Electric Grid Power Flow Model Camouflage Against Topology Leaking Attacks

Ian Markwood[†], Yao Liu[†], Kevin Kwiat[‡], and Charles Kamhoua[‡]

[†]University of South Florida, Tampa, FL, U.S.A
[‡]Air Force Research Laboratory, Information Directorate, Cyber Assurance Branch, Rome, NY, U.S.A.

***Abstract*—The power flow model for DC power grids has been used theoretically to launch false data injection attacks (FDIAs) against state estimation. We recognize FDIAs are just one possible attack using the power flow model and that the grid topology information within the model implies its discovery may also facilitate topology-based attacks. We show attackers can derive the power flow model, and thus the topology also. Indeed, with incomplete data, attackers can accurately reconstruct regions of the model, or topology, all that is necessary to launch an attack. We also illustrate how to cause such attackers to derive instead a convincing fake model by camouflaging the real model. Consequently, no sensitive information will leak, so attacks based on this fake model will be ineffective, rather alerting grid administrators to the attacker's efforts. Using five test cases included in the MATLAB power flow analysis tool MATPOWER, ranging from 9 to 300 buses, an average 67.0% of the topology may be derived with a 69.1% model accuracy. Lastly, we find reconstructions of small portions of the model sufficient for performing FDIAs with 75% success, and that camouflage prevents 93% of them in all but the 9-bus case.**

I. Introduction

Fundamental to the productivity and stability of a developed nation is its electrical power grid, which is responsible for delivering generated power over great distances to individuals, businesses, and services, thereby maintaining modern life. The cascading outages that in 2003 affected some 50 million people in the northeastern US into Canada are now a few years past, but physical and cyber-attacks against the power grid occur on the order of every four days in the US [1], making a robust defense highly imperative. In a publicized 2013 attack, gunmen caused $15 million in damage to a northern California substation and were not caught. Such physical attacks are necessarily localized, while indeed cyber-attacks have no such constraint and consequently far higher potential impact. A strong defense in this realm is thus of the highest priority.

We identify that a class exists of cyber-attacks which target the grid from a knowledge of its power flow model, which we call *model-specific attacks*. Within power system monitoring, the state estimation (SE) process uses this power flow model to provide the control center an approximate understanding of the power flow throughout the grid and thereby a means to make corrections and preserve stability [2]. The first model-specific attacks discovered are the well-researched false data injection attacks against SE [3], where an attacker with knowledge of the power flow model can corrupt the SE accuracy by injecting false errors into select power meters, without these errors being detected [3]. Additionally, as the power flow model contains the topology (i.e. the interconnection network) of the grid [4], we identify and evaluate an additional model-specific attack in the form of grid topology information leakage, a compromise of proprietary information and potential security risk.

Beyond topology leakage and false data injection attacks, the sensitive nature of the information within the power flow model indicates a potential for additional model-specific attacks with various targets and impacts. Rather than defending against each of these other attacks as they arise, a more proactive research approach will develop a thorough protection of the model. Power grid administrators should suppress access to the model, but it is important to identify any other means of discovering or deriving it. Accordingly, this research finds that an attacker may reconstruct the model in Direct Current (DC) systems using information from the SE process. This information may be intercepted during communications between the grid administrators and reporting or distribution centers, for example, while the model itself would not need to be present in such communications. This side channel attack places a lighter data-collection requirement on the attacker than to directly acquire the sensitive power flow model. Furthermore, we also explain how an attacker can use incomplete data to recover portions of the model. This may be all that is necessary for some model-specific attacks; indeed, in a case study we find that a small partial model is sufficient for successful false data injection attacks with high probability.

Central to this reverse engineering attack is the realization that the power flow model is described by a sparse matrix, which allows a reasonable starting approximation even from noisy data where an exact reproduction using traditional algebraic methods fails entirely. However, this approximation is quite crude, so we have developed novel post-processing techniques such that the attacker may uncover substantially more accurate information, using knowledge of the data arrangement methodology fundamental to all DC power flow models. For example, the symmetry of this matrix and the guaranteed zero-sum nature of each row/column provide a powerful knowledge base supporting the attacker in this refinement process.

The power flow model structure also yields an interesting protection scheme. The integrity and availability of the power grid being of the utmost importance, an adversary attempting this attack should not merely be prevented from doing so,

Approved for Public Release; Distribution Unlimited 88ABW-2016-3300, dated 30 Jun 2016.

but also should be identified as having tried. As the attack requires only passive eavesdropping, its discovery is difficult unless the attacker believes the attack successful and so carries on to use the model in some way. Therefore, we construct a structurally accurate fake power flow model which we use to camouflage the real model during any transmission of SE data such that an attacker performing this reconstruction method will derive the fake model instead. The attacker is inspired to believe the derivation successful, while any subsequent topology or model-specific attacks will necessarily fail due to their incorrect basis. However, the failed attacks can alert grid administrators to the threat, such that proper defensive measures can be taken and law enforcement can be notified.

In this research, we find that an attacker may uncover large portions of the DC power flow model, and show how this can be used in our novel model-specific attack of grid topology information leakage. Additionally, we perform a brief case study finding that the other currently existing model-specific attack, the false data injection attack, is feasible using the reverse engineered power flow model. We then exhibit a means to camouflage the power flow model so an attacker will not successfully discover it but instead a fake version. We test these developments using a MATLAB power flow analysis tool suite called MATPOWER, finding that an attacker can accurately derive a large portion of the unprotected power flow model in each of the IEEE 9, 14, 30, 118, and 300-bus test cases included therein. Specifically, an attacker may successfully recover the model with an average 69.1% accuracy for the systems represented in these test cases, allowing for unauthorized discovery of 67.0% of their topology after using the model for topology leakage. Under camouflage, however, the full model derivation attack is reduced to an average 10.1% accuracy. Without camouflage, the derived models prove sufficient to successfully launch false data injection attacks with 75% probability. While the camouflage proves unsuitable for preventing these attacks in the 9-bus system, it does however prevent 93% of them in the other systems.

II. Background Information

Not only is uninterrupted power conveyance mandatory for grid stability, but so also is the quality of this power, requiring minimal fluctuation in voltage magnitude and frequency to prevent malfunction of connected electronics. System monitoring is the process of managing information collected by meters in pertinent locations within the grid and stored in a Supervisory Control and Data Acquisition (SCADA) telemetry system. These measurements may include bus voltages, power injections, and power flows in the various subsystems of a power grid, and the control center uses these to understand the system context, or state. In particular, state estimation (SE) is that part of system monitoring in which an approximation of the current state is derived from meter measurements and a known power flow model. The resulting series of state variables is used as input to contingency analysis, which modifies the use of components within the grid to preserve its proper function despite possible equipment failures [2].

In DC SE, the state variables and meter measurements are related through a linear regression model as

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}. \tag{1}$$

AC SE entails a nonlinear relation between these values and so is often approximated by DC SE for analysis purposes. For long distance bulk power transmission, DC power is used over high-voltage, direct current (HVDC) systems [5]. We focus on the DC SE environment for this segment of our research; since DC SE is fundamental to AC SE, the attack and defense we explore here can serve as a basis to extend to the AC model. In Equation 1, $\mathbf{z}$ is the vector of meter measurements of length m, $\mathbf{e}$ is the length m vector of errors for each meter, $\mathbf{x}$ is the vector of state variables of length n, and $\mathbf{H}$ is the power flow model, in the form of an $m \times n$ full rank matrix [2].

This information has two direct implications, that the power flow model $\mathbf{H}$ may be used to cause harm, and that it may be derived from other information in the absence of its own availability. False data injection attacks, introduced previously as an example of model-specific attacks, follow the first implication, and build an attack vector as a linear combination of the columns of $\mathbf{H}$, which is added to $\mathbf{z}$ in the form of injections at each applicable meter. These injections undetectably cause the estimated state variables $\mathbf{x}$ to deviate from their correct values which could destabilize the system monitoring process [3]. This research considers the second implication, the possibility of determining $\mathbf{H}$ from $\mathbf{z}$ and $\mathbf{x}$, given that as a result of research in false data injection attacks, $\mathbf{H}$ is now considered sensitive information.

Specifically, for study of the DC power flow, state variables $\mathbf{x}$ are the voltage angles $\boldsymbol{\theta}$ of length n, and the measurements $\mathbf{z}$ are the net injections $\mathbf{P}$ measured at each of the n buses in the power system. They are related by the power flow model $\mathbf{H}$, which for DC is the nodal admittance matrix $\mathbf{Y}$ describing the electrical admittance over any branches between buses [6]. $\mathbf{Y}$ is of size $n \times n$ to hold data for possible branches between each of the n buses and every other, except itself; the diagonal elements of $\mathbf{Y}$ are called self-admittances and describe the total admittance ending at that bus, to satisfy Kirchhoff's circuit laws. In all, each element of $\mathbf{Y}$ is given by

$$Y_{ij} = \begin{cases} y_{ii} + \sum_{k \neq i} y_{ik} & j = i \\ -y_{ij} & j \neq i \end{cases}$$

where y_{ij} is the admittance between buses i and j, and y_{ii} is the admittance to ground at bus i, typically zero [7].

After this introduction of their physical meaning, this paper will refer to these values in their general terms of $\mathbf{H}$, $\mathbf{x}$, and $\mathbf{z}$. With this construction, each branch admittance appears twice (Y_{ij} and Y_{ji}, $i \neq j$), so $\mathbf{H} = \mathbf{Y}$ is necessarily symmetric. The diagonal elements are always positive, and the off-diagonal elements are all negative. Finally, because each bus' self-admittance (diagonal element) is the sum of all admittances for branches connected to that bus (off-diagonal elements in that row/column), each row/column sums to zero. These properties will aid in the reconstruction process through post-processing steps we designed to enforce them upon the initial estimation.

III. Power Flow Model Derivation Attack Process

Recovery of the power flow model may be performed in two major components, with the second refining the results of the initial calculation. Additionally, this section describes this process for deriving a partial model using incomplete data.

A. Initial Reconstruction

As already introduced, we derive the power flow model for DC systems using pairs of meter measurements and corresponding state variables. A ready formulation appends n sets of measurements together as columns of an $m \times n$ matrix $\mathbf{Z}$ (with corresponding $m \times n$ matrix of measurement errors $\mathbf{E}$) and likewise n sets of state variables as an $n \times n$ matrix $\mathbf{X}$. The relation

$$\mathbf{Z} = \mathbf{HX} + \mathbf{E} \tag{2}$$

now holds, but the error $\mathbf{E}$ is irrecoverable from the measurements $\mathbf{Z}$. As this error, normally assumed to be white Gaussian noise, is suffered by all measurements uniquely, Equation 2 is better described as

$$\mathbf{Z}_E = \mathbf{HX} \tag{3}$$

with $\mathbf{Z}_E$ and $\mathbf{X}$ known by the attacker. However, due to the random error hidden in $\mathbf{Z}_E$, a simple right multiplication of $\mathbf{X}^{-1}$ in Equation 3 returns a very inaccurate result for $\mathbf{H}$. With the inaccurate result of the traditional algebraic method, the attacker's challenge is to find another way to solve for $\mathbf{H}$. The sparsity of $\mathbf{H}$ is now relevant, as it enables an approximation via compressive sensing, despite the error in $\mathbf{Z}_E$.

Compressive sensing [8] is the process of recovering sparse signals using a dictionary, or sensing matrix, enabling much more expedient conveyance than that required by the Shannon/Nyquist sampling theorem. Considering the dictionary $\mathbf{A}$ and sparse signal $\mathbf{x}$ generating $\mathbf{b}$ by relation $\mathbf{Ax} = \mathbf{b}$, $\mathbf{b}$ may be sent instead of $\mathbf{x}$, requiring less sampling [8]. This is made possible by the fact that minimizing the L^1-norm (through basis pursuit, for example) in reconstruction of $\mathbf{x}$ from $\mathbf{b}$ and $\mathbf{A}$ results in the sparsest solution with high probability [9].

Building from this technique, we show the derivation of $\mathbf{H}$ row by row, in the presence of the aforementioned error. Equation 3 implies

$$\mathbf{z}_i = \mathbf{h}_i\mathbf{X}$$

where $\mathbf{z}_i$ and $\mathbf{h}_i$ are the i-th rows of $\mathbf{Z}_E$ and $\mathbf{H}$, respectively. The transpose property ensures

$$\mathbf{z}_i^T = \mathbf{X}^T\mathbf{h}_i^T$$

which is of the form $\mathbf{Ax} = \mathbf{b}$ indicating the sparse row $\mathbf{h}_i$ may be solved as in compressive sensing, with $\mathbf{X}^T$ as the dictionary. Accordingly, each row of $\mathbf{H}$ is estimated through basis pursuit [10], which minimizes the L^1-norm, and recompiled into an overall approximation $\hat{\mathbf{H}}$.

The recovered approximation is more resilient to error than the direct calculation, naturally, as it is an approximation, but it requires considerable refinement to converge more closely to the original. As visible in Tables II and I in our Evaluation, the initial reconstruction is highly inaccurate and will be unsuitable for use in any further attacks. It is, however, a good starting point, and the following section details a series of methods we have developed for refinement.

B. Post-Processing

We detail here a four-step process for augmenting the initial reconstruction, based on the properties of the DC power flow model detailed in the background information in Section II:

1) H is symmetrical
2) Diagonal entries are always positive
3) Off-diagonal non-zero entries are all negative
4) Each row (and column) sums to zero

The second property always holds immediately, due to the strength (high magnitude) of these values relative to the rest in each vector; the rest do not. In the four steps below, we ensure they do, while increasing accuracy over the initial reconstruction $\hat{\mathbf{H}}$. The first and second steps run in parallel, and their results are merged in the third. Thresholds appear in the first and second steps, and are optimized in Section V-B.

1) Item-specific symmetry enforcement: This step addresses the first and third properties above, from a focus on each pair of entries across the diagonal. It begins with a simple threshold that zeros out off-diagonal values of $\hat{\mathbf{H}}$ above a small negative threshold t_1. This immediately satisfies Property 3. Also, while basis pursuit finds the sparsest solution for each $\mathbf{h}_i$ in $\mathbf{z}_i^T = \mathbf{X}^T\mathbf{h}_i^T$, in presence of error in $\mathbf{Z}_E$ this results in most values in the estimated $\hat{\mathbf{h}_i}$ being near (but not equal) to and necessarily changed to zero, which this threshold achieves. Next, for symmetry across the diagonal (Property 1), entries that are zero on one side of the diagonal of $\hat{\mathbf{H}}$ are zeroed out on the other side if necessary, while entries that are non-zero on both sides of the diagonal are averaged. The results of this step are called $\hat{\mathbf{H}}_1$. This step has a weakness in the form of non-zero entries which are removed by the threshold for being smaller than the noise the initial threshold is calibrated to, so the following parallel step works to recover these.

2) Full row/column symmetry enforcement: Also building from the initial reconstruction, this step exploits the necessary symmetry between the full i-th row and i-th column, where the previous treated only the symmetry of each entry across the diagonal. We observe that actual non-zero elements of $\mathbf{H}$ often appear as local minima in both the i-th row and i-th column of $\hat{\mathbf{H}}$, while actual zero elements may appear as a minima in the row, but not the column, for example. This observation is again due to their relative strength in the compressive sensing reconstruction process, and is visible in Figure 1 and enables a noise-agnostic recovery of those non-zero elements of $\mathbf{H}$ which are closer to zero than the average noise in $\hat{\mathbf{H}}$. Note in Figure 1 the wildly dissimilar nature of the row and column plotted together, except at the four points where non-zero off-diagonal values should appear.

Therefore, for each row and column pair, positive values in the initial reconstruction are set to zero for sake of Property 3, after which any matching local minima are identified as the non-zero elements of that pair. "Matching" local minima are

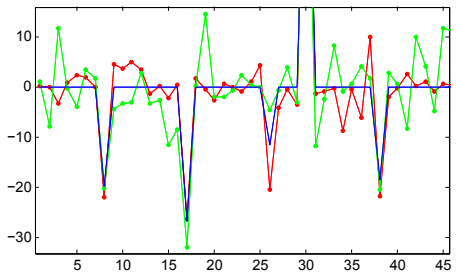


Fig. 1: A sample IEEE 118-bus model reconstruction undergoing post-processing Step 2 on column/row 30. Red and green dotted lines are the initial reconstruction of the 30th column and row, respectively; the blue line with four negative peaks shows the four actual non-zero off-diagonal entries of $\mathbf{H}$ for column/row 30.

defined as those within some percentage (threshold t_2) of each other. These are averaged, as in Step 1, for Property 1. Due to the possibility of adjacent non-zero entries, one of which would necessarily not appear as a local minima due to the other, this process is repeated, with identified non-zero entries temporarily raised to zero, until no new matching minima are found. Finally, of the potential non-zero values found, those are kept whose entries on the row and column are sufficiently close in magnitude to each other. The results are named $\hat{\mathbf{H}}_2$.

3) Merging $\hat{\boldsymbol{H}}_1$ and $\hat{\boldsymbol{H}}_2$: As mentioned, $\hat{\mathbf{H}}_1$ has a dearth of the near-zero values which Step 1 cannot discern from noise, some of which may appear in $\hat{\mathbf{H}}_2$. Also, while the requirement in Step 2 that derived non-zero values are close in magnitude between their row and column reconstructions is necessary to remove many false positives, it also removes some true non-zero values in $\hat{\mathbf{H}}_2$, which $\hat{\mathbf{H}}_1$ may contain. Consequently, we rewrite $\hat{\mathbf{H}}$ as $\hat{\mathbf{H}}_1$ and $\hat{\mathbf{H}}_2$ merged in a simple union. This is realized as any zero elements of $\hat{\mathbf{H}}_1$ which are non-zero in $\hat{\mathbf{H}}_2$ being replaced with the corresponding non-zero values from $\hat{\mathbf{H}}_2$. As any non-zero entries appearing in both $\hat{\mathbf{H}}_1$ and $\hat{\mathbf{H}}_2$ will necessarily be of the same magnitude, having been derived in the same manner (averaging over the diagonal), this has the same result as if $\hat{\mathbf{H}}_1$ and $\hat{\mathbf{H}}_2$ are reversed.

4) Row/column summation enforcement: Having now ensured the symmetry of $\hat{\mathbf{H}}$ and the negativity of its off-diagonal elements, Property 4 must next be satisfied, which requires each row and column to sum to zero. Step 4 iterates over the columns of $\hat{\mathbf{H}}$, finding for each the subset (of any cardinality) of non-zero off-diagonal elements which most closely sums in magnitude to the diagonal entry, such that the entire column sum is nearest zero. Elements outside of this subset are subsequently set to zero, in the column and also the corresponding transposed row to preserve Property 1. Obviously, this means that some elements of columns yet to be processed are decided by the results of previous columns, namely, those elements above the diagonal, which reflect over the diagonal to columns already processed. Each subsequent column hence contains a "safe set" consisting of its elements above the diagonal, which may not be zeroed.

The output of these four post-processing steps is an $\hat{\mathbf{H}}$ forced to satisfy the structure of the power flow model. Section V-B details the accuracy gains from post-processing, but briefly the number of correctly identified zero/non-zero entries increases by an average of 47.8% additively for the five IEEE test cases, and the entries' magnitudes become 44.1% more accurate.

C. Partial Model Derivation

Some attacks making use of the power flow model may not require the full model, and with some extension this attack method can reconstruct robust partial models using only a portion of the existing meters and only a portion of the desired number of measurement/state variable pairs specified in Section III-A. This imposes a still smaller requirement on the attacker while our case study (Section V-E) illustrates the utility of the partial model in false data injection attacks, finding that just a small portion of the model is sufficient to perform these attacks successfully with high probability.

With incomplete information, the model derivation process differs slightly. From Equation 3, again, we derive $\hat{\mathbf{H}}$ row by row, using $\mathbf{z}_i = \mathbf{h}_i\mathbf{X}$, where $\mathbf{z}_i$ and $\mathbf{h}_i$ are the i-th rows of $\mathbf{Z}$ and $\mathbf{H}$, respectively. We estimate $\mathbf{h}_i$ as before, but in the case of access only to a subset of power meters and corresponding state variables, we can determine each row i only if i is the index of a meter and state variable which is known. For example, if only meters and state variables 1 through 35 of a 118 bus system are available, we can only recreate rows 1 through 35 of $\mathbf{H}$. The result will simply differ in that $\mathbf{h}_i$ will only have $n-k$ entries which are presumably accurate, with the rest assigned zero.

In this manner, using some portion m_1 of m meter measurements/state variables and some portion n_1 of n sets of these, the same overall methodology presented in Section III-A results in $\hat{\mathbf{H}}'$ of size $m_1 \times n_1$. This is padded with zeros to form a correctly dimensioned $\hat{\mathbf{H}}$, but the non-zero values of $\mathbf{H}$ dimensionally outside of $\hat{\mathbf{H}}'$ are not recoverable. Due to this fact, Step 4 of the post processing method presented in Section III-B cannot be performed as it will generate incorrect results if actual non-zero values of $\mathbf{H}$ appear outside of the recoverable $\hat{\mathbf{H}}'$. The others remain applicable however, and able to improve the accuracy of $\hat{\mathbf{H}}$.

IV. Power Flow Model Camouflage

Problem space: Preventing power flow model information leakage will require protecting state variables and meter measurements wherever they appear. Communications from grid administrators to logging or reporting centers may include this information in its aggregate (which may be misused for this attack) for records of past performance or review in case of later unexpected situations. The aggregate information is thus the subject of our protection efforts, but an attacker's ability should not be ignored to compile the ingredients we have specified for deriving the full model or just a small portion. For example, meter measurements need not be retrieved from some repository if they can be viewed at the physical meters. In addition, power companies have been loath to provide

encryption at these endpoints due to the necessary addition of specialized hardware and software encryption solutions. This reluctance may fade with the continued development of the smart grid, but the DC systems addressed by this work are the larger backbone to the substations supplying the end user, and will consequently require some other mitigation technique for some time. For this reason, the ability to protect meter measurements through encryption should not be assumed.

In contrast, for transmission between components of the grid administration, the compiled meter measurements and state variables may be encrypted, as the endpoints of these communications can be presumed to have the necessary hardware and software. Simply encrypting the data will protect it but will offer no ability to detect that an attacker is eavesdropping. The sensitive environment of power systems security would benefit from that knowledge, however, as some attackers are likely supported by governments. Where a casual attacker would easily yield to failure, such a motivated attacker will have the resources necessary to exhaust this attack space and proceed to another. To effectively defeat this class of attacker is only possible through their identification and incarceration.

Our approach: We offer an opportunity to identify and prevent the attack before the attacker proceeds to another attack space, by providing a false indication of success. We reason that if an attacker is able to derive a fake power flow model $\hat{\mathbf{H}}_f$ having all the correct characteristics, an attack based thereupon will both fail and alert the grid administrators. In the case of FDIAs, the attack vector will not pass bad data detection, and existing methods for discovering faulty meters will in so doing physically locate the attacker (who must physically compromise the applicable meters [3]).

An attacker using our power flow model derivation technique will need pairs of meter measurements and state variables, but these should be altered in some way so as to lead to the solution of $\mathbf{H}_f$ instead of $\mathbf{H}$. However, legitimate administrative duties involving the sending and receiving of real measurements and variables should persist under this camouflage, so the data sent should contain the real $\mathbf{Z}$ and $\mathbf{X}$, irretrievable to attackers but not to authorized personnel. This may be achieved by creating for the attacker to find an $\mathbf{H}_f$ equal to $\mathbf{H}$ multiplied with some matrix $\mathbf{F}$, effectively camouflaging $\mathbf{H}$ for all those who do not know $\mathbf{F}$. This $\mathbf{F}$ will also be used to encode/decode $\mathbf{Z}$ and/or $\mathbf{X}$, depending on how $\mathbf{F}$ is multiplied with $\mathbf{H}$. Multiple options exist, encoding $\mathbf{Z}$, $\mathbf{X}$, or both, but the meter measurements $\mathbf{Z}$ should be untouched. Because encryption of power meters is not immediately possible, the attacker can presumably view the measurements at the physical meter locations. If these values do not match the attacker's eavesdropped data, the camouflage will be obvious.

Accordingly, we inject $\mathbf{F}$ into Equation 3 as:

$$\mathbf{Z} = (\mathbf{HF})\mathbf{F}^{-1}\mathbf{X}. \tag{4}$$

In Equation 4, $\mathbf{H}_f = \mathbf{HF}$ will be the modified form of $\mathbf{H}$ which the attacker will derive instead of the true $\mathbf{H}$, using the unchanged $\mathbf{Z}$ and the modified $\mathbf{X}_f = \mathbf{F}^{-1}\mathbf{X}$. Without knowing $\mathbf{F}$, it will be impossible to discern the real $\mathbf{H}$, but with $\mathbf{F}$ as a pre-shared secret among authorized parties, the original meter measurements and state variables will be retrievable. We find that it takes a very specific yet not generalizable $\mathbf{F}$ to form $\mathbf{H}_f$ with the correct properties of the DC power flow model, so we obtain $\mathbf{F}$ from a pre-contrived $\mathbf{H}_f$ and the real $\mathbf{H}$ as $\mathbf{F} = \mathbf{H}^{-1}\mathbf{H}_f$. Creating this fake power flow model $\mathbf{H}_f$ is as simple as constructing a sparse matrix satisfying the properties of a real model as enumerated in Section III-B. A convincing fake will also have non-zero values in the same distribution as the real non-zero values, but in different locations; we give a sample method of constructing a realistic $\mathbf{H}_f$ in Section V-D where we evaluate its ability to disrupt the derivation process.

Ultimately, by encoding $\mathbf{X}$ as above and leaving $\mathbf{Z}$ as is, the attacker may solve $\mathbf{Z} = \mathbf{H}_f\mathbf{X}_f$ and discover the fake power flow model $\mathbf{H}_f$, which will be unsuitable for topology or model-based attacks. We verify the effects of model camouflage in Section V-D and in our case study on FDIAs in Section V-E.

V. Evaluation and Case Study

A. Setup

We assess our derivation of the power flow model on five IEEE test cases provided in a collection of MATLAB code entitled MATPOWER. MATPOWER simulates power flow calculations in AC and, in our case, DC systems, and provides sample data for power systems including 9, 14, 30, 118, and 300-bus cases, which we examine here. However, this case data only comprises the system in one state, while we require pairs of meter measurements and state variables to form our matrices $\mathbf{Z}$ and $\mathbf{X}$. Also, MATPOWER does not perform state estimation, due to no simulation of meter measurements and their errors, but calculates the state variables (voltage angles) directly as part of the power flow analysis.

To create the collection of state variable sets $\mathbf{X}$, we perturb each of the calculated state variables to form n slightly different copies, by adding to each variable random Gaussian noise of magnitude based on a specified noise level (0.1). This formulation corresponds to the case where the power grid state does not fluctuate wildly for some time, that is, when generation and load are fairly constant throughout the grid. This scenario is commonplace during night hours when most people are sleeping or weekday morning hours when most are working, for example. We then create the collection of n measurement sets $\mathbf{Z}$ by the equation $\mathbf{Z} = \mathbf{HX}$. Now, $\hat{\mathbf{H}}$ may be calculated directly from $\mathbf{Z}$ and $\mathbf{X}$, which is not possible in practice, so we add noise to each entry of $\mathbf{Z}$ to simulate measurement error, using random Gaussian noise as before.

To simulate systems under different supply and demand scenarios is prudent for a thorough understanding of the applicability of this attack in different environments, so we also construct new case data for each of the five aforementioned system sizes. Here, we alter the generation capacity for the generators in the system, as well as branch admittances and loads. Practically, this is done by filling out the MATPOWER case structure with values reproducing the distribution of values in the provided cases. These are held within the necessary

Case	No Post-Processing	Step 1 only	Step 2 only	Step 3	Step 4
9	89.3% (37.0%)	69.8% (4.3%)	74.5% (10.6%)	69.8% (4.3%)	70.1% (4.1%)
14	90.4% (30.1%)	76.1% (4.3%)	69.1% (7.1%)	76.1% (4.3%)	71.6% (3.2%)
30	91.1% (27.0%)	71.0% (1.7%)	60.2% (2.7%)	71.0% (1.7%)	70.0% (1.3%)
118	93.2% (32.0%)	66.0% (0.37%)	34.4% (0.29%)	66.0% (0.37%)	65.4% (0.28%)
300	86.9% (5.3%)	64.5% (0.04%)	26.9% (0.0%)	64.5% (.04%)	58.0% (0.02%)

TABLE I: Power system topology reconstruction accuracy: valid connections found (M_1) and wrongly identified connections (M_2, parenthetical)

Case	No PP	Step 1 only	Step 2 only	Step 3	Step 4
9	37.2%	49.5%	49.0%	49.5%	49.9%
14	39.0%	62.1%	61.6%	62.1%	62.1%
30	27.6%	68.5%	65.0%	68.5%	69.1%
118	0.0%	70.7%	52.3%	70.7%	71.9%
300	58.0%	91.8%	87.3%	91.8%	92.4%

TABLE II: Power flow model reconstruction accuracy: similarity of $\mathbf{H}$ and $\hat{\mathbf{H}}$ (M_3)

constraints, such as power drawn from a generator being limited to the range supported by that generator. Unique power flow analyses may then be performed on these constructed cases to provide different power flow models to test.

A variety of approaches could be taken to optimize the thresholds used in post-processing the results of the initial reconstruction. We focus on the topology accuracy and calculate a confusion matrix, minimizing the sum of the false positives (identified connections which do not exist) and false negatives (actual connections undiscovered) to optimize. Depending on an attacker's goals one or the other type of error could be given precedence. To present resultant accuracies, we employ three metrics, the first two describing topology accuracy, and the third representing the overall accuracy of the reconstructed power flow model. Working from the confusion matrix, the first metric M_1 is the percentage of valid connections found, and the second metric M_2 is the percentage of unconnected buses incorrectly derived as connected ("false connections"):

$$M_1 = \frac{TP}{TP+FN}, M_2 = \frac{FP}{FP+TN}$$

Then, an accurate knowledge of the power flow model in its entirety relies upon the non-zero values having the right magnitude, to correctly describe the impedance on each branch. The third metric M_3 is defined as the total difference between $\hat{\mathbf{H}}$ and $\mathbf{H}$, standardized as a percentage of the total magnitude of $\mathbf{H}$, and subtracted from 1 to indicate similarity. That is,

$$M_3 = 1 - \frac{\sum_{i,j} |\hat{\mathbf{H}}_{ij} - \mathbf{H}_{ij}|}{\sum_{i,j} |\mathbf{H}_{ij}|}$$

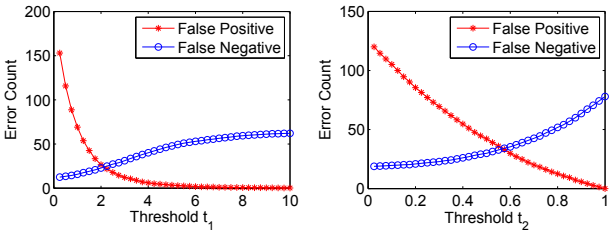
In all further figures and tables, Topology accuracy refers to a maximal M_1 and minimal M_2, and Full Model accuracy refers to a maximal M_3.

B. Power Flow Model Derivation Accuracy Optimization

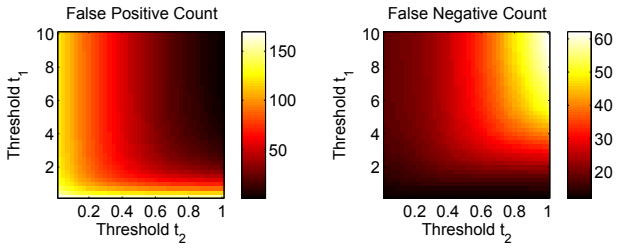
For each of our five test case sizes, we generate ten power flow models as described in our Evaluation Setup (Section V-A). We then generate paired measurements and state variables matrices $\mathbf{Z}$ and $\mathbf{X}$ ten times for each model. Using these 100 $\mathbf{Z}$ and $\mathbf{X}$ pairs, we test our post-processing methods and vary the two thresholds involved, to verify the methods work to refine the results, and find the thresholds which best achieve this. For comparison, we present the accuracy obtained by the bare reconstruction process with only a small threshold applied to take near-zero values to zero. (Without this threshold, false positives - non-zeros, or wrongly hypothesized connections - will be nearly 100%, so we consider this "no post-processing" in our reported numbers for more informative comparison.) We then test the post-processing as Step 1 only, Step 2 only, Steps 1 and 2 combined (Step 3), and Full post-processing (Steps 1-4). Steps 1 and 2 each have one threshold to optimize, so in performing these tests we optimize whichever one or both are applicable. We illustrate in Figure 2 this process for the IEEE 30-bus case, but report all results in Tables I and II.

Without post-processing, 91.1% of valid connections are found for the 30-bus case. While this is high, 27.0% of unconnected pairs of buses are falsely identified as connected, so the overall model accuracy is only 27.6%. For reference, the sparsity of the IEEE 30-bus case is roughly 88%, so a 27.0% false positive rate results in around 213 invalid hypothesized connections, which is far more than the 112 valid connections. Minimizing the false positive rate is of high importance, hence the post-processing. Threshold optimization for Step 1, as shown in Figure 2a, finds the lowest error sum with threshold $t_1 = 3$, resulting in 71% of valid connections found and a much-reduced 1.7% of non-connections assumed connected. The optimization for Step 2 is performed similarly to Step 1, with a threshold $t_2 = 0.7$ resulting in the lowest error, as illustrated in Figure 2b. Figures 2c and 2d show the optimization of both thresholds for Step 3 and Step 4, with the heat maps representing the same threshold optimization process but in two dimensions for the two applicable thresholds. Ultimately, with thresholds $t_1 = 2.75$ and $t_2 = 1.0$, the lowest cumulative error is found for Step 4, with 70% of valid connections found and only 1.3% of non-connections mis-attributed.

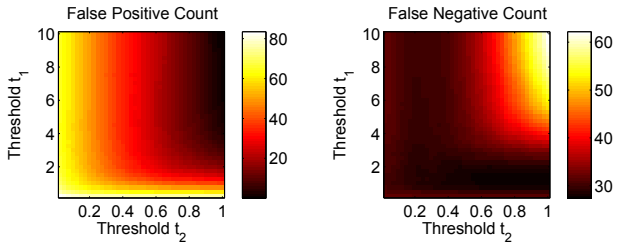
Examination of the achieved accuracies shows a higher overall model reconstruction accuracy found for larger systems. Table II exhibits this trend clearly, though it may be expected that larger systems would be more difficult to reconstruct. However, compressive sensing requires a certain degree of sparsity for an accurate reconstruction of the data [9]. The smaller cases are far less sparse than the larger cases, and so are necessarily harder to reconstruct. Additionally, while topology reconstruction accuracy decreases slightly for the large systems (Table I), it is important to note the tiny amount of wrongly identified connections for those systems. With the sparsity of the 300 bus power flow model around



(a) Post-processing Step 1 (b) Post-processing Step 2



(c) Post-processing Step 3 (Steps 1 and 2 combined)



(d) Post-processing Step 4 (full post-processing)

Fig. 2: Threshold optimization for each of the four post-processing steps in the IEEE 30-bus test case

98%, this corresponds to an average of less than five wrongly identified connections. Optimizing the two thresholds as we have is logical for an attacker interested in the topology of the power grid, ensuring that while not all of the grid topology is discovered, few identified connections are fictitious.

C. Partial Model Derivation Accuracy

We again generate ten power flow models for each of our five test case sizes, this time only creating one pair of meter measurements $\mathbf{Z}$ and state variables $\mathbf{X}$ for each. Then, we experiment with reconstructing differing submatrices of $\mathbf{H}$ for each of these ten power flow models. In this experiment, all tested submatrix sizes begin at the top left entry and extend to some varied percentage of rows and columns based on the percentage of meter measurements and state variables simulated as available. Consequently, in the heat maps present in Figure 3, each entry corresponds to the accuracy of reconstructing the submatrix of $\mathbf{H}$ starting at the top left element and extending to that entry (normalized by percentage of data used). Only the 300-bus case is shown, due to space limitations, but it is representative of the trends visible in the other cases. Its heat maps are broken up into 4% partitions in each direction, for a fine-grained division of its large number of buses, while the smaller cases have fewer partitions.

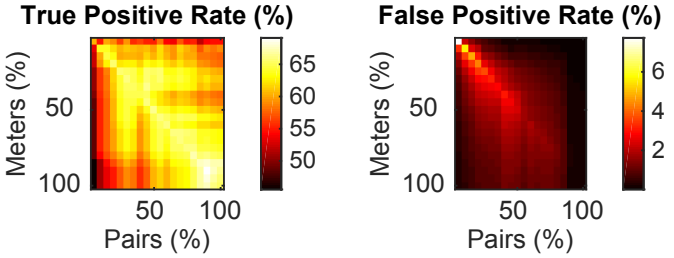


Fig. 3: Partial reconstruction accuracy for the IEEE 300 bus test case, representative of the rest. Each accuracy entry on the heat map corresponds to the partial reconstruction with dimensions starting in the top left and spanning to that entry.

The heat maps in Figure 3 depict accuracy measured according to the two metrics M_1 and M_2, as before representing the percentage of valid topological connections found and the percentage of unconnected buses incorrectly presumed to be connections, respectively. All post-processing is used except for Step 4, because with incomplete rows/columns, it cannot be assumed that non-zero values will not appear outside of the recoverable $\hat{\mathbf{H}}$. Consequently the thresholds t_1 and t_2 are set to the values optimizing Step 3 as shown in Tables I and II. The true positive heat map exhibits a trend of more valid connections found nearing 100% of data used, where it is comparable to the values reported in those tables for Step 3.

Viewing the false positive heat map, portions of the model having disproportional numbers of columns to rows tend to have a lower rate of false connections. This is largely due to the greater proportion of zero to non-zero elements, causing a larger denominator for the false connections metric M_2. As visible in the corresponding valid connections heat map, M_1 suffers for those more rectangular model subgraphs. Also, the corresponding full model accuracy (metric M_3, not pictured) is lower for these as well, because an imbalanced ratio of rows to columns means there is less data for the post-processing steps to work with in refining the actual values. As the unprocessed $\hat{\mathbf{H}}$ is not naturally symmetric, Steps 1 and 2 are designed to recreate the symmetry of $\mathbf{H}$, but if the portion examined has more rows than columns (or vice versa) then Step 2 will not be able to address the extra rows (or columns). This is also why the heat maps are not perfectly symmetric while $\mathbf{H}$ is.

D. Derivation Accuracy under Model Camouflage

The camouflage process begins with the construction of a fake power flow model $\mathbf{H}_f$. As this needs to be a convincing fake, beyond fitting the properties of a power flow model as enumerated in Section II it should have values of the same order as a real model, and it should have a similar degree of sparsity. Consequently, in constructing $\mathbf{H}_f$ for a given test case, we derive a list of values reproducing the distribution of off-diagonal values in $\mathbf{H}$ for that case, and with these fill the strictly upper triangular portion of $\mathbf{H}_f$. These values are added at random according to a probability matching the number of off-diagonal values in $\mathbf{H}$, to maintain a similar degree of

Case	Without Camouflage		With Camouflage	
	Topology	Full Model	Topology	Full Model
9	70.1%	49.9%	42.1%	14.4%
14	71.6%	62.1%	40.0%	16.1%
30	70.0%	69.1%	30.3%	8.1%
118	65.4%	71.9%	31.0%	11.7%
300	58.0%	92.4%	42.0%	0.0%

TABLE III: System topology and power flow model reconstruction accuracy without and with the effects of camouflage

Case	Size of $\hat{\mathbf{H}}$ Compared to $\mathbf{H}$	#Meters to Compromise	State Variable Differences	
			Mean	Maximum
9	25%	2	4.2%	14.2%
14	11.1%	2	0.5%	0.8%
30	6.25%	3	4.2%	23.5%
118	6.25%	10	0.9%	3.4%
300	6.25%	25	8.9%	498.1%

TABLE IV: Attacker requirements and resulting average and maximum state variable alterations

sparsity. The upper triangular portion is mirrored over the diagonal to instill symmetry, and finally the values of each row/column are summed with the absolute value of this sum set as the diagonal. The resulting $\mathbf{H}_f$ is now a valid power flow model, but for a system that does not exist.

We again generate ten power flow models for each of our five test cases, each with a pair of $\mathbf{Z}$ and $\mathbf{X}$. Then, each model is given camouflage with $\mathbf{H}_f$ calculated as above. Stemming from Equation 4, we then generate a fake set of state variables by $\mathbf{X}_f = \mathbf{F}^{-1}\mathbf{X} = (\mathbf{H}^{-1}\mathbf{H}_f)^{-1}\mathbf{X} = \mathbf{H}_f^{-1}\mathbf{H}\mathbf{X}$, while $\mathbf{Z}$ is left as is. The model derivation process is launched for each model, using full post-processing with the thresholds optimized in Section V-B. The corresponding highly inaccurate results are displayed in Table III in contrast to the more accurate results without camouflage. The percentage of topology accuracy that does appear during camouflage owes essentially to the portion of non-zero values which comprise the diagonal of every power flow model, and is therefore unavoidable from a defense standpoint, but meaningless to the attacker. For example, in the 118-bus system, an average of 143.4 non-zero entries of $\mathbf{H}$ were found, but 118 of them are the diagonal elements and an additional average 319.6 non-zero entries were camouflaged. Furthermore, the full model accuracy is extremely low, especially for large systems with more off-diagonal values. The following case study shows false data injection attacks using a camouflaged fake model largely fail.

E. Case Study: False Data Injection Attacks Using A Derived Power Flow Model

The probability of successfully carrying out FDIAs is derived in the original source material [3], but this is again based on the assumption of having the correct power flow model. In this research, the power flow model is calculated, and not without error, and so its suitability for a model-specific attack such as this should be tested. Furthermore, we examine the efficacy of carrying out these attacks using only a portion of the existing meters and only a portion of the desired number of measurement/state variable pairs. We lastly illustrate the results of this attack when the model is under camouflage.

To reduce the impact of sporadic large noise in meter measurements or their malicious alteration, the bad data detection process first identifies and removes any measurements deviating strongly from their expected values. The calculation common to the various existent methods is a residual error comparison between the vector of observed measurements and that of hypothetical measurements calculated using the estimated state variables [4]. That is, after generating the estimate $\hat{\mathbf{x}}$ of state variables $\mathbf{x}$, the expected measurements are calculated as $\hat{\mathbf{z}} = \mathbf{H}\hat{\mathbf{x}}$ and compared with the actual measurements $\mathbf{z}$. If the 2-Norm of this difference (the residual) is below a specified threshold, the measurements are assumed correct, as are the corresponding estimated state variables. False data injection attacks form an attack vector $\mathbf{a}$ as a linear combination of some columns of $\mathbf{H}$, which is added to $\mathbf{z}$ by corrupting the meters corresponding to non-zero elements in $\mathbf{a}$ [3]. An attack vector may be generated in the same way using $\hat{\mathbf{H}}$ (as a padded $\hat{\mathbf{H}}'$), but as $\hat{\mathbf{H}}$ is merely an estimation of $\mathbf{H}$, this attack vector is not guaranteed to pass bad data detection.

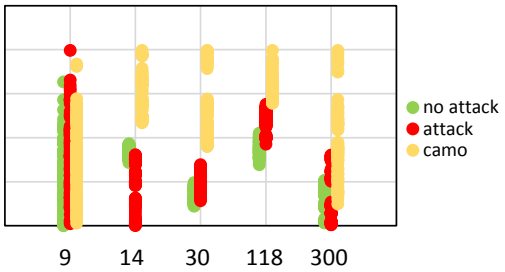


Fig. 4: Normalized error residuals without and within the presence of false data injection attacks using the derived power flow model $\hat{\mathbf{H}}$, for each of the IEEE test cases.

Nevertheless, this research finds these attacks very viable with this limited knowledge. Figure 4 illustrates the residuals calculated when the system is provided unaltered meter measurements ("no attack") and those having been modified using $\hat{\mathbf{H}}$ ("attack"), for each test case and normalized for display. To set a threshold during bad data detection allowing for all of the legitimate residuals will evidently also allow for a large majority (75%) of the compromised measurement residuals to pass as well; compromised measurements cannot be distinguished from natural measurements.

Table IV states the requirements on the attacker to perform the FDIAs in Figure 4, as well as the resulting impact of these attacks on the accuracy of the state variables. These results use the default test cases with $\mathbf{Z}$ and $\mathbf{X}$ pairs generated ten times per, and an attack vector formed by a linear combination of the first third of available reconstructed columns, with random weights and constructed ten times per $\mathbf{Z}$ and $\mathbf{X}$ pair. This means the average state variable difference is an average over 100 attacks, per case, and the maximum variable difference is the average of each attack's maximum effect. Evidently, at

least one state variable was altered by a factor of five, on average, during 100 attacks against the 300-bus case, and they were all, on average, modified by 9%.

Returning to Figure 4, the residuals labeled "camo" are those calculated when the power flow model is under camouflage. Except for the 9-bus case and 28% of the 300-bus case, these residuals are of clearly different magnitudes than those for the unaltered meter measurements. The small size and relative lack of sparsity for the 9-bus system indicates that it may not be an appropriate candidate for application of camouflage, but the larger systems clearly benefit. Indeed, all attacks against 14, 30, and 118-bus systems are prevented.

VI. Related Work

Despite extensive research into FDIAs against state estimation and their limits and impacts, the underlying requirement for the attacker to know the power flow model remains largely unexplored. The potential is similarly unexplored for the power flow model, or the grid topology contained therein, to be misused in ways other than the former with FDIAs. Deriving the model is deemed impossible using just a collection of measurements in [11], but the topology of the system is attained using a joint estimation of both the model $\mathbf{H}$ and the state variables $\mathbf{X}$, made convex using an iterative method switching between these two. Having been tested only on the IEEE 14-bus system simulation data, and presenting the derivation accuracy largely graphically, [11] is not particularly tenable. In addition to the more comprehensive test cases, our research estimates only the power flow model $\mathbf{H}$, which is more computationally expedient, and we achieve this as well as leaking the network topology with high accuracy.

In parallel publications [12] and [13], the authors perform a re-creation of the power flow model with a very similar theme as the previously discussed related work. As publicly available local pricing calculations are performed using the power flow model, a collection of prices local to every meter were used rather than their measurements. Analogous to [11], these authors perform a joint estimation of the model and another component used in the price estimation. In this case, the authors state the power flow model as the output of this process, though the accuracy is again only graphically represented, and only for the IEEE 30-bus case. From this view it appears the topology information is accurate, but the power flow model information is less so. We accordingly strive for more thorough testing and concrete portrayal of results.

The problem of network topology derivation on the part of power grid administrators has spawned slightly more work recently, for the sake of performing accurate state estimation in an environment that changes with time. Recent trends toward distributed generation on the part of customers adding solar power to their homes cause topological changes grid administrators need to incorporate, as state estimation is dependent upon these aspects they do not govern. Frequency domain reflectometry is used over the power line communications channel by [14], which has the same topology as the power grid, to determine lengths of branches and where they separate. A similar strategy is used in [15], instead compiling end-to-end distance measurements and working out distances of interior nodes from these. These efforts focus on small "micro grids" with the latter requiring considerable foreknowledge. Another work suggests an announcement protocol to be carried out when a new endpoint is added to the topology, for a decentralized approach [16]. Clearly, these ideas will not facilitate a leak of the power flow model by an attacker.

VII. Conclusion

This research illustrates an attacker's ability to reverse engineer a large portion of DC power flow model to an accuracy of 69.1% (averaged across 5 IEEE system sizes), including 67.0% of its topology, and illustrates its use in performing false data injection attacks against state estimation. The residual error caused by false data injection attacks using this reconstruction is well hidden within the inherent error, but demonstrate a novel camouflage technique able to prevent them 93% of the time in systems other than the 9-bus case, by reducing the inferred system model accuracy to 10.1%.

Acknowledgments

This paper is supported by the Air Force Research Lab's Visiting Faculty Research Program.

References

[1] S. Rielly and R. Sabalow, "Bracing for a big power grid attack: 'one is too many'."

[2] A. Wood and B. Wollenberg, *Power Generation, Operation, and Control*. A Wiley-Interscience publication, Wiley, 1996.

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

[4] A. Monticelli, *State Estimation in Electric Power Systems*. Kluwer's Power Electronics and Power Systems Series, Springer, 1999.

[5] J. Arrillaga, *High voltage direct current transmission*. No. 29, Iet, 1998.

[6] G. Andersson, "Modelling and analysis of electric power systems," *ETH Zurich, september*, 2008.

[7] J. Grainger, *Power system analysis*. New York: McGraw-Hill, 1994.

[8] E. J. Candè and M. B. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21–30, 2008.

[9] E. van den Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[10] E. van den Berg and M. P. Friedlander, "SPGL1: A solver for large-scale sparse reconstruction," June 2007.

[11] X. Li, H. Poor, and A. Scaglione, "Blind topology identification for power systems," in *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, pp. 91–96, Oct 2013.

[12] V. Kekatos, G. B. Giannakis, and R. Baldick, "Online energy price matrix factorization for power grid topology tracking," *arXiv preprint arXiv:1410.6095*, 2014.

[13] V. Kekatos, G. Giannakis, and R. Baldick, "Grid topology identification using electricity prices," in *PES General Meeting— Conference & Exposition, 2014 IEEE*, pp. 1–5, IEEE, 2014.

[14] M. Ahmed and L. Lampe, "Power line network topology inference using frequency domain reflectometry," in *Communications (ICC), 2012 IEEE International Conference on*, pp. 3419–3423, June 2012.

[15] L. Lampe and M. Ahmed, "Power grid topology inference using power line communications," in *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, pp. 336–341, Oct 2013.

[16] N. Honeth, A. Saleem, K. Zhu, L. Vanfretti, and L. Nordstrom, "Decentralized topology inference of electrical distribution networks," in *Innovative Smart Grid Technologies (ISGT), 2012 IEEE PES*, pp. 1–8, Jan 2012.