figure 4, show that there is periodicity in the data, and the there is significant correlation at short time lags. Based on this observation, intuitive data recovery mechanisms may be developed that use a meter's historical data to recover any lost data. In our temporal correlation based data recovery algorithm, a window of past w measurements is maintained for each user. On the arrival of each new data sample, the window slides and the auto-correlation values of the user

Algorithm 1 Data Replacement Using Spatio-Temporal Correlations

parameters: w, η, θ;
loop
  for at each time instant (hour) n **do**
    slide window for each user;
    update $r_i(k)$, $0 \leq k \leq 168$, $\forall i$;
    if data missing at user i **then**
      $\mathcal{M} = \phi$, $\mathcal{N} = \phi$;
      evaluate $\rho_{i,j}$, $\forall j, j \neq i$;
      if $(x_{j,n} \sim \text{missing}) \wedge (\rho_{i,j} \geq \eta)$ **then**
        $\mathcal{M} = \{j\} \cup \mathcal{M}$;
      end if
      if $r_i(x)$ is among θ largest values of $r_i(k)$, $0 \leq k \leq 168$ **then**
        $\mathcal{N} = \{x\} \cup \mathcal{N}$;
      end if
      $\hat{x}_{i,n} = \frac{1}{\sum_{j \in \mathcal{M}} \rho_{i,j} + \sum_{k \in \mathcal{N}} r_i(k)} \times \left[\sum_{j \in \mathcal{M}} \rho_{i,j} x_{j,n} + \sum_{k \in \mathcal{N}} r_i(k) x_{i,n-k} \right]$;
    end if
  end for
end loop when session is terminated

(as given in 3) is updated. We evaluate the auto-correlation values for lags of 0 to 168 (i.e. a maximum lag of seven days) to capture any weekly patterns in the gas consumption data, such as different consumer behavior patterns in weekdays and weekends. The missing data is then computed as

$$\hat{x}_{i,n} = \frac{1}{\sum_{k \in \mathcal{M}} r_i(k)} \sum_{j \in \mathcal{M}} r_i(j) x_{n-j} \tag{5}$$

where $\mathcal{M}$ is the set of lags for which previous data samples of user i are used for data recovery. The set of lags in $\mathcal{M}$ is chosen based on the value of the auto-correlation function. Similar to the spatial correlation case, a lag value may be included in $\mathcal{M}$ if the auto-correlation value at that lag exceeds a threshold value, or if the auto-correlation at that lag is one of the largest k values of the auto-correlation function for user i. Note that the formulation above requires $w \geq 168$. Also, for users whose auto-correlation function decays quickly with lag, constructing $\mathcal{M}$ with the lags that have the largest values ensures that a imputed value for the lost data is generated, albeit with some error.

5.3 Data Recovery using Spatio-Temporal Correlation

While the two data recovery methodologies described above work reasonably well, an approach that exploits both the spatial and temporal correlations may provide better results. This is particularly true for users that exhibit either lower values of spatial correlation with other users or an auto-correlation function than decays sharply. Thus, our final approach for data replacement uses both spatial and temporal correlations and the proposed methodology is given in Algorithm 1.

The proposed approach is based on keeping a window of past w meter readings for each user. At each instant of meter readings, the auto-correlation function for each user is updated, for a range of lags from 0 to 168. If the latest reading is missing at any user (say user i), we first evaluate this user's correlation coefficient with all other users. Then, we create the set $\mathcal{M}$ of users whose correlation coefficient exceeds a threshold η and whose latest meter reading is not missing. Similarly, we create a set $\mathcal{N}$ of lag values that correspond to the largest θ ($\theta \in \mathbb{N}^+$) values of the auto-correlation function $r_i(k), 0 \leq k \leq 168$. The imputed data value is then computed using a weighted average of the previous values of user i as well as the change in the values of the other users with whom user i exhibits high levels of correlated behavior.

Note that the purely spatial and purely temporal correlation based data recovery mechanisms proposed in Sections 5.1 and 5.2 respectively, are special cases of the methodology shown in Algorithm 1. The purely spatial correlation based data recovery mechanism is obtained by setting $\theta = 0$ in the algorithm, while the purely temporal correlation based data recovery mechanism is obtained by setting $\eta > 1$.

5.4 Performance Analysis

This section evaluates the accuracy of the proposed data imputation techniques. In addition to the spatial (SC), temporal (TC), and joint spatio-temporal (STC) correlation based techniques proposed in this paper, we consider two other data imputation methodologies from existing literature. The first benchmark is mean imputation (MI) where a missing value is replaced by the mean value of that meter from the recent past (a 24 hour period in our case). The second benchmark is k-nearest neighbor based data imputation (kNNI) [30, 31]. With kNNI, the k nearest neighbors of a meter with missing data are first selected (in terms of the Euclidean distance of their measurement vectors). The missing value is then estimated by taking the weighted (inversely proportional to the distance) average of the readings of these k neighbors. We use $k = 10$ in this paper based on empirical observations on our data. For the data imputation methodologies proposed in this paper, we use $w = 168$, $\eta = 0.9$ and $\theta = 1$ for our results. The values of η and θ were chosen empirically. For the SC method, the set $\mathcal{M}$ was constructed by using the threshold method (using $\eta = 0.9$). If a user did not have a correlation coefficient greater than η with any other user, its missing values were replaced with 0. For the TC method, the set $\mathcal{M}$ was constructed by considering the lags corresponding to the θ largest values of the auto-correlation function. We use $\theta = 1$ as it was empirically observed to give the best results.

We use the smart gas meter data from the Pecan street project as the input to the data imputation techniques. To simulate the impact of attack scenarios that result in the loss of data availability, we delete some of the data. We assume that the attacker randomly and independently drops data from each user as per a two-state Markov model. In the good state, no data is dropped and all data is dropped in the bad state. The average time spent in the bad state is 24 hours, and the state transition probabilities are calculated based on the fraction of the overall data that is dropped (we consider three cases: 10and 30% of the total data is dropped).

5.4.1 Performance Metrics

The performance of the proposed data imputation methodologies and the benchmark schemes is evaluated using three metrics. Let $X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,n}\}$ denote the original set of meter readings for user i, and let $\hat{X}_i = \{\hat{x}_{i,1}, \hat{x}_{i,2}, \cdots, \hat{x}_{i,n}\}$

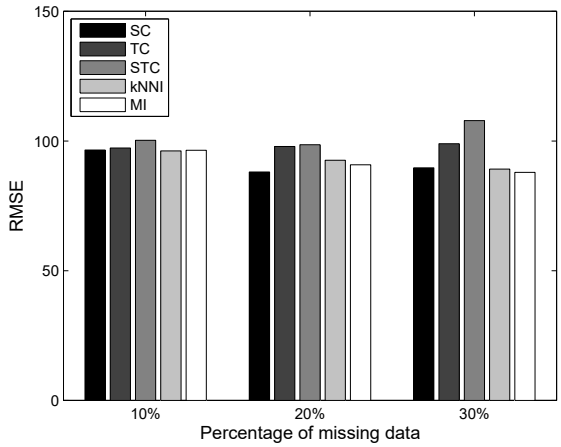

Figure 5: Comparison of the RMSE for the data recovery algorithms.

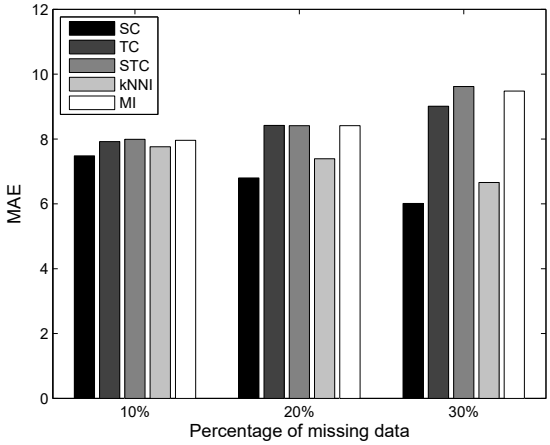

Figure 6: Comparison of the MAE for the data recovery algorithms.

be the corresponding data set where missing values have been replaced. The first metric is root mean square error (RMSE) which is a commonly used measure of the difference between estimated and real values. The RMSE of a data imputation algorithm (for user i) is defined as

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n}(x_{i,k} - \hat{x}_{i,k})^2}{n}}. \quad (6)$$

The second metric is mean absolute error (MAE) that measures the closeness of the estimated values to the real values. The MAE is given by

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|x_{i,k} - \hat{x}_{i,k}|. \quad (7)$$

The variation in the estimation error can be analyzed by jointly considering the MAE and the RMSE. While the RMSE is always greater than or equal to the MAE, the magnitude of their difference is directly proportional to the variance in the errors.

The third metric is the integral of absolute error (IAE), which is defined as

$$IAE = \int_0^t |x_i(t') - \hat{x}_i(t')|dt' \quad (8)$$

where $x_i(t')$ and $\hat{x}_i(t')$ denote the measured and estimated values for user i at time t', and t is the total time for which the data has been collected. The IAE metric is commonly used in control systems and in general, a larger IAE value indicated a poorer performance of the control algorithm.

5.4.2 Results

The metrics defined above are evaluated for each of the 131 users in the database. The results reported in this section are the averaged values for all users. Figures 5, 6 and 7 show the RMSE, MAE and IAE for the five data imputation techniques under consideration, for the three cases of data loss rates.

As can be seen, in general, the methodology based on purely spatial correlations gives the best results. Also, the methodologies based on the use of both the spatial and temporal correlations, as well as the one based on using the mean values tend to perform the worst. The mean value based methodology does not perform well since the mean value does not track the short term variations in the gas consumption of a user with sufficiently fine granularity. The poor performance of the joint use of temporal and spatial correlation based methodology is counter-intuitive. However, this can be explained by noting that the spatial correlation part of methodology (as well as the purely spatial correlation based methodology) uses a high threshold value ($\eta = 0.9$). Our experimental results show that only about 25% of the users have at least one neighbor with such a high correlation value and the accuracy in these cases is quite high. For the rest, the missing value is replaced with 0, and given that gas consumption is 0 in households for most hours of the day, the imputed data values with the purely spatial correlation based methodology have the highest accuracy. However, when this approach is combined with the temporal correlations, for the cases where none of the neighbors has a correlation that exceeds the threshold, the user's past data is used as a replacement. In many cases, this leads to a non-zero value for the imputed data, thereby contributing to errors. Among the benchmarks, kNNI performs well compared to the other methodologies, although it has higher computational complexity.

It is worth noting that the lag with the highest value of auto-correlation in our experiments is usually a lag of 1 (i.e. the last known value). Thus the TC method can be simplified to an algorithm that simply replaces any lost data for a user with its last known meter reading. Such a method has the advantage of simplicity over the other algorithms and works well for cases where the loss durations are short (and also due to the fact that gas consumption values are zero for frequent periods of time).

Finally, we note that the proposed methodologies in this paper use a linear function of the cross and auto correlation values when assigning the weights during data imputation. On the other hand, the kNNI based algorithm uses a non-linear function. Using a non-linear weight in the proposed methodologies may improve their accuracy and we leave this as future work.

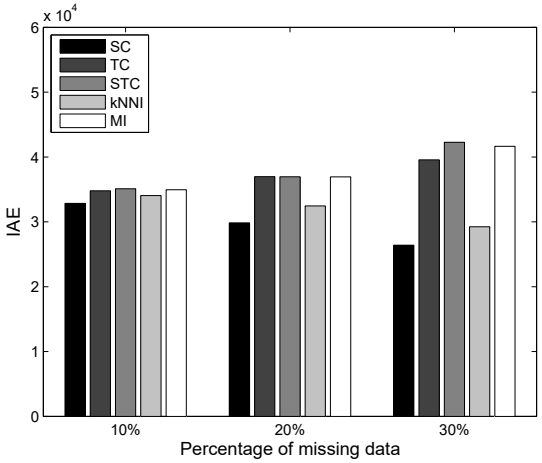

Figure 7: Comparison of the IAE for the data recovery algorithms.

6. CONCLUSIONS

The availability of real-time data from sensors, meters and other devices is critical to the operation of cyber-physical systems. Consequently, attacks that hamper the flow of these streams of information are an attractive strategy for malicious adversaries. To ensure the continued operation of the underlying CPS under such attacks, this paper proposed a methodology for the reconstruction of lost or modified data. Considering the specific case of smart gas meters, the paper first demonstrated the existence of significant spatio-temporal correlations in the meter readings. Then, we proposed methodologies that exploit these correlations to reconstruct any missing data. Our results indicate that the use of temporal correlations can provide good accuracy in the reconstructed values at low computational cost. This is a work-in-progress paper and we envision that it possible to further reduce the errors associated with our algorithms.

7. REFERENCES

[1] S. Sridhar and M. Govindarasu, "Model-based attack detection and mitigation for automatic generation control," *IEEE Transactions on Smart Grid,* vol. 5, no. 2, pp. 580-591, 2014.

[2] Y. Liu, P. Ning and M. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security,* vol. 14, no. 1, article 13, 2011.

[3] S. Amin, X. Litrico, S. Sastry and M. Bayen, "Stealthy deception attacks on water SCADA systems," *Proc. ACM International Conference on Hybrid Systems: Computation and Control,* pp. 161-170, Stockholm, Sweden, Apr. 2010.

[4] D. Eliades and M. Polycarpou, "A fault diagnosis and security framework for water systems," *IEEE Transactions on Control Systems Technology,* vol. 18, no. 6, pp. 1254âĂŞ1265, 2010.

[5] S. Amin, X. Litrico, S. Sastry and A. Bayen, "Cyber security of water SCADA systems; Part I: Analysis and experimentation of stealthy deception attacks," *IEEE Transactions on Control Systems Technology,* vol. 21, no. 5, pp. 1963-1970, 2013.

[6] S. Amin, X. Litrico, S. Sastry and A. Bayen, "Cyber security of water SCADA systems; Part II: Attack detection using enhanced hydrodynamic models," *IEEE Transactions on Control Systems Technology,* vol. 21, no. 5, pp. 1679-1693, 2013.

[7] D. Holstein and J. Diaz, "Cyber security management for utility operations," *Proc. HICSS,* pp. 241c-241c, Kauai, HI, Jan. 2006.

[8] A. Teixeira et al., "Cyber security analysis of state estimators in electric power systems," *Proc. IEEE CDC,* pp. 5991âĂŞ5998, Atlanta, GA, Dec. 2010,

[9] V. Do, L. Fillatre and I. Nikiforov, "A statistical method for detecting cyber/physical attacks on SCADA systems," *Proc. IEEE CCA,* pp. 364-369, 2014.

[10] S. Pal, B. Sikdar and J. Chow, "Detecting data integrity attacks on SCADA systems using limited PMUs," *Proc. IEEE SmartGridComm,* pp. 545-550, Sydney, Australia, Nov. 2016.

[11] Y. Mo and B. Sinopoli, "Secure control against replay attacks," *Proc. Allerton Conf. on Communications, Control and Computing,* pp. 911-918, Monticello, IL, Sep. 2010.

[12] M. Zhu and S. Martinez, "On the performance analysis of resilient networked control systems under replay attacks," *IEEE Transactions on Automatic Control,* vol. 59, no. 3, pp. 804-808, 2014.

[13] S. Amin, A. Cardenas, and S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," *Proc. Hybrid Systems: Computation and Control,* pp. 31-45, Apr. 2009.

[14] S. Pal, B. Sikdar and J. Chow, "Real-time Detection of Packet Drop Attacks on Synchrophasor data," *Proc. IEEE SmartGridComm,* Venice, Italy, November 2014.

[15] M. Zhu and S. Martinez, "Stackelberg-game analysis of correlated attacks in cyber-physical systems," *Prof. American Control Conference,* pp. 4063âĂŞ4068, San Francisco, CA, Jul. 2011.

[16] O. Kosut, L. Jia, R. Thomas and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid,* vol. 2, no. 4, pp.645-658, 2011.

[17] S. Bi and Y. Zhang, "Defending mechanisms against false-data injection attacks in the power system state estimation," *in Proc. IEEE GLOBECOM Workshops,* pp. 1162âĂŞ1167, Houston, TX, Dec. 2011

[18] F. Hamza, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," *Proc. Allerton Conference on Communications, Control and Computing,* pp. 337-344, Urbana Champaign, IL, Sep. 2011.

[19] T. Kim and H. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Transactions on Smart Grid,* vol. 2, no. 2, pp. 326-333, 2011.

[20] R. Little and D. Rubin, *Statistical Analysis with Missing Data, 2nd Ed.,* Wiley-Interscience, New York, 2002.

[21] E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. on Information Theory,* vol. 52, no. 12, pp. 5406-5425, December 2006.

[22] D. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory,* vol. 52, no. 4, pp. 1289-1306, April 2006.

[23] P. Longley, M. Goodchild, D. Maguire and D. Rhind, *Geographic Information Systems and Science,* Wiley: New York, NY, 2005.

[24] G. Lu and D. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Computational Geosciences,* vol. 34, pp. 1044-1055, 2008.

[25] D. Guo, X. Qu, L. Huang and Y. Yao, "Sparsity-based spatial interpolation in wireless sensor networks," *Sensors,* vol. 11, no. 3, pp. 2385âĂŞ2407, 2011.

[26] A. Azadeh, S. Asadzadeh, R. Marandi, S. Shirkouhi, G. Khoshjhou and S. Talebi, "Optimal estimation of missing values in randomized complete block design by genetic algorithm," *Knowledge-Based Systems,* vol. 37, pp. 37âĂŞ47, 2013.

[27] G. Nasr and N. Connor, *Natural Gas Engineering and Safety Challenges: Downstream Process, Analysis, Utilization and Safety,* Springer, 2014.

[28] Pecan Street Inc., `https://dataport.pecanstreet.org/`, Last accessed: 30 Dec., 2016.

[29] D. Rubin, "Inference and missing data," *Biometrika,* vol. 61, pp. 581âĂŞ592, 1976.

[30] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. on Information Theory,* vol. 13, no. 1, pp. 21-27, January 1967.

[31] E. Kocaguneli, T. Menzies and J. Keung, "On the Value of Ensemble Effort Estimation," *IEEE Trans. Software Engineering,* vol. 38, no. 6, pp. 1403-1416, Nov./Dec. 2012.