

# A Hybrid Method for Non-Technical Loss Detection in Smart Distribution Grids

George M. Messinis<sup>1b</sup>, Alexandros E. Rigas, *Student Member, IEEE*, and Nikos D. Hatziargyriou<sup>1b</sup>, *Fellow, IEEE*

**Abstract**—Non technical losses (NTLs) are a major problem faced by DSOs all over the world. Traditional meter inspection techniques are not effective in detecting NTL, while the wide adoption of smart meters and smart grid technologies provide new opportunities for solving the problem. The proposed NTL detection system utilizes voltage sensitivity analysis, power system optimization, and support vector machines (SVMs) for correctly detecting NTLs in the distribution grid under various conditions. Breakout detection is used for extracting features from consumption time series and training an SVM classifier. Generalized least squares is used for estimating the network voltage self-sensitivities. Finally, the NTL detection problem is formed as a non-linear non-convex optimization problem solved with semi-definite programming relaxation. The three modules can be operated either autonomously or combined. The proposed NTL detection system is demonstrated under different scenarios to test its effectiveness.

**Index Terms**—Electricity theft, fraud detection, non-technical losses, SVM, voltage sensitivity.

## NOMENCLATURE

### Parameters

$a_t$	The percentage of energy stolen over actual energy consumed at time $t$ .
$b$	The slope (%/day) of the “Smart” attack indicating how fast attack intensity increases
$C$	Number of monitored nodes
$C^{Insp}$	Cost of meter inspection
$C^{SVM}$	SVM cost parameter
$E$	Consumption time series
$E_n$	Root node voltage
$F$	Consumption time series length
$f$	Percentage of malicious consumers
$Insp$	Number of inspections
$N$	Number of distribution network nodes
$PF_{ref}$	Reference power factor
$p$	Maximum power factor deviation
$P_k^G, Q_k^G$	Active, reactive power generation on bus $k$

$P_k^D, Q_k^D$	Active, reactive power demand measurement on bus $k$
$q$	Charging price for stolen energy
$r$	Rate of days with fraud to total number of days
$R^{TP}$	Reward for true positive (TP) case
$R_{hk}, X_{hk}$	Line resistance, reactance between buses $h, k$
$T_{start}$	The time a malicious behavior is initiated
$T^{FRAUD}$	The set of days with fraud
$T_{max}$	The time the maximum attack intensity value occurs
$v_k^{meter}$	Voltage magnitude as measured on bus $k$
$V_k$	Complex voltage at bus $k$
$Y_{kj}^*$	Complex conjugate of the $kj$ element of the network admittance matrix $Y$
$\gamma^{SVM}$	SVM gamma parameter
$\Delta P, \Delta Q$	Active, reactive power variation matrix
$\Delta E$	Voltage variation matrix
$\Delta E_i^{t_j}$	Voltage magnitude variation at node $i$ between two consecutive time steps indicated as $t_j$ .
$\Delta P_i^{t_j}$	Active power variation at node $i$ between two consecutive time steps indicated as $t_j$
$\Delta Q_i^{t_j}$	Reactive power variation at node $i$ between two consecutive time steps indicated as $t_j$
$\varepsilon$	Voltage deviation tolerance
$\omega$	Voltage measurement errors.

### Variables

$E_i^{stolen}$	Energy stolen by consumer $i$ until being detected
$I$	Total compensation of the NTL detection system
$P_k, Q_k$	Active, reactive power mismatch for bus $k$
$P_k^{fraud}$	Amount of fraud for bus $k$
$P_{Grid}^{Loss}$	Network active power losses (including NTL)
$Q_k^{fraud}$	Reactive power fraud for bus $k$
$S_P, S_Q$	Active, reactive power to voltage sensitivity matrix
$S_{P_i,w}^{err}$	The weight output $w_i^{Sens}$ of the SENS module per window $w$ and consumer $i$
$S_{P_i,w}^{scal}$	Calculated self-sensitivity term of the node consumer $i$ is connected to
$S_{P_i}^{real}$	Self-sensitivity term of the node consumer $i$ is connected to, extracted from network topology
$w_i^{NTL-MIN}$	NTL-MIN module weight output for consumer $i$
$w_i^{NTL-SYS}$	NTL detection system weight output for consumer $i$
$w_i^{SENS}$	SENS module weight output for consumer $i$
$w_i^{SVM}$	SVM module weight output for consumer $i$

Manuscript received April 12, 2018; revised September 3, 2018 and December 17, 2018; accepted January 22, 2019. Date of publication January 30, 2019; date of current version October 30, 2019. Paper no. TSG-00563-2018. (*Corresponding author: George M. Messinis.*)

The authors are with the Electric Power division, School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece (e-mail: gmessinis@power.ece.ntua.gr; arigas@mail.ntua.gr; nh@power.ece.ntua.gr).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSG.2019.2896381

$\Delta W^{SENS}$	Difference between the minimum weight produced by the SENS module for malicious consumers and the maximum weight produced for benign ones
$\Delta \bar{E}$	Normalized change in mean feature
$\Delta \sigma$	Change in standard deviation feature
$\hat{\tau}$	Breakout detection time.

## I. INTRODUCTION

**N**ON-TECHNICAL losses (NTL) can have large impact on the optimal operation and management of distribution grids, increasing costs and reducing power quality. NTL is defined as the sum of unaccounted energy resulting from electricity theft, meter malfunction, measurement errors etc. and is calculated as the difference between the energy injected in the network and the measured energy consumed including technical losses. Electricity theft is the part of NTL defined as the sum of unaccounted energy due to malicious measurement manipulation by consumers or other parties. Since the major component of NTL is electricity theft (fraud), the terms NTL, electricity theft and fraud are interchangeably used by the authors throughout the manuscript, while mainly focusing on electricity theft. Fraud in the electricity domain in some cases reaches even 50% of the estimated consumption. Overall, it has been reported that the cost of electricity theft worldwide reaches more than \$25 billion every year [1].

NTL detection techniques can be classified as data oriented, network oriented and hybrids [2]. All methods require energy consumption measurements. Network oriented methods further utilize network data, such as topology and measurements from RTUs (Remote Terminal Unit) and observer meters (meters on the secondary side of the MV/LV transformer). Data oriented methods utilize only consumer related data like technical characteristics, personal/spatial, social or financial information. Hybrid methods are combinations of the previous two.

Data oriented methods include supervised and unsupervised learning techniques. Support Vector Machines (SVM) are most frequently used for supervised electricity theft detection [3]. One of the earliest works suggests the use of the Extreme Learning Machine (ELM) [4] while Optimum Path Forest was proposed in [5]. Unsupervised methods do not require labeled samples, but their performance is inferior [6].

Network oriented methods use three different concepts: load flows, state estimation and sensor placement [2]. The load flow concept includes checking energy balances and performing power flow studies for detecting NTL. Tariq and Poor [7] use Stochastic Petri Nets and Singular Value Decomposition for detecting NTLs in microgrids by estimating the line resistance connected to each consumer. Voltage sensitivity analysis is utilized in [8] and [9], where the results (voltages) of power flows are compared with smart meter measurements. Both approaches require the installation of an observer meter and assume that large voltage errors indicate possible fraud. State estimation approaches have been applied in [10] where a distributed solution of the Kalman filter preserving privacy is proposed.

Hybrid methods share characteristics from both data and network oriented methods. SVM is used together with observer

meters for checking energy balance in [11], while in [12] SVM is used with Decision Trees and observer meters. A combination of MV state estimation and ANOVA (Analysis of Variance) was proposed in [13]. A distribution state estimator is implemented and anomalous consumption at LV transformer level is detected with the normalized residual test. The authors then use ANOVA to enhance detection by comparing the results with previously validated (i.e., without NTL) usage baselines. Guo *et al.* [14] utilize real time measurements from feeder remote terminal units as well as clustering techniques and SVM in order to detect smart meter cyber tampering. Finally, in [15] state estimation, multivariate control charts and the A\* path search algorithm are combined for detecting frauds.

This paper presents a novel hybrid system for detecting NTLs consisting of three modules based on different principles. This ensures that a large variety of frauds can be detected. The first module (entitled SVM module) utilizes breakout detection and other features to detect frauds in yearly consumption time series by training an SVM. The second module (entitled SENS module) calculates voltage self-sensitivities from meter measurements and compares them to their theoretical values (extracted from network topology). Finally, the third module (entitled NTL-MIN module) solves an optimization problem where the objective is to minimize losses. The decision variables are consumers' active energy consumptions while voltage magnitude measurements are introduced as constraints.

The main contributions of this paper are:

- The calculation of voltage sensitivities from leaf node measurements and their use for detecting NTLs which has not been presented before to the authors' knowledge.
- The formulation of an SDP (semi-definite programming) optimization problem for detecting the location, extent and initiation time of NTL per consumer in distribution networks, which to the authors' knowledge has not appeared in literature. Apart from detection, such information can be used in order to accurately calculate compensations.
- The combination of the previous two concepts with an SVM approach which utilizes features extracted via the Twitter breakout detection algorithm, thus forming a hybrid NTL detection system capable of operating under various environments and detecting a variety of fraudulent behaviors.
- The proposed approach is tested on a real open data set, by transparently simulating different types of fraud while accounting for a number of parameters, which characterize consumers and the networks they are connected to.

## II. DATA ANALYTICS WITH SVM FOR NTL DETECTION

In this section, a module for detecting NTL solely based on smart meter active energy consumption measurements is presented. The module utilizes the SVM for classifying each consumer's annual consumption time series as malicious or not, after extracting relevant features. The objective of the module is to detect abrupt changes in consumption and indicate the time those changes take place. The output of the

TABLE I  
ELECTRICITY THEFT MODEL PARAMETERS

Parameter	Definition
<b>Introduction to Fraud (<math>T_{start}</math>)</b>	The day of year (1-365) a consumer starts committing fraud. Days before $T_{start}$ are free of frauds.
<b>Fraud Rate (<math>r</math>)</b>	The rate (%) of days with fraud to total number of days.
<b>Attack Intensity (<math>a_t</math>)</b>	The percentage of energy stolen over actual energy consumed at time $t$ .
<b>Ramp slope (<math>b</math>)</b>	The slope (%/day) of the “Smart” attack indicating how fast attack intensity increases.

module expresses the probability of fraud per consumer and the estimated fraud initiation time.

#### A. Modelling Non-Technical Losses

The data set used comes from the Irish Smart Energy Trial [16]. It includes half hourly active energy consumption of about 5000 residential and commercial consumers during 2009 and 2010. Since consumers volunteered to participate in this smart metering trial, it is reasonable to assume that the data set is free of NTLs. Fraud is thus simulated by modelling the behavior of a consumer’s attack intensity through time by (1)-(4). This model ensures that the NTL detection system will be tested under a wide range of fraud types and intensities. Fraud simulation parameters are defined in Table I. A set of data altering attacks can be defined based on these parameters. These attacks are used for validating the operation of all three modules. Let  $E_t$  be the original energy consumption of a consumer at time  $t$ . A time series with fraud is simulated by multiplying  $E_t$  with the factor  $(1 - a_t)$ , where attack intensity  $a_t$  is calculated by (1)-(4) for each type of attack:

$$BASE: a_t = \begin{cases} a_{max}, & t \geq T_{start} \\ 0, & t < T_{start} \end{cases} \quad (1)$$

$$INTERRUPT: a_t = \begin{cases} a_{max}, & t \in T^{FRAUD} \\ 0, & t \notin T^{FRAUD} \text{ AND } t < T_{start} \end{cases} \quad (2)$$

$$SMART: a_t = \begin{cases} a_{max}, & t \geq T_{max} \\ b(t - T_{start}), & T_{start} \leq t < T_{max} \\ 0, & t < T_{start} \end{cases} \quad (3)$$

$$COMBINED: a_t = \begin{cases} a_{max}, & t \geq T_{max} \text{ AND } t \in T^{FRAUD} \\ b(t - T_{start}), & T_{start} \leq t < T_{max} \text{ AND } t \in T^{FRAUD} \\ 0, & t < T_{start} \text{ AND } t \notin T^{FRAUD} \end{cases} \quad (4)$$

$a_{max}$  is the maximum value of  $a_t$ ,  $T_{max} > T_{start}$  is the time when the maximum attack intensity value occurs and  $b$  is the ramp slope parameter.  $T^{FRAUD}$  is the set of days with fraud, calculated by randomly sampling a number of  $r \cdot (365 - T_{start})$  days from the set  $[T_{start}, T_{start} + 1, \dots, 365]$ , where  $r$  is the fraud rate.

#### B. Extracting Features From Consumption Time Series

In 2014, the Breakout Detection package was introduced by Twitter, as an open source R package for detecting breakouts in cloud data [17]. Breakouts are defined as mean shifts (sudden jumps in the time series) or as ramp-ups (gradual increase/decrease from one steady state to another) within the

time series. This concept strongly applies to most cases of energy fraud, too. The relative algorithm is presented in [18] and is briefly described here.

For each time series the following problem is solved in order to specify the time of a possible breakout: Let  $Z_1, Z_2, \dots, Z_F$  be the observations of an energy consumption time series  $\mathbf{Z}$  of size  $F$ . Let  $\tau, \kappa$  be constants such that  $1 \leq \tau < \kappa \leq F$ . Sets  $A_\tau = \{Z_1, Z_2, \dots, Z_\tau\}$  and  $B_\tau(\kappa) = \{Z_{\tau+1}, Z_{\tau+2}, \dots, Z_\kappa\}$  are defined such that  $A_\tau, B_\tau(\kappa) \subseteq \mathbf{Z}$ . A breakout location  $\hat{\tau}$  within the time series is then estimated by solving the following problem:

$$(\hat{\tau}, \hat{\kappa}) = \underset{\tau, \kappa}{\operatorname{argmax}} \tilde{Q}(A_\tau, B_\tau(\kappa)) \quad (5)$$

where

$$\tilde{q} = \tilde{Q}(A_\tau, B_\tau(\kappa)) = \frac{mn}{m+n} (2m_{AB} - m_{AA} - m_{BB}) \quad (5a)$$

$$m_{AB} = \operatorname{median} \left\{ |x_i - y_j|^2 : 1 \leq i \leq n, 1 \leq j \leq m \right\}, \\ x \in A_\tau, y \in B_\tau(\kappa) \quad (5b)$$

$$m_{AA} = \operatorname{median} \left\{ |x_i - x_j|^2 : 1 \leq i < j \leq n \right\}, x \in A_\tau \quad (5c)$$

$$m_{BB} = \operatorname{median} \left\{ |y_i - y_j|^2 : 1 \leq i < j \leq m \right\}, y \in B_\tau(\kappa) \quad (5d)$$

$n, m$ : the cardinality of sets  $A_\tau, B_\tau(\kappa)$  respectively.

The maximization of  $\tilde{Q}$  indicates that a location  $\hat{\tau}$  is found for which the difference between the distributions of sets  $A_\tau$  and  $B_\tau(\kappa)$  is maximized. It can thus be inferred that apart from abrupt changes in average consumption, gradual changes can also be detected (thus enabling identification of “Smart” attacks). In addition, James *et al.* [18] proved that the use of the median in (5b)-(5d) makes the algorithm robust to noise, making it also possible to detect “Interrupt” attacks. The result of the algorithm is a possible breakout location  $\hat{\tau}$  for each consumer’s time series. The following features are then calculated:

1. Normalized change in mean: The difference between the mean of consumption  $E$  before and after breakout detection time  $\hat{\tau}$ . This difference is normalized by dividing with the mean yearly consumption.

$$\Delta \bar{E} = \frac{\operatorname{mean}(E[1, \hat{\tau}]) - \operatorname{mean}(E[\hat{\tau}, F])}{\operatorname{mean}(E)} \quad (6)$$

2. Change in standard deviation: The standard deviation of the consumption after a fraud event is expected to be lower than the one before.

$$\Delta \sigma = \frac{sd(E[1, \hat{\tau}]) - sd(E[\hat{\tau}, F])}{sd(E[1, \hat{\tau}])} \quad (7)$$

The following features are also extracted:

3. Im: Imaginary part of the normalized time series Fourier transformation
4. Slope: The annual consumption time series linear fit slope
5. Symmetry: Difference (percentage) of energy consumption between the last month and first month of the year.

Feature distributions for various values of the most important electricity theft parameters (attack intensity, fraud rate and ramp slope as defined in Table I) are presented in Fig. 1.

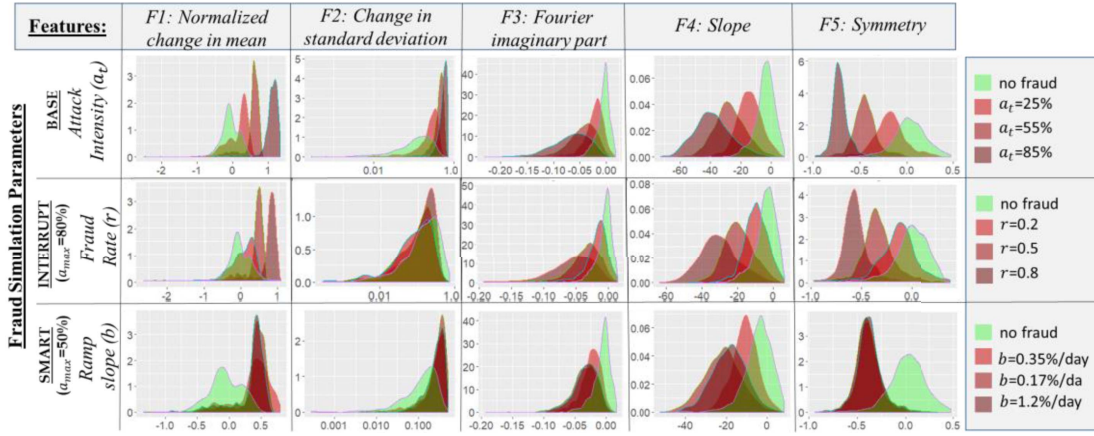


Fig. 1. Distribution of features as attack intensity, fraud rate and ramp slope increase (green distribution indicates cases without fraud and darker red distributions indicate increased values of the relative parameter).

The green density plots represent the values of the features when no fraud is simulated, while red density plots represent the case fraud is simulated on each consumer (three indicative values are chosen for each parameter, with darker red indicating a larger parameter value). The first row of Fig. 1 (“Base” attack) suggests that increasing the attack intensity impacts all features, making it easier to discriminate between benign and malicious samples. The second row of Fig. 1 (“Interrupt” attacks) suggests that decreasing the fraud rate hinders class separation. Finally, the ramp slope parameter (“Smart” attack) does not seem to affect significantly the discriminative nature of the chosen features (third row of Fig. 1).

### C. Supervised Learning With SVM

SVMs have been widely used for NTL detection, showing good performance under various data sets and types of fraud. Apart from its wide use in NTL detection, SVM has been successfully utilized before for a variety of power systems applications, including demand response [19] and intrusion detection [20]. The ‘LIBSVM’ [21] library and the relative R implementation (the ‘e1071’ R package [22]) are used in this work. The library also includes a function for tuning the SVM (calculating optimal  $C^{SVM}$  and  $\gamma^{SVM}$  parameters for the RBF kernel).

A total number of 3546 residential consumers’ yearly consumption data are available. Out of these, 1000 consumers are randomly selected for the validation of the NTL detection system (SVM, SENS and NTL-MIN module combination). The rest 2546 consumers are used for training and testing the SVM.

The SVM output is usually a binary classification of the input data. Although this output is useful for calculating performance metrics (through the confusion matrix), it does not fit the proposed concept, where weights are assigned to consumers. The class probability  $w_i^{SVM}$  of each consumer  $i$  is the final output of the module instead (together with the estimated time  $t_i^{fraud}$  a fraudulent behavior was initiated). This probability is calculated as described in [21] by passing the

SVM decision values to a logistic function whose parameters are optimized using the SVM training set.

The following metrics are used for evaluating the SVM:

- Area Under Curve (AUC): The area under the Receiver Operating Characteristic (ROC) curve, expressing the probability with which the classifier ranks a random malicious sample higher than a random benign sample.
- Accuracy: The sum of TP (True Positive) and TN (True Negative) samples to the total number of samples.
- Detection Day Error: The absolute deviation between the time fraud is actually initiated and the fraud initiation time calculated by the module.

## III. VOLTAGE SENSITIVITY ESTIMATION FOR DETECTING NTL

In this section the network voltage sensitivity terms are estimated in order to detect consumers with NTL. The manipulation of active power measurements is expected to impact the calculation of sensitivities.

### A. Voltage Sensitivity Analysis

The tree graph representation is chosen for the grid since most distribution networks are usually operated in a radial configuration. Nodes are divided in three categories: the root node (MV busbar), intermediate nodes (nodes of degree 2 or higher) where no generation or consumption exists and leaf nodes (nodes of degree 1) with consumption/generation.

The voltage sensitivity analysis proposed for MV distribution grids in [23] is extended to LV grids. PQ models are used for representing the loads and lines are modeled by the RL-direct sequence equivalent circuit. According to [23] a change in active and reactive power at any node will result in a voltage change that can be linearized as:

$$[\Delta E] = [S_P][\Delta P] + [S_Q][\Delta Q] \quad (8)$$

where  $S_P$ ,  $S_Q$  are the active and reactive power to voltage sensitivity matrices and  $\Delta P$ ,  $\Delta Q$ ,  $\Delta E$  are active power, reactive power and voltage variation matrices respectively.

The sensitivity terms  $S_{P_{ij}}, S_{Q_{ij}}$  for nodes  $i, j$  are calculated as:

$$\begin{aligned} S_{P_{ij}} &= \frac{\partial E_i}{\partial P_j} = -\frac{1}{E_n} \left[ \sum_{hk \in PT_{i,j}} R_{hk} \right], \\ S_{Q_{ij}} &= \frac{\partial E_i}{\partial Q_j} = -\frac{1}{E_n} \left[ \sum_{hk \in PT_{i,j}} X_{hk} \right] \end{aligned} \quad (9)$$

where  $E_n$  is the root node voltage.  $PT_{i,j}$  is the set of nodes contained in the path between the common ancestor of nodes  $i, j$  and the root node. Finally,  $R_{hk}, X_{hk}$  are the line resistance and reactance between buses  $h, k$  respectively.

Equation (9) reveals that voltage sensitivities are highly dependent on network structure. They can either be estimated directly from network topology or inferred from active/reactive power and voltage magnitude measurements, as will be demonstrated in Section IV-B. However, (8) implies that manipulating active/reactive power measurements without manipulating voltage measurements will result in sensitivities that do not reflect the physical characteristics of the network.

### B. Voltage Sensitivity Matrix Estimation

A method for calculating voltage sensitivities from measurements in distribution networks has been presented in [24] given availability of measurements on all nodes of the network. A modified approach is followed here, assuming availability of voltage magnitude, active and reactive power measurements on leaf nodes only, since this is where smart meters are placed in LV distribution grids. For each monitored bus  $i$  the following system of linear equations can be formed:

$$\begin{pmatrix} \Delta E_i^{t_1} \\ \vdots \\ \Delta E_i^{t_j} \\ \vdots \\ \Delta E_i^{t_m} \end{pmatrix} \approx \begin{pmatrix} \Delta P_1^{t_1} \cdots \Delta P_i^{t_1} \cdots \Delta P_C^{t_1} & \Delta Q_1^{t_1} \cdots \Delta Q_i^{t_1} \cdots \Delta Q_C^{t_1} \\ \vdots & \vdots \\ \Delta P_1^{t_j} \cdots \Delta P_i^{t_j} \cdots \Delta P_C^{t_j} & \Delta Q_1^{t_j} \cdots \Delta Q_i^{t_j} \cdots \Delta Q_C^{t_j} \\ \vdots & \vdots \\ \Delta P_1^{t_m} \cdots \Delta P_i^{t_m} \cdots \Delta P_C^{t_m} & \Delta Q_1^{t_m} \cdots \Delta Q_i^{t_m} \cdots \Delta Q_C^{t_m} \end{pmatrix} \times \begin{pmatrix} S_{P_{i1}} \\ \vdots \\ S_{P_{ii}} \\ \vdots \\ S_{P_{iC}} \\ S_{Q_{i1}} \\ \vdots \\ S_{Q_{ii}} \\ \vdots \\ S_{Q_{iC}} \end{pmatrix} \Rightarrow \Delta E_i = (\Delta PQ)(SPQ_i) + \omega \quad (10)$$

where:

$\Delta E_i^{t_j}$ : Voltage magnitude variation at node  $i$  between two consecutive time steps indicated as  $t_j$ .

$\Delta P_i^{t_j}, \Delta Q_i^{t_j}$ : Active, reactive power variation at node  $i$  between two consecutive time steps indicated as  $t_j$ .

$\omega$ : measurement errors.

For the estimation of voltage sensitivities, (10) can be solved by applying Generalized Least Squares, assuming enough

measurements ( $t_m > 2C$ ) to define an over-determined system:

$$SPQ_i = \left( \Delta PQ^T \Sigma^{-1} \Delta PQ \right)^{-1} \Delta PQ^T \Sigma^{-1} \Delta E_i \quad (11)$$

where the correlation matrix  $\Sigma$  has the following structure [24]:

$$\Sigma = \begin{pmatrix} 1 & -0.5 & 0 & \cdots & 0 \\ -0.5 & 1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & -0.5 \\ 0 & \cdots & 0 & -0.5 & 1 \end{pmatrix}$$

The data requirements of the voltage sensitivity module are defined next. According to the grid model assumed, measurements are only available on leaf nodes (MV busbar voltage is considered constant and regulated at 1 p.u. unless observer meters are present). High frequency (1 minute-15 minutes) active/reactive power and voltage magnitude measurements must be available. Finally, it must be noted that SENS module may fail in case  $\Delta PQ^T \Sigma^{-1} \Delta PQ$  presents linear dependencies.

The data set used consists of half hour measurements of active energy which need to be transformed (by assuming constant consumption) to instantaneous active power measurements due to the absence of suitable large open data sets. Reactive power measurements are then simulated by assuming random power factors in the range [0.85, 1], typical for residential consumers. Voltage magnitude measurements are simulated by running power flows in MATLAB/Matpower and Gaussian noise is added to the voltage measurements as described in [24], to generate realistic scenarios.

The voltage sensitivity module calculates consumers' voltage sensitivities by using (11) with a sliding window of size larger than  $2 \cdot C$ . Sensitivities are estimated per window, and real time calculation is possible, thus making real time detection feasible too. The module for NTL detection requires only self-sensitivities. Given the network topology, the error  $S_{P,w}^{err}$  between calculated self-sensitivity of each node and its theoretical value is calculated per window  $w$ . Furthermore, only the active power voltage sensitivities  $S_P$  are utilized due to their increased calculation accuracy resulting from the dominance of the resistive components of the network. Finally, the resulting  $S_{P,w}^{err}$  per window  $w$  and consumer  $i$  is the weight output  $w_i^{Sens}$  of the SENS module:

$$w_i^{SENS} = S_{P,w}^{err} = \frac{|S_{P,w}^{calc} - S_{P_i}^{real}|}{S_{P_i}^{real}} \quad (12)$$

where  $S_{P,w}^{calc}$  is the self-sensitivity term of the node consumer  $i$  is connected to, calculated by (11) for time window  $w$  and  $S_{P_i}^{real}$  is the respective self-sensitivity term extracted from network topology.

### IV. NTL MINIMIZATION FOR DETECTING FRAUDS

In this section, the third module (NTL-MIN module) of the NTL detection system is described, incorporating optimization techniques for accurate NTL detection. The module detects

which consumers commit fraud, while calculating the amount of stolen power in kW per time step.

#### A. Optimization Problem Formulation

Let  $U$  be the set of vertices representing all network buses and  $L$  a subset of  $U$  representing the leaf nodes. Given the network topology and voltage magnitude measurements on leaf nodes, it is possible to express the NTL detection problem as an optimization problem with the objective to minimize losses. An optimization problem for the Bus Injection Model (BIM) is formed:

$$\min J = P_{Loss}^{Grid} = \sum_{k=1}^{|U|} (P_k^G - P_k^D - P_k^{fraud}) \quad (13)$$

$$s.t. P_k + iQ_k = V_k \sum_{j=1}^{|U|} V_j^* Y_{kj}^*, \forall k \in U \quad (13a)$$

$$P_k = P_k^G - P_k^D - P_k^{fraud}, \forall k \in U \quad (13b)$$

$$Q_k = Q_k^G - Q_k^D - Q_k^{fraud}, \forall k \in U \quad (13c)$$

$$v_k^{meter^2} - \varepsilon \leq |V_k|^2 \leq v_k^{meter^2} + \varepsilon, \forall k \in L \quad (13d)$$

$$P_k^{fraud} \geq 0, Q_k^{fraud} \geq 0, \forall k \in L \quad (13e)$$

$$P_k^{fraud} = 0, Q_k^{fraud} = 0, \forall k \notin L \quad (13f)$$

$$P_k^G = 0, Q_k^G = 0, \forall k \neq 1 \quad (13g)$$

$$P_k^D = 0, Q_k^D = 0, \forall k \notin L \quad (13h)$$

$P_{Loss}^{Grid}$  represents the network active power losses (technical and non-technical),  $P_k^G$  the active power generation on bus  $k$ ,  $P_k^D$  the active power consumption as measured at bus  $k$ ,  $P_k^{fraud}$  the amount of fraud for bus  $k$  (control variable) while  $|U|$  is the cardinality of set  $U$ , equal to the number of nodes  $N$ .

Constraints (13a)-(13c) express power balances at every bus.  $P_k, P_k^G, P_k^D, P_k^{fraud}$  are the active power mismatch, generation, demand measurement and fraud at bus  $k$  respectively.  $Q_k, Q_k^G, Q_k^D, Q_k^{fraud}$  are the reactive power mismatch, generation, demand measurement and fraud at bus  $k$  respectively.  $Q_k^{fraud}$  is utilized in order to ensure that cases where reactive power is influenced by fraud are included.  $V_k$  is the complex voltage at bus  $k$  while  $Y_{kj}^*$  represents the complex conjugate of the  $kj$  element of the network admittance matrix  $Y$ .  $v_k^{meter}$  denotes the voltage magnitude as measured on bus  $k$ . Constraint (13d) drives the problem solution towards NTL detection by forcing voltage magnitudes to be close to the voltage magnitude measurements acquired by smart meters on leaf nodes, while  $\varepsilon$  is a small number representing voltage deviation tolerance used to allow problem solution under voltage measurement noise. Constraint (13e) suggests that malicious consumers will always try to benefit from fraudulent activity. Constraint (13f) expresses the fact that fraud may take place only on leaf nodes, since this is where consumers are connected (13h). Finally, (13g) assumes production only at the slack bus ( $k = 1$ ), but may be easily removed to consider distributed generation within the network.

Given the network topology (admittance matrix  $Y$ ) and the constraints presented above, (13) will drive  $P_k^{fraud}$  to include

all non-technical losses. At this point the value of the objective function will be equal to the sum of technical losses.

#### B. Convex Relaxation

The formulated problem is non-linear and non-convex and thus convex relaxation can be used to obtain a close approximation of the optimal solution. SDP relaxation [25] is applied while constraints (13a)-(13d) are relaxed taking into account the analysis presented in [26]:

$$Tr(\underline{Y}_k W) = P_k^G - P_k^D - P_k^{fraud}, \forall k \in U \quad (14a)$$

$$Tr(\bar{Y}_k W) = Q_k^G - Q_k^D - Q_k^{fraud}, \forall k \in U \quad (14b)$$

$$v_k^{meter^2} - \varepsilon \leq Tr(M_k W) \leq v_k^{meter^2} + \varepsilon, \forall k \in L \quad (14c)$$

$$W \geq 0 \quad (14d)$$

where

$$M_k = \begin{bmatrix} e_l e_l^T & 0 \\ 0 & e_l e_l^T \end{bmatrix} \quad (15)$$

$$y_k = e_l e_l^T Y \quad (16)$$

$$\underline{Y}_k = \frac{1}{2} \begin{bmatrix} Re(y_k + y_k^T) & Im(y_k^T - y_k) \\ Im(y_k - y_k^T) & Re(y_k + y_k^T) \end{bmatrix} \quad (17)$$

$$\bar{Y}_k = -\frac{1}{2} \begin{bmatrix} Im(y_k + y_k^T) & Re(y_k - y_k^T) \\ Re(y_k^T - y_k) & Im(y_k + y_k^T) \end{bmatrix} \quad (18)$$

$$X = [Re(V)^T Im(V)^T]^T \quad (19)$$

$$W = XX^T \quad (20)$$

$e_l$  denotes the  $l$ th standard basis vector of  $R^k$  for  $k \in U$ , and  $V = [V_1, V_2, \dots, V_N]^T$ . The solution of the relaxed problem is an exact solution of the original problem if  $rank(W) = 1$  is satisfied [26]. Hence, this constraint has to be checked after solving the relaxed convex problem [27].

The NTL-MIN module described here uses the active/reactive power and voltage magnitude measurements on leaf nodes. The optimization problem (13), (13e-h), (14a-d) is a single period SDP problem solved per time slot (e.g., per 15 minutes). In order to ensure that the optimization problem is feasible, it is assumed that voltage measurements cannot be manipulated by malicious users. It should be noted that the module calculates the exact amount of fraud  $P_{i,t}^{fraud}$  per consumer  $i$  and time slot  $t$ , thus it is possible to estimate the time and the fraud power  $P_{i,t}^{fraud}$ . The output of the module is a weight attached to each consumer:

$$w_i^{NTL-MIN} = \frac{1}{T} \sum_{t=1}^{t_{end}} \frac{P_{i,t}^{fraud}}{P_{i,t}^{fraud} + P_{i,t}^D} \quad (21)$$

where  $P_{i,t}^D$  is the meter  $i$  active power measurement at time  $t$ .

#### V. NTL DETECTION SYSTEM OVERVIEW

In this section the integration of the SVM, SENS and NTL-MIN modules into a single NTL detection system (NTL-SYS) is presented. The result is a system more effective than the application of the individual modules, since:

- The SENS and NTL-MIN modules can also detect “Smart” attacks (breakouts are not present in the consumer time series).

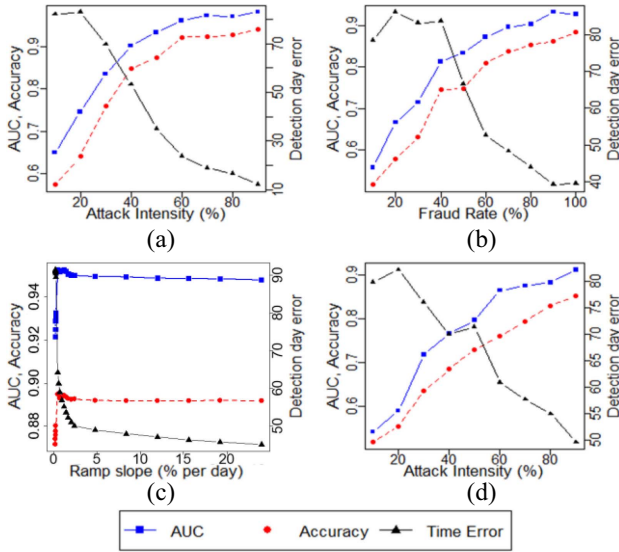


Fig. 2. SVM module performance metrics for (a) BASE, (b) INTERRUPT, (c) SMART and (d) COMBINED attack types.

- The SVM module can detect frauds under corrupted voltage measurements (either due to increased voltage noise or deliberately corrupted by a malicious user).

The three modules have different properties and requirements and cover a wide range of possible frauds. The SVM module has the lowest data requirements based only on time series of individual active energy consumptions, while the SENS and NTL-MIN modules require network topology and P, Q and V measurements at leaf nodes. The advantages of each module can be combined into a single system by assigning a suitable weight to each of their outputs.

The SVM module output,  $w_i^{SVM}$ , expresses the probability of a consumer belonging either to the malicious or benign class and receives values between 0 and 1. The SENS module output  $w_i^{SENS}$  expresses the error between the theoretical and calculated values of voltage self-sensitivity per consumer. This weight is then normalized in the range of [0, 1] to allow the combination of the SENS module with the SVM and NTL-MIN modules. Finally, the NTL-MIN module produces a weight  $w_i^{NTL-MIN}$  per consumer  $i$  which expresses the percentage of electricity stolen and also receives values in [0, 1]. Each of these weights can be evaluated independently in order to assess if a specific consumer is malicious or not. The three modules can also be combined into a single system by multiplying the three aforementioned weights and calculating a single weight  $w_i^{NTL-SYS}$  which is used to rank consumers for inspection. Weight multiplication (22) was chosen as a simple rule which minimizes FPR, but other approaches can be followed [28] according to the needs of the NTL detection system user.

$$w_i^{NTL-SYS} = w_i^{SVM} \cdot \text{normalize}(w_i^{SENS}) \cdot w_i^{NTL-MIN} \quad (22)$$

Given a list of consumers ranked for inspection with a priority determined by weights  $w_i^{NTL-SYS}$ , the NTL detection system user defines the number of physical inspections  $Ins_p$ , to be performed. Starting from the top of the ranked

list, a number of  $Ins_p$  consumers are considered malicious. Performance metrics, such as False Positive Rate (FPR), Detection Rate (DR) and Accuracy (ACC) can then be calculated. In addition, each inspection is associated with a relative cost  $C^{Ins_p}$  (visits of technical personnel, etc) and produces a reward  $R^{TP}$  for true positive (TP) cases. The definition of  $R^{TP}$  varies and is usually defined by estimating the amount of stolen energy charge at a higher price  $q$  that includes some penalty. A simple model would thus define the total compensation  $I$  of the NTL detection system user as:

$$I = R^{TP} - Ins_p \cdot C^{Ins_p} = \sum_{i \in TP} E_i^{stolen} \cdot q - Ins_p \cdot C^{Ins_p}. \quad (23)$$

## VI. SIMULATION RESULTS

The operation of each module is first simulated separately in order to demonstrate the influence of various parameters using the Irish Smart Energy Trial data set. After that, the operation of the integrated NTL detection system given a number of fraud scenarios is simulated.

### A. SVM Module

The SVM module calculates a weight expressing the probability with which a consumer commits fraud given his annual active energy consumption curve. The features presented earlier are extracted per consumer and a radial basis function (RBF) SVM is trained to detect frauds. Grid search is used for optimizing SVM-RBF parameters ( $C^{SVM} = 10$ ,  $\gamma^{SVM} = 0.15$ ) and 5-fold cross-validation is used in order to ensure that the classifier generalizes well and avoids overfitting. SVM is trained on a mixture of all fraud types and tested on each type separately to better evaluate its performance. 70% of the data set (1782 consumers out of the initial 2546) is used as a base for constructing the training set, while the rest 30% (764 consumers) is used for the test set. The training set is produced by randomly partitioning the 1782 consumers into two equal sets of 891 benign and malicious consumers each. A number of 48 different attacks (Base, Interrupt, Smart and Combined attacks of different parameter values) are then simulated on the malicious users set, resulting in a malicious users set of 42768 samples, while benign users are equal to 891. In order to avoid class imbalance issues during the SVM training, oversampling is performed on what is now the minority class (benign class), by replicating all benign samples 48 times. Thus, a balanced training set of 85536 samples is produced. The test set is produced by keeping the number of benign and malicious consumers equal and by simulating various attack strategies (combinations of parameters  $a_t$ ,  $r$  and  $b$ ), not present in the training set.

Fig. 2 (a) presents the performance of SVM on the “Base” attack type (1) with respect to attack intensity. The “Interrupt” attack type (2) is more difficult to detect (although decreasing fraud rate  $r$  means less NTL). Fig. 2 (b) presents the SVM model performance metrics with respect to fraud rate (random attack intensities between 30% and 100%). The effect of the ramp slope  $b$  for the “Smart” attack type (3) is investigated in Fig. 2 (c) (attack intensity receives random values between 30% and 100%). Simulations suggest that the ramp slope does

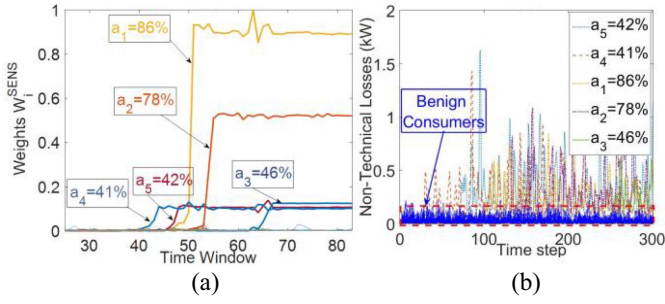


Fig. 3. Example of (a) the SENS and (b) the NTL-MIN module operation.

not affect significantly the SVM performance. However, for small values of the ramp slope (lower than 0.5%/day) the performance slightly deteriorates. The impact of the ramp slope is reflected on the detection day error, since for low ramp slopes it can reach 90 days. Finally, the SVM is tested on the “Combined” attack type (4) where fraud rate receives random values between 30% and 100% and ramp slope receives random values between 0.5%/day and 10%/day. Performance metrics are presented in Fig. 2 (d). Both AUC and accuracy are lower than the three cases earlier presented indicating the difficulty to detect such frauds. The SVM module seems to perform reasonably well in most cases though, with AUC greater than 70% for attack intensities over 30%.

### B. Voltage Sensitivity Module

The Voltage Sensitivity (SENS) module calculates a weight per consumer per time window. The window length in this case is set to  $5 \cdot N$ , where  $N$  is the number of network nodes. The main parameters expected to influence the module’s performance are network size  $N$ , attack intensity  $a_i$  and malicious consumer rate  $f$  (percentage of consumers committing frauds).

The main metric calculated for the evaluation of the module is the difference  $\Delta W^{SENS}$  between the minimum weight produced for malicious consumers and the maximum weight produced for benign ones. Large differences (over 10%) indicate that the module can detect NTLs.

The module operation is demonstrated by generating a realistic scenario, with a random LV network of  $N = 87$  nodes,  $C = 48$  consumers, where  $f = 5\%$  of consumers commit frauds (“Base” attack type) with attack intensities between 20%-90% and random fraud start times. Fig. 3 (a) presents the real time operation of the module, showing how the consumer weights are calculated per time window. The abrupt changes of  $w_i^{SENS}$  indicate the initiation of malicious behavior while the level on which  $w_i^{SENS}$  stabilizes expresses attack intensity.

The effect of network size and attack intensity on the module’s operation is investigated next by producing random networks of sizes ranging between 30 and 210 nodes. “Base” attack is again simulated. In this case, malicious consumer rate  $f$  is set equal to 50% and introduction to fraud time ( $T_{start}$ ) is random. The difference  $\Delta W^{SENS}$  is then calculated and presented in Fig. 4 (a). It may be concluded that the voltage sensitivity module performance is not affected by the network

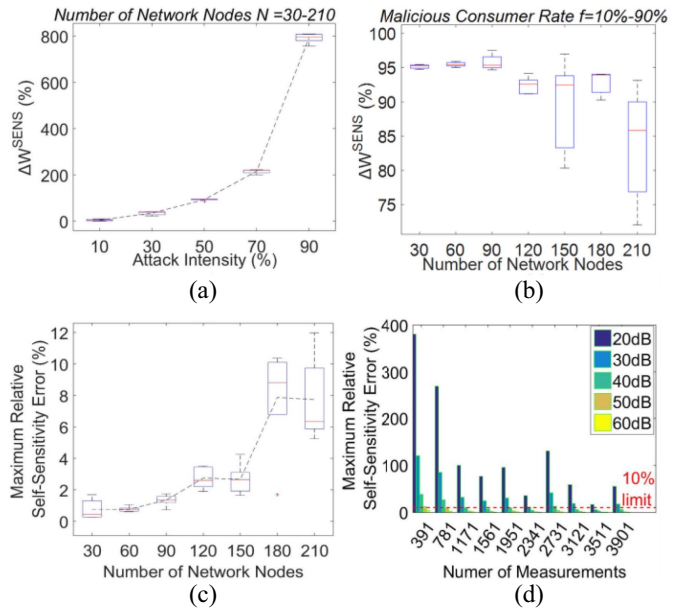


Fig. 4. Impact of (a) number of network nodes, (b) malicious consumer rate, (c) sensitivity estimation accuracy and (d) voltage measurement noise on the SENS module operation.

size and that the module can easily detect frauds with attack intensities over 30%.

The same analysis is performed in order to evaluate the influence of the number of malicious consumers by varying  $f$  between 10% and 90% and setting attack intensity equal to 50%. The number of nodes  $N$  ranges between 30 and 210 with a step of 30. For each value of  $N$ , 5 random networks are simulated. Again, performance is not greatly affected by the number of malicious consumers or network nodes since  $\Delta W^{SENS}$  remains over 72% (Fig. 4 (b)).

In order to assess the module performance correctly, the accuracy in estimating sensitivities is quantified. The maximum relative errors between theoretical and calculated self-sensitivity values are presented in Fig. 4 (c). The same number of nodes  $N$ , ranging between 30 and 210 with a step of 30, is assumed. For each value of  $N$ , 5 random networks are again generated. The maximum error in calculating self-sensitivities remains low for networks up to 150 nodes (less than 4%), while it is acceptable even for larger networks.

The effect of voltage measurement noise (signal-to-noise ratio) is depicted in Fig. 4 (d). The SENS module performance may deteriorate with increased voltage noise. In such cases, increasing the number of measurements used to calculate sensitivities limits the maximum relative self-sensitivity error below 10% for signal-to-noise ratios over 40 dB.

A small variance of the power factor within a certain time window is required in order for the SENS module to operate efficiently. In order to quantify this variance, a network of 62 nodes is used and “Base” attacks are simulated ( $a_i = 50\%$ ,  $f = 50\%$ ). The consumers’ power factor receives random values in the range of  $PF_{ref} \pm p$ , where  $p$  is the maximum power factor deviation) and  $PF_{ref}$  receives values between 0.87 and 0.98. Fig. 5(a) suggests that the SENS module operates well for a power factor variance larger than  $p = 0.0015$  and is not affected by the value of the power factor itself ( $PF_{ref}$ ).

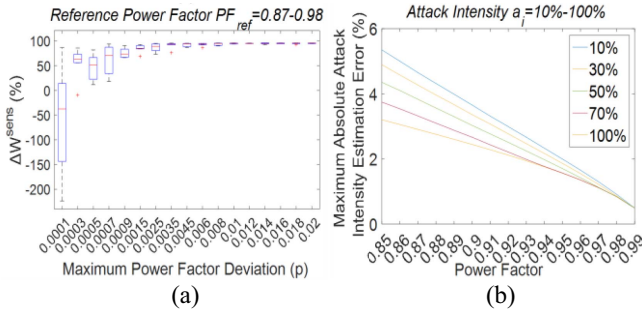


Fig. 5. Power factor effect on (a) SENS, (b) NTL-MIN module performance.

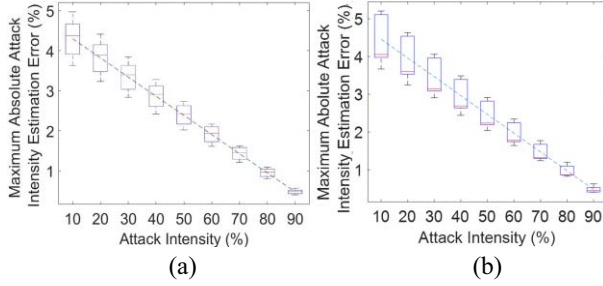


Fig. 6. Impact of attack intensity and (a) number of network nodes, (b) malicious consumer rate on NTL-MIN module operation.

### C. NTL-Minimization Module

The NTL-MIN module performance may also be affected by the network size, malicious consumer rate  $f$  and attack intensity  $a_t$ . The maximum absolute attack intensity estimation error is used to assess the module's accuracy.

First, the realistic scenario of a random low voltage network with  $N = 87$  nodes,  $f = 5\%$  and  $a_t$  between 20%-70% is presented to demonstrate the module's operation. Fig. 3 (b) presents the estimated fraud power (kW) per consumer and time step. The effect of attack intensity and number of nodes is presented in Fig. 6 (a) for malicious consumer rate equal to 50%. It can safely be concluded that the NTL-MIN module performance is not affected by the number of network nodes. Similar results are derived from Fig. 6 (b) where the effect of malicious consumer rate  $f$  is investigated for a network of  $N = 63$  nodes. In fact, the main parameter affecting the module is attack intensity. Error increases for lower intensities although it remains low (less than 5%) in all cases.

The effect of voltage measurement noise is investigated in Fig. 7 by calculating the maximum absolute attack intensity estimation error for various voltage noise levels (cases of 39 and 63 node networks). Increasing voltage noise (signal-to-noise ratio from 40 dB to 20 dB) results in reduced performance and may also lead to infeasible solutions (white areas). Dark blue areas indicate lower errors and the simulation results suggest that the voltage deviation tolerance  $\varepsilon$  must be carefully chosen according to the voltage measurement noise.

The effect of consumer power factor on the NTL-MIN module is presented in Fig. 5 (b). A network of 63 nodes is used, where 50% of the consumers perform "Base" attacks of varying attack intensities (10%-100%). Power factor varies between 0.85-0.99 per consumer and simulations show that

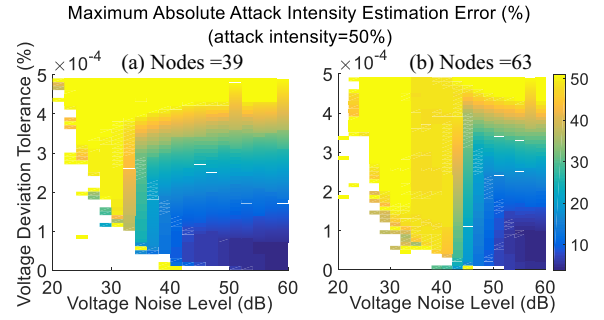


Fig. 7. Impact of voltage measurement signal to noise ratio on NTL-MIN module operation for a network of (a) 39 nodes and (b) 63 nodes.

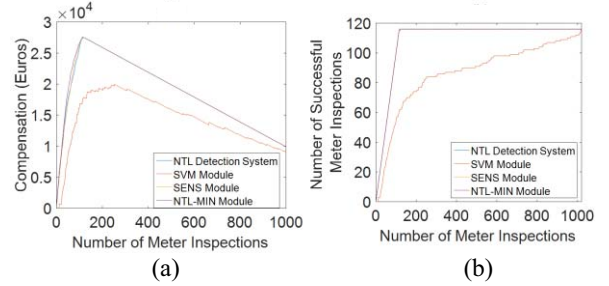


Fig. 8. Impact of number of meter inspections on (a) the compensation produced by the NTL detectors and (b) number of successful meter inspections.

a small error is introduced as the power factor value decreases. The maximum absolute attack intensity estimation error is 5.2%, for attack intensities as low as 10%, indicating that the power factor does not significantly affect the NTL-MIN module performance.

### D. NTL Detection System Operation

The integrated NTL detection system and its modules are tested under a mixture of 40 scenarios including different types of fraud with randomly selected parameters. The results are summarized in Table II. Each scenario includes approximately 500 randomly chosen consumers out of which 50% commit fraud. The consumers are distributed to 10 randomly generated distribution networks of approximately 70 nodes each. For each scenario and module, FPR, DR, Accuracy and AUC are calculated given that the number of meter inspections  $Insp$  is equal to the number of malicious consumers (consumer fraud rate is equal to 50% in this case). The averaged metric values are presented in Table II. The NTL detection system produces good results for all scenarios with the SENS and NTL-MIN modules providing excellent precision. The SVM module exhibits lower performance, which is still acceptable given its modest requirements. The weights assigned to consumers by each of the modules are presented in Fig. 9 for a "Combined" type fraud case with low attack intensity, fraud rate and ramp slope. This is one of the most difficult NTL scenarios simulated, however the NTL detection system still manages to discriminate successfully between malicious and benign users. Finally, a larger scenario is simulated, with 1000 consumers distributed to 20 randomly generated networks where malicious consumer rate  $f = 10\%$ . The effect of the number of

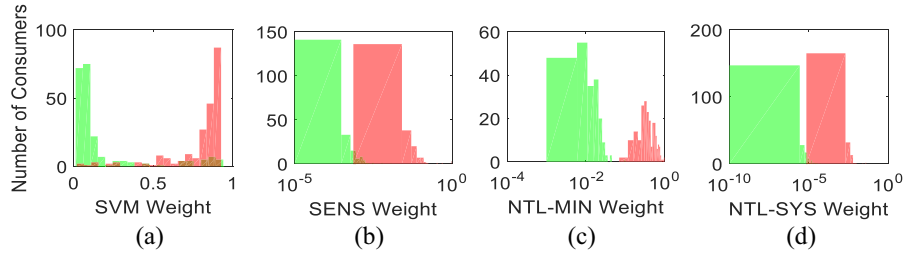


Fig. 9. Module outputs for a hard “Combined” fraud scenario: (a) SVM, (b) SENS, (c) NTL-MIN and (d) NTL detection system.

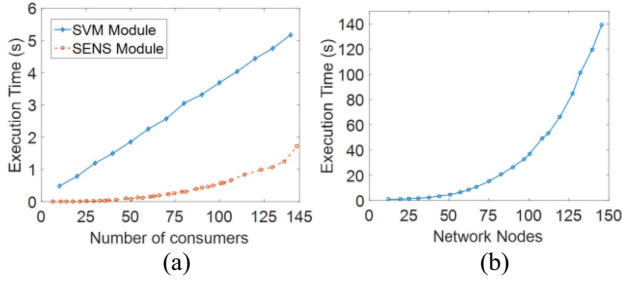


Fig. 10. Simulation run time of (a) SVM, SENS and (b) NTL-MIN module.

TABLE II  
NTL DETECTION SYSTEM PERFORMANCE METRICS

	FPR	DR	ACC	AUC
SVM Module	9.7	88.8	91.2	93.7
SENS Module	~0	98.9	99.4	99.9
NTL-MIN Module	~0	98.9	99.4	99.9
NTL detection system	~0	98.8	99.4	99.9

meter inspections  $Insp$  on performance (true positives) and compensation  $I$  of the NTL detection system is presented in Fig. 8 ( $C^{Insp} = 20\text{Euros}/inspection$ ,  $q = 0.17\text{Euros}/kWh$ ). As the number of meter inspections increases, compensation also increases for the NTL-SYS and its composing modules, up to the point when all malicious users have been detected and inspections do not result in any reward (just cost). For the NTL-SYS, SENS and NTL-MIN modules this number is 10%, equal to the number of fraud consumers. The SVM module produces less compensation due to false positive cases. Furthermore, SVM compensation is maximized when approximately 20% of the population is inspected. Fig. 8 (b) shows that when the SVM module generates maximum compensation, about 67% of the malicious users are detected, a number which could be increased if lower compensation is allowed. The NTL-SYS, SENS and NTL-MIN modules on the other hand, detect 100% of the malicious users with the minimum number of inspections (10%) and maximum compensation.

## VII. COMPARISONS AND DISCUSSION

The scalability of the proposed system and modules is discussed here in terms of complexity. Fig. 10 presents the execution time of the three algorithms as a function of the number of consumers (SVM and SENS modules) and the number of network nodes (NTL-MIN module). The complexity of the SVM module is linear with respect to the number of consumers, and most of the required time is spent on

data processing (feature extraction), rather than operation of the actual SVM. The SVM module is thus quite scalable. The SENS module complexity is bound by the complexity of inverting the matrix  $\Delta P Q^T \Sigma^{-1} \Delta P Q$  (whose size is equal to  $2C \cdot 2C$ ). This complexity is typically polynomial ( $O(n^3)$  in worst case). Finally, the NTL-MIN module complexity is mainly defined by the optimization problem solver (in this case Sedumi [29]) and is again polynomial ( $O(n^2 m^{2.5} + m^{3.5})$ , where  $n$  is the number of decision variables and  $m$  the number of constraints). The main difference with the SVM and SENS modules is that the complexity is not a function of the number of consumers but a function of the number of network nodes. The NTL-MIN module is more complex than the other two and execution times are larger. The most important disadvantage of the NTL-MIN module though, is not time complexity but memory requirements. A standard Windows 7, Intel Core i7-7500U @ 2.7 GHz processor with 16GB of RAM laptop runs out of memory when trying to solve the optimization problem for networks of more than 150 nodes.

The above considerations are summed up in Table III, where the proposed concept is compared with most recent works on NTL detection. DR, accuracy and AUC are chosen as comparison criteria. The proposed scheme has higher AUC values than those reported in recent publications, although it is tested under different data sets and conditions. The SVM module shows performance close to what is reported in similar machine learning and data analytics schemes [11], [12], [30]–[32] with lower DR and accuracy, but higher AUC. The SENS and NTL-MIN modules show performance close to power systems oriented techniques [7], [10]. Other criteria used for comparison is the detection delay (time required for the NTL detection system to decide if a consumer is malicious or not), robustness to new attacks, privacy, cost, resources and scalability. The proposed system is flexible, with detection delay spanning from a few hours to months and high robustness to new attacks, mainly attributed to the SENS and NTL-MIN modules. Privacy can be an issue, especially for those parts utilizing high resolution active/reactive power and voltage measurements. Cost includes any installation, maintenance and operation costs. The proposed scheme does not require the installation of any specialized devices, assuming smart meters are in operation. Instant active/reactive and voltage measurements are required though, which may increase the system’s cost. These data requirements, together with the requirement of LV network topology, make the system more demanding concerning resources. Finally, the system is also

TABLE III  
COMPARISON OF DIFFERENT NTL DETECTION SYSTEMS

System	SVM	SENS	NTL-MIN	NTL-SYS	[30]	[12]	[11]	[31]	[10]	[7]	[32]
DR (%)	88.8	98.9	98.8	98.8	-	-	94.0	65	99-100	98-100	-
ACC (%)	91.2	99.4	99.4	99.4	-	92.5	-	94.4	-	-	-
AUC (%)	93.7	99.9	99.9	99.9	91.0	-	-	81.9	-	-	80
Detection Delay	months	hours	hours	hours-months	months	days	days	months	hours	hours	months
Robustness to new attacks	low	high	high	high	high	high	med.	med.	high	high	med.
Privacy	med.	low	low	low	med.	low	low	med.	high	high	med.
Cost	low	med.	high	med.	low	med.	med.	low	high	high	low
Resources	low	med.	high	med.	med.	med.	low	low	high	high	low
Scalability	high	med.	low	med.	high	high	high	high	low	low	high

flexible in terms of scalability. The SVM module can easily process thousands of consumers at once. The SENS and NTL-MIN modules are operated per network, meaning that a large number of small systems will be required in order to cover a large area.

### VIII. CONCLUSION

A novel concept for detecting NTL in distribution grids has been presented. The proposed NTL detection system consists of three modules based on SVM, sensitivity analysis and optimization. The SVM module has the ability to detect frauds with minimum data requirements and satisfactory performance metrics. The SENS module computes self-sensitivities and successfully detects all types of NTL given voltage magnitude and power measurements. Finally, the NTL-MIN module makes use of optimization techniques and SDP relaxation for estimating both the time and extent of fraud in kW per time step and consumer. The NTL-MIN module accurately detects frauds under all fraud types of NTL assumed. The three modules are analyzed and integrated into a single NTL detection system. A number of simulation scenarios are defined in order to assess the performance of each module separately and of the NTL detection system as a whole. Simulation results show that the system performs well even under stealthy cases of fraud. Finally, the operation of the system is evaluated by estimating compensations generated by detecting NTL.

### REFERENCES

- [1] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft," *Energy Policy*, vol. 39, no. 2, pp. 1007–1015, 2011.
- [2] G. M. Messinis and N. D. Hatziaargyriou, "Review of non-technical loss detection methods," *Elect. Power Syst. Res.*, vol. 158, pp. 250–266, May 2018.
- [3] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010.
- [4] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, Aug. 2008.
- [5] C. C. O. Ramos, A. N. De Sousa, J. P. Papa, and A. X. Falcao, "A new approach for nontechnical losses detection based on optimum-path forest," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181–189, Feb. 2011.
- [6] G. M. Messinis and N. D. Hatziaargyriou, "Unsupervised classification for non-technical loss detection," in *Proc. Power Syst. Comput. Conf. (PSCC)*, 2018, pp. 1–7.
- [7] M. Tariq and H. V. Poor, "Electricity theft detection and localization in grid-tied microgrids," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1920–1929, May 2018.
- [8] S. Weckx, C. Gonzalez, J. Tant, T. De Rybel, and J. Driesen, "Parameter identification of unknown radial grids for theft detection," in *Proc. 3rd IEEE PES Innov. Smart Grid Technol. Europe (ISGT Europe)*, 2012, pp. 1–6.
- [9] P. Kadurek, J. Blom, J. F. G. Cobben, and W. L. Kling, "Theft detection and smart metering practices and expectations in the Netherlands," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. Europe (ISGT Europe)*, 2010, pp. 1–6.
- [10] S. A. Salinas and P. Li, "Privacy-preserving energy theft detection in microgrids: A state estimation approach," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 883–894, Mar. 2016.
- [11] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.
- [12] A. Jindal *et al.*, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1005–1016, Jun. 2016.
- [13] S.-C. Huang, Y.-L. Lo, and C.-N. Lu, "Non-technical loss detection using state estimation and analysis of variance," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2959–2966, Aug. 2013.
- [14] Y. Guo, C.-W. Ten, and P. Jirutitijaroen, "Online data validation for distribution operations against cyber tampering," *IEEE Trans. Power Syst.*, vol. 29, no. 2, pp. 550–560, Mar. 2014.
- [15] J. B. Leite and J. R. S. Mantovani, "Detecting and locating non-technical losses in modern distribution networks," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1023–1032, Mar. 2018.
- [16] *Irish Social Science Data Archive*. Accessed: Mar. 27, 2018. [Online]. Available: <http://www.ucd.ie/issda/data/commissionforenergyregulationcer/>
- [17] *Breakout Detection R Package*. Accessed: Nov. 23, 2018. [Online]. Available: <https://github.com/twitter/BreakoutDetection>
- [18] N. A. James, A. Kejariwal, and D. S. Matteson, "Leveraging cloud data to mitigate user experience from 'breaking bad,'" in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2016, pp. 3499–3508.
- [19] A. Jindal, N. Kumar, and M. Singh, "Internet of energy-based demand response management scheme for smart homes and PHEVs using SVM," *Future Gener. Comput. Syst.*, Apr. 2018.
- [20] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, "Machine learning methods for attack detection in the smart grid," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 8, pp. 1773–1786, Aug. 2016.
- [21] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.
- [22] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R Package Version 1.6-6*. [Online]. Available: <http://CRAN.R-project.org/p>
- [23] M. Brenna *et al.*, "Automatic distributed voltage control algorithm in smart grids applications," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 877–885, Jun. 2013.
- [24] C. Mugnier *et al.*, "Model-less/measurement-based computation of voltage sensitivities in unbalanced electrical distribution networks," in *Proc. Power Syst. Comput. Conf. (PSCC)*, 2016, pp. 1–7.
- [25] S. H. Low, "Convex relaxation of optimal power flow—Part II: Exactness," *IEEE Trans. Control Netw. Syst.*, vol. 1, no. 2, pp. 177–189, Jun. 2014.
- [26] J. Lavaei and S. H. Low, "Zero duality gap in optimal power flow problem," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 92–107, Feb. 2012.

- [27] A. Giannitrapani, S. Paoletti, A. Vicino, and D. Zarrilli, "Optimal allocation of energy storage systems for voltage control in LV distribution networks," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2859–2870, Nov. 2017.
- [28] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mac. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
- [29] Y. Labit, D. Peaucelle, and D. Henrion, "SEDUMI INTERFACE 1.02: A tool for solving LMI problems with SEDUMI," in *Proc. IEEE Int. Symp. Comput.-Aided Control Syst. Design*, Sep. 2002, pp. 272–277.
- [30] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Trans. Smart Grid*, to be published.
- [31] N. F. Avila, G. Figueroa, and C.-C. Chu, "NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7171–7180, Nov. 2018.
- [32] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.



**George M. Messinis** received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Greece, in 2011, where he is currently pursuing the Ph.D. degree with Electric Power Division focusing on smart grids. His research interests include smart grid operation, smart meters, fraud detection, and data science. He is a member of the Technical Chamber of Greece.



**Alexandros E. Rigas** received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Greece, in 2011, where he is currently pursuing the Ph.D. degree with Electric Power Division focusing on power system optimization. His research interests include power system optimization, distributed generation, and storage systems. He is a member of the Technical Chamber of Greece.



**Nikos D. Hatziaargyriou** (SM'90–F'09) has been a Faculty Member with the Electrical and Computer Engineering School, National Technical University of Athens, Greece, since 1984. Since 2015, he has been the Chairman of the Hellenic Distribution Network Operator. He has authored the book entitled *Microgrids: Architectures and Control* and over 200 journal publications and 500 conference proceedings papers. He is the Past Chair of the Power System Dynamic Performance Committee. He is an Honorary Member of CIGRE and the Past Chair of CIGRE SC C6 "Distribution Systems and Distributed Generation." He is the Chair of the EU Technology and Innovation Platform on Smart Networks for Energy Transition. He has participated in over 60 RD&D projects performed for the EU Commission, electric utilities and manufacturers in Europe, for both fundamental research and practical applications. He is included in the 2016 and 2017 lists of the top 1% most cited researchers.