RESEARCH

Geographic coordinate validation and assignment using an edge-constrained layout

Adam B. Birchfield[1*]

*Correspondence:
abirchfield@tamu.edu

[1] Department of Electrical
and Computer Engineering,
Texas A&M University, 188 Bizzell
St., College Station, TX 77840,
USA

Abstract

Electric grids with buses that are mapped to geographic latitude and longitude are useful for a growing number of applications, such as data visualization, geomagnetically induced current calculations, and multi-energy coupled infrastructure simulations. This paper presents a methodology for validating the quality of geographic coordinates for a power system model, and to assign coordinates to buses with missing or low-quality coordinates. This method takes advantage of geographic indicators already intrinsic to a grid model, such as branch length as implied by impedance and susceptance parameters. The coordinate assignment process uses an approach inspired by graph drawing, that lays out the vertices (buses) and edges (transmission lines), formulated as a nonlinear programming problem with soft edge length constraints. The layout method is very computationally fast and scalable to large power system cases. The method is demonstrated in this paper using a 37-bus test case and a 6717-bus test case, both publicly available, along with a large actual grid model. The results show that, for cases with only a few errors in the coordinates, cases with no coordinates known beforehand, and others in between, this method is able to assign reasonable geographic coordinates to best match known data about the grid.

Keywords: Geographic coordinates, Power system visualization, Graph drawing, Geographic information systems (GIS), One-line diagrams

Introduction

Geographic coordinates are not directly necessary to solve power flow solutions, optimal power flow, and transient stability simulation on electric power grids. Hence for the sake of simplicity and the economy of data storage, traditionally power flow cases have not contained information about the latitude and longitude of the physical substations in which the electrical buses are located.

There is a growing trend, however, in recent times, toward more cases including geographic coordinates, for several reasons. First, data visualization: geography is a natural starting point for representing a power system in a single-line diagram, or showing other data that varies over a wide area [1–3] (though not the only way [4]). Second, geomagnetically induced current (GIC) analysis, such as for geomagnetic

disturbances (GMD) and electromagnetic pulse (EMP), requires geography to compute the impact of wide-area electric fields [5–7]. Third, geographic embedding of power flow cases opens up opportunities to coordinate the analysis with other co-located information, such as locational weather data (especially cloud coverage and wind speeds for renewable integration) [8, 9], communications networks [10], natural gas pipeline networks [11], transportation [12], and water [13]. As a result, there are some examples of regions in the North American Electric Reliability Corporation (NERC) that require submitting geographic coordinates for some applications related to network planning [14, 15].

A common challenge in working with wide-area electric transmission grid analysis is that, in many cases, engineers and researchers do not have readily available mapping of buses to high-quality geographic coordinates. In some cases, partial coordinates exist for a region of the case, or for the highest voltage network. In other cases, the coordinates are given at a very low level of precision. At a minimum, usually there are some buses in a system for which the coordinates are not given or are incorrect. If not properly considered, flagged, and if possible corrected, errors will propagate into analysis methods such as GIC calculations, leading to wrong conclusions, and data visualization may be misleading or not look right.

In this paper, we present a method to evaluate and improve the quality of a geographic embedding for an electric transmission system dataset, using information already intrinsic to the power flow case. The first part of this work presents validation metrics to assess a set of geographic coordinates, whether estimated from an algorithm or provided in advance. The metrics show ways in which the power flow data (specifically the transmission line and other branch parameters that indicate the length of the line) are or are not consistent with the given geographic coordinates. These metrics allow for assessing not only the quality of the geographic embedding as a whole but also the flagging of specific substation and line data that may contain errors.

The second, related contribution of this paper is that we introduce a graph-layout-based methodology to assign new geographic coordinates to some or all buses that will match the underlying case data and satisfy the validation metrics well. This task is formulated as a nonlinear optimization problem with soft constraints that can be solved with an interior point solver quickly even for large systems. This method can apply whether most buses already have assumed coordinates or no coordinates are known at all, or anywhere in between. The paper demonstrates the effectiveness of these methods on example cases for different scenarios up to 6717 buses.

There are a number of potential benefits of this work. First, it can aid error detection and correction in network planning and the development of power flow cases. This would include correcting substation mapping and geographic coordinates, but in some cases could include updates to the power flow data itself if, for example, a line's shunt susceptance was flagged as too large for its (correct) geographic length. Second, in cases where low-quality or limited geographic information is available (or at least where some buses are not mapped), this method provides a quick way to get coordinates estimated for the rest of the buses, to allow for analysis that requires geography such as GIC simulations. Third, it supports the creation of better one-line diagrams and other data visualization, even for cases that have no pre-assigned geographic coordinates, by creating an

initial set of coordinates that approximate the underlying actual geographic coordinates in the sense that they are consistent with the power flow data.

The outline of this paper is as follows. After a background survey of related work ("Background" section), the proposed methodology is presented in two parts: first, the framework for validating coordinates and identifying errors ("Assessing the quality of geographic mappings for electric grid cases" section), and then the optimization-based algorithm for laying out bus coordinates optimally ("Length-constrained graph layout" section). Demonstrations of applying the method to a variety of realistic scenarios and showing the method's effectiveness are given ("Results and discussion" section), wrapping up with some concluding thoughts ("Conclusions" section).

Background

Digital geo-mapping of energy infrastructure data has its origins in the mid-twentieth century with advances in computational power and the advent of geographic information systems (GIS), [16, 17]. As has already been mentioned, interest in GIS for power systems includes the ability to correlate with other geo-mapped data [8–13], which has implications for planning and operations, particularly with the growth of distributed generation (as in [18]) and increased attention to recovery from natural disasters [19].

Many of the efforts in the power system literature toward validating and improving geographic coordinates have been relatively recent and targeted at the distribution level. In [20], the researchers take advantage of a larger volume of smart meter data and the observation that voltages will be correlated with GIS information. Similar to other efforts to identify network topology and load phase connections, the voltage patterns over time can show errors in GIS data. In [21], graph theory processes are used to detect GIS errors for distribution, and similarly, in [22], the objective is to find errors in GIS data at the distribution level (particularly at the secondary level near the customers), using analysis that considers GIS data, network topology data, and customer physical addresses with a clustering-based procedure. A related effort applies image processing algorithms to improve system mapping [23].

A strong motivation for better transmission-level geographic coordinate mapping is improved visualization. Good diagrams augment system analysis data with a visual context and help engineers and others better diagnose problems and communicate results effectively. Graph drawing, generically, is a challenging problem, because in order to represent a network of edge-vertex relationships on a two-dimensional plane, the vertices must be assigned cartesian coordinates that match a number of visual metrics that are often directly in tension [24–26]. For large-scale graphs, one family of visualization techniques is the force-directed approach, where graphs are modeled as physical systems with spring attraction along edges and electro-static repulsion between neighboring vertices [27]. Another technique in generic graph drawing at a large scale is modeling the system with a hierarchical structure [28]. With specific reference to electric grids, extensive effort has gone into automating transmission system visualization, with some of the earliest work recognizing that a unique aspect is that there are local substation diagrams that are then connected over a wide area [29]. Automatic network layout algorithms include [30–34] and the author's work in [1], which use a variety of methods but typically involve either a modified force-directed approach, geographic coordinates as

a baseline, or both. Work that specifically looks at visual quality without regard to geographic layout includes [35, 36]. These layout methods can apply not only to network diagrams but also to visualization of other datasets, as in mosaic tile displays [37, 38]. In addition, recent work has shown additional work in automatic network layouts using parallel fast methods [39] and linear programming [40].

The ubiquitous IEEE test cases did not originally contain geographic coordinates [41], although at least one recent effort has assigned coordinates to the RTS-96 case after the fact [42]. In more recent development of electric transmission system test cases, there has been more of a focus on including them. Although much actual power system data is not available for public release due to its designation by the US Federal Energy Regulatory Commission (FERC) as Critical Energy Infrastructure Information (CEII) [43], some information merely about the location of critical infrastructure is more widely accessible, such as the location of generators greater than 1 MW from the EIA form 860 [44]. New public test cases that include geographic information include the 20-buse GIC test case [45], recent large-scale synthetic grids [46, 47], and the California test system targeted at extreme weather and wildfire studies [48].

Assessing the quality of geographic mappings for electric grid cases

How accurate geographic coordinates need to be in an electric grid case, and the consequences of inaccuracies, depends on the application. For wide-area system network visualization, small errors in mapping or location will not be noticeable. In fact, a bit of distortion may be introduced intentionally to show the electric structure more clearly. But any substations not mapped, or those mapped drastically wrong, will be either missing from the diagram or a distraction making it appear as if transmission lines are cutting thousands of miles across a case. For visualization applications in conjunction with other infrastructure and particularly with satellite or mapping datasets, a higher level of precision will be required.

GIC applications have a very similar pattern. Given the amount of uncertainty in the input data to GIC studies (see [49]), high levels of coordinate precision are not required. However, because the GIC levels are highly related to the length of the line, it is important to keep the geographic length of the line relatively consistent with the line's actual length and in the same general region. Uncaught major mapping errors can unintentionally inject large currents into the network and throw off the results.

Before continuing, one note should be made about geographic projections. Coordinates are usually given as degrees of latitude and longitude, which correspond to locations on the spheroid representing Earth. For the purposes of this paper, such coordinates are converted to planar cartesian coordinates using the Universal Transverse Mercator (UTM) projection [50, 51], so that they are given as x, y where x is the "easting" in meters and y is the "northing" in meters. This allows direct calculation of distances. Across the scale of even the largest grids, errors in using UTM are significantly smaller than the typical size of a substation, not to mention the other sources of uncertainty in power system data. For "Length-constrained graph layout" section, coordinates are generated in x, y, and then are projected back into latitude and longitude by the inverse UTM projection. This is merely a choice of convenience; other projections such as state plane coordinate systems could be used as well—but using latitude and longitude as

if they were cartesian coordinates is not good because the distance metrics would be invalid.

The rest of this section outlines the validation analysis for a case with some given geographic coordinates. First, the analytical observations are given; then, they are quantified into metrics that are used to assign quality flags to various power system data. The quality flag integer variable q is defined such that $q = 0$ indicates zero confidence in the accuracy of the associated data, with higher values of q indicating better data. The maximum value of q is 3 for branches and 5 for buses.

Given geographic coordinates and substation mapping

Ordinarily, geographic coordinates are not assigned directly to buses, but buses are identified with an associated substation, which in turn has a geographic latitude and longitude (converted then to x, y for this analysis as described above). The first observation to be made in validating coordinates is that some coordinates can be immediately identified as incorrect. The validation process starts with finding the middle location of the system, defined as the median latitude and median longitude. All valid coordinates should be within a certain radius of that spot (depending on the known size of the system, say 1000 km). In particular, coordinates with (0, 0) are obviously missing. Any coordinates outside the acceptable range are marked with $q = 0$ from the beginning.In addition to this, another indicator is the apparent decimal points of precision of data. While it is possible that the substation may be exactly located on a whole-number line of latitude and longitude, more probably this is an indication that the quality is low to begin with.

Once these preliminary indicators have been assessed, the attention of the validation method turns to the network branches and their apparent length, $\ell = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ (where the two buses are located at (x_1, y_1) and (x_2, y_2)). Ultimately, the validation of the bus geographic coordinates (absent other data like checking satellite imagery) is dependent upon the bus's relationship to other known coordinates via the branches. Buses not connected to any branches cannot be further validated, but these play little role in the system.

Network branches that are not transmission lines

Most branches in a bus-branch power system model represent either lines or transformers. Transformers are often directly labeled as such, but if not they can be quickly identified as those branches which connect buses that are labeled with different nominal voltage levels. Transformers ought to have a length of essentially zero (in some cases a transformer might include a small line to a neighboring substation and may have a short distance associated with it). There are also a (usually small) number of branches which do not represent transformers but also are not ordinary transmission lines. Although they connect two buses of the same voltage level, they are short connections within a substation and hence should also have a length of essentially zero. These can be sometimes difficult to distinguish from actual transmission lines, but can usually be identified by relatively low reactance (X), in many cases zero resistance (R), and in essentially all cases, zero shunt susceptance (B).

If a case has been created using a Ward equivalent or similar approach from a larger case, there may be equivalent lines modeled. In these cases, there should be no

expectation of the parameters corresponding to the geographic separation between the buses for equivalent lines. Unusual values like negative series impedances can be obtained through equivalencing. For the purposes of this analysis, the goal is to identify and ignore equivalent lines, as they do not provide any insight into the geographic accuracy of bus coordinates. Often equivalent lines are marked, for example by a circuit identifier of "EQ" or "99". Even if not, one typical feature is a large positive or negative series impedance with zero shunt susceptance (B).

Transmission lines

Transmission lines are the remaining branches, specified by per-unit reactance X, resistance R, and admittance B. Transmission lines will have a physical length which the parameters X, R, and B all correspond to. Prior work [52] has surveyed actual North American electric grids and provided a starting point for the expected per-unit, per-distance length for various categories of voltage levels. The crucial part of the corresponding table is repeated in Table 1. While the range is relatively wide (due in part to variations in line construction), this regularizing data can help to flag obviously invalid coordinates.

But the main way to know a line's length L is via the propagation time τ, if the values for X and B are relatively accurately known.

$$L = v_{\text{prop}} \cdot \tau = v_{\text{prop}} \sqrt{L_{\text{line}} C_{\text{line}}} = \frac{v_{\text{prop}}}{\omega} \sqrt{XB}$$

where we assume that the line's propagation speed v_{prop} is very near the speed of light $c = 3.0E8 \frac{\text{m}}{\text{s}}$. (Inductance and capacitance L_{line} and C_{line} marked with subscripts to distinguish from length L.) Note that there is no need to convert from per-unit X and B in this equation as the base values for impedance and admittance will cancel.

A crucial caveat in calculating transmission line length L and comparing it to the straight line distance between the two buses is that transmission lines do not in general follow a straight line path. The actual length L is always longer than $\ell = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. So at least this analysis provides a maximum value for the distance between two buses that are connected with a transmission line. Given that transmission lines are often run as straight as reasonably possible, given constraints

Table 1 Typical per-distance transmission line reactance (*X*) values [52]

Voltage class	10th percentile, per-unit, per-km	50th percentile, per-unit, per-km	90th percentile, per-unit, per-km
765 kV	0.000058	0.000070	0.000086
500 kV	0.000121	0.000155	0.000210
345 kV	0.000198	0.000360	0.000518
230 kV	0.000343	0.000948	0.001550
161 kV	0.000517	0.001828	0.003780
138 kV	0.000596	0.002471	0.006295
115 kV	0.000796	0.003398	0.006387
100 kV	0.00265	0.00480	0.00768
69 kV	0.00146	0.00804	0.01860

associated with geographic features and right-of-way access, buses too close together can be identified as well.

Quantifying bus coordinate quality

The following heuristic rules were put into place for branch validation related to ℓ, the distance between the substations, and L, the expected length of the line as determined by parameters. (See prior sections. For example, transformers have $L = 0$.) Any lines that do not fit into the following categories are given $q = 0$.

- If $L \leq 1\,\text{km}$

  o   If $l \leq 2\ km$, $q = 3$
  o   Else If $l \leq 5$ km, $q = 2$
  o   Else If $l \leq 1000$ km $q = 1$

- Els If $L \leq 40$ km

  o   If $0.4 \leq l/L \leq 1.05$, $q=3$
  o   Else If $0.2 \leq l/L \leq 1.$ $q=2$
  o   Else If $l/L \leq 5$ $q=1$

- Else if $L \leq 3000$ km

  o   *If* $0.7 \leq l/L \leq 1.05$, $q = 3$
  o   Else If $0.5 \leq l/L \leq 1.1$, $q = 2$
  o   Else If $l/L \leq 3$, $q=1$

These thresholds are heuristic and come from observations in the quality of real datasets. The 1 km threshold is the threshold below which quantifying the length and distinguishing from internal substation branches becomes more challenging, so these lines merely target the two buses being separated by a small distance ℓ. The threshold 40 km separates shorter length lines, where the line length could easily be double the straight-line distance, from longer ones that will tend to be straighter as a whole. The 3000 km threshold helps to eliminate unrealistically long lines, regardless of whether they match the line parameters. Notice that more room is given for the straight-line to actual length ratio to be less than 1.0 than greater than 1.0.

Next, each bus is assigned a coordinate quality flag based on the branches connected to it (except any buses which are set automatically to $q = 0$ via the criteria in III.A). First, buses are grouped into sets that are connected by branches where $L \leq 1\text{km}$ and $\ell \leq 2\text{km}$. Within a group (either a single substation or a cluster of very close nearby substations), the quality flags for all other connected branches are considered. The quality flag for the buses in that group is then set to the median value of the quality flags in the group, plus 1. The reason for using a median metric is that even one branch with $q = 3$ indicates that the spacing between a substation and at least one neighbor is within a good range. If all the $q = 3$ for every branch connected to the group, the group is set to $q = 5$.

Length-constrained graph layout

The goal of the layout algorithm is to assign geographic coordinates (x, y) to all the buses in a power flow case, given input case data and potentially some or all input geographic coordinate data, taking into account the quality flags described in the previous section (together with any a priori knowledge about which coordinates or other data are more reliable). The method must be generally applicable and computationally feasible even for very large systems because the target application is engineers needing to assign or clean up coordinates before making a visual diagram or running a GIC or infrastructure study.

We structure this problem as a graph drawing problem, where there is an assumed graph topology (branches connecting buses). Broadly, four assumptions underpin the framework of our formulation:

1. Feasibility

The system this graph represents is a physical system, so there exists some correct set of geographic coordinates (x_i, y_i) for bus i that satisfies all legitimate branch constraints. This solution may not be unique, and it might not be essential to reach it exactly.

2. Regularization

Some or all buses have a guess for their geographic coordinate $(\widehat{x}_i, \widehat{y}_i)$ with some confidence c_i (for bus i). Some buses have unknown locations where $c_i = 0$. If starting with no coordinates, pick one anchor bus and give it arbitrary coordinates.

3. Edge length constraints

Branches (edges) have an expected length L_{ij}, which the separation ℓ_{ij} between buses i and j should ideally approximate. Lines also have a scalar confidence level c_{ij}, which could be zero, for example, for equivalent lines. The length L_{ij} for transmission lines could be set to 80% of the line right-of-way path distance, to better represent the range of values ℓ could take.

4. Spread out

Subject to other constraints, the graph layout is spread out. Adjacent edges emanating from a bus have maximal angle separation. Buses that are far from each other by traversing the graph should also be far from each other spatially.

These assumptions are well-suited to formulation as a nonlinear programming problem, with soft constraints. First, define the objective portion for z_i for any node i in the set of nodes $i \in \mathcal{N}$.

$$z_i = c_i \left(\left(x_i - \widehat{x}_i \right)^2 + \left(y_i - \widehat{y}_i \right)^2 \right)$$

Next, for any branch (i, j) in the set of actual branches $(i, j) \in \mathcal{E}_1$, define the objective portion based on deviation in distance.

$$\ell_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$$z_{ij} = c_{ij}(L_{ij} - \ell_{ij})^2$$

The way we approach assumption 4 is in two parts: local spread out and global spread out. Each results in creating new sets of edges. For local spread out we define new branches $(i,j) \in \mathcal{E}_2$ that are second neighbors in the graph with edges $\mathcal{E}_1$. That is, two buses i and j are connected in $\mathcal{E}_2$ if there is some bus k such that $(i,k) \in \mathcal{E}_1$ and $(j,k) \in \mathcal{E}_1$ but $(i,j) \notin \mathcal{E}_1$. For the local spread parameter a_{ij} just use the length, with a single local spread scaling factor α for the whole system.

$$a_{ij} = \alpha \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Similarly, the global spread-out assumption is handled with a third set of edges $\mathcal{E}_3$. These edges are formed by recursive binary partition of the graph associated with $\mathcal{E}_1$.

The binary partition works as follows. In each recursive iteration consider the node-set $\mathcal{N}_k$. If there are less than 5 nodes in the set, return. Otherwise, select the two extreme points of the graph, i and j, which are the two nodes which are separated by the longest path length l_{ij} along the length of the graph (found or approximated using Dijkstra's algorithm). Add (i,j) to $\mathcal{E}_3$, then partition the buses in $\mathcal{N}_k$ into two subsets, $\mathcal{N}_{ki}$ for buses closer to i along the length of the graph (again using Dijkstra) and $\mathcal{N}_{kj}$ for buses closer to j. Recursively repeat for each subset until completed.

With the partition done, create objective components b_{ij} for (i,j) proportional to both the length l_{ij} and a system-wide global spreading factor β.

$$b_{ij} = \beta \cdot l_{ij} \cdot \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

The reason these global scaling factors are proportional to length is that the partitions higher on the binary tree should spread out further (for example, the first pair will be the two furthest nodes on the graph).

With these pieces in place, the final optimization problem can be formulated as:

$$\min_{x,y} \left(\sum_{i \in \mathcal{N}} z_i + \sum_{(i,j) \in \mathcal{E}_1} z_{ij} - \sum_{(i,j) \in \mathcal{E}_2} a_{ij} - \sum_{(i,j) \in \mathcal{E}_3} b_{ij} \right)$$

with no hard constraints in a "subject to" clause. Hence we reduce the initial constrained coordinate assignment problem to an equivalent, unconstrained problem, optimizing over the control variables x_i and y_i (horizontal and vertical positions of all nodes $(i \in \mathcal{N})$ with the objective function minimizing two functions, separation from reference coordinates (z_i) and expected edge length (z_{ij}), defined above, weighted by parameters c. Simultaneously, the objective function seeks to maximize both local and global spreading with the a and b functions, defined above as well. This unconstrained, continuously differentiable problem is excellently suited to a standard nonlinear optimizer such as IPOPT, as implemented below.

A few observations can be made about this formulation. First, it is tunable depending on the system and confidence levels in the different data. The parameters c and β are unitless, whereas α would have length units (like meters). Higher values of α and β will cause the coordinates to spread out more, at the expense of deviating more from the known branch lengths. The second observation is that both the regularization and edge length terms are quadratic and tend to pull the system together, whereas both of the spread terms are linear and tend to push the system apart. Very broadly speaking, the user picks α and β to establish a constant "force" that sets the tolerable deviation from a priori coordinates and branch lengths, analogously to force-directed graph layout methods. The third observation is that the system is quite sparse. None of the three edge sets will have a size much greater than the original number of branches. Unlike a typical force-directed graph layout method, there is no need to calculate the distances between every pair of points. This has the effect of keeping the computational complexity low.

Results and discussion

In this section, we demonstrate the ability of the geographic quality assessment method (shortened to GQA in this section) to identify errors in bus coordinates, and for the edge length-constrained graph layout (shortened to LCL for this section) to determine new coordinates that are of high quality in terms of consistency with the power flow data.

The implementation of these methods is on a laptop with an 11th Gen Intel i7 processor at 2.5 GHz clock speed and 64 GB RAM. The non-linear optimization problem was formulated with the Pyomo platform and solved with the Interior Point Optimizer (IPOPT) [53, 54].

A variety of test scenarios are used for the results in this section, with three main base grids:

1. *Hawaii40*. This 37-bus case is synthetic, built with an algorithm according to the methods described in [46] and [55]. It does not correspond to any actual grid or contain CEII, so its data is made available at [47]. It is geographically located on the Hawaiian island of Oahu. Since it is synthetic, it has ground truth coordinates that are consistent with the line parameters.
2. *Texas7k*. This 6716-bus case is also synthetic [46, 55], geo-located on the portion of the U.S. state of Texas served by the Electric Reliability Council of Texas (ERCOT). Like the Hawaii40 case, its data is available at [47], it does not contain CEII, and it has ground truth latitude and longitude. As this case is a realistic size, variations on it are used for the majority of the results in this paper.
3. *Grid3*. This is an actual model of a portion of the electric grid located in North America, with about 5000 buses. It is used in the last section of results to verify the methodology against real data. Only high-level results can be given because the case contains CEII. It has a priori assumed coordinates, but they are not ground truth coordinates and there are some known issues with the data, which this algorithm is shown to address (see "Results for Grid3" section).

The scenarios for this analysis have been selected to mimic potential applications to real situations and to demonstrate the effectiveness of the GQA and LCL methods.
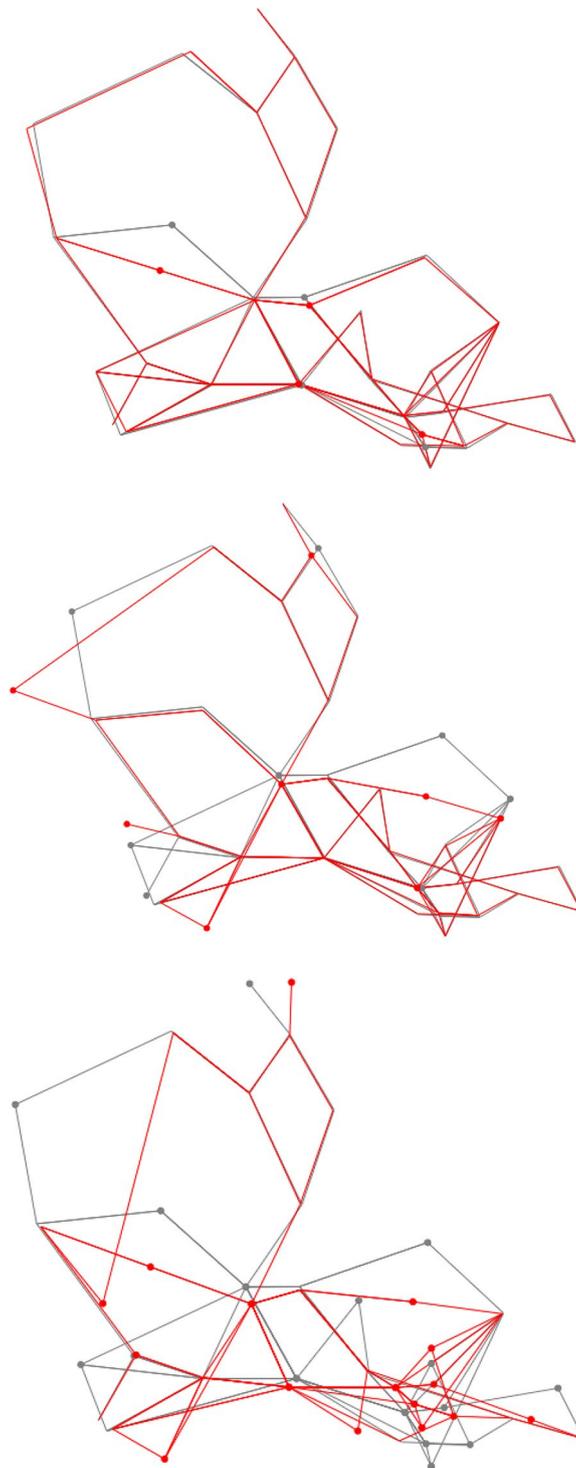
Missing coordinates in Hawaii40

The first examples are shown in the Hawaii40 case because it is small enough that individual nodes can be distinguished in the figures. The base case has ground truth coordinates which are known because of the way the synthetic grids are designed. In the design of the transmission lines (as described in [46]), the length is assumed to be 1.0 to 1.5 times the straight-line path between the substations, with parameters X and B set correspondingly depending on the tower design. So it is no surprise that the GQA process scores essentially all of the lines with a perfect quality flag of 3, and essentially all of the buses with a perfect quality flag of 5. There are a few exceptions for three lines added later without the correct process, which GQA flagged with $q = 1$.

The scenarios tested involved assuming that there was missing substation coordinate data for a subset of the substations in the case. The GQA was run on the case with these missing coordinates, and then the LCL method was run to attempt to provide estimated coordinates for these substations, based only on the known line data and the relation to the remaining, correct substation coordinates. For the LCL algorithm in these cases, the branch confidence constants c_{ij} were set to 1 for all branches, and the bus confidence constant c_i was set to 10 for buses with $q \geq 1$. The spread constraints were included with $\alpha = \beta = 0.001$. After the LCL was complete, for each case the GQA was run again to check the improvement in coordinate quality.

Table 2 and Fig. 1 show the results for the base case and three scenarios: one with coordinates assumed to be missing for 4 substations, one with 8 missing, and one with 16 missing. In Table 2, each scenario is shown with two rows, before and after the LCL algorithm assigns new coordinates to the buses. In all cases, the algorithm manages to find coordinates for the buses such that all the lines (except the three with known data challenges) have lengths that reflect their parameters, and as a result, nearly all the buses have perfect quality flags $q = 5$. Figure 1 shows where these coordinates are set in each case. Of course, there are many possible, valid solutions as the hints from power flow variables do not uniquely specify the coordinates. In most cases, the estimated coordinate is quite near the ground truth coordinate, separated sometimes by just a degree of freedom such as flipping over an axis. For the purposes of visualization or GIC calculations, these estimated coordinates would be better than having no coordinates or very wrong coordinates.

Table 2 Hawaii40 results of GQA

Scenario	Number of branches with quality flag $q =$				Number of buses with quality flag $q =$					
	0	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**4**	**5**
Base	0	3	0	86	0	0	0	0	3	34
after LCL	0	3	0	86	0	0	0	0	3	34
Missing 4	10	4	0	75	6	0	0	0	11	20
after LCL	0	3	0	86	0	0	0	0	3	34
Missing 8	33	8	0	48	11	2	3	0	15	6
after LCL	0	3	0	86	0	0	0	0	3	34
Missing 16	49	16	0	24	21	9	0	0	6	1
after LCL	0	1	2	84	0	0	0	0	5	33

Fig. 1 Results of coordinate estimation with LCL in Hawaii40. Gray is ground truth, red is estimation. Bolder dots highlight the substations whose coordinates were assumed to be missing. Top: 4 substations with missing coordinates. Middle: 8 substations with missing coordinates. Bottom: 16 substations with missing coordinates

Fixing coordinate mapping errors in Texas7k

For the next set of scenarios, Texas7k is used, which has a size more commensurate with actual electric grid models. From the base case which is ground truth, varying levels of different types of errors in the coordinate mapping are considered.

Four different types of errors in coordinates are considered. First, for some substations, the coordinates are assumed to be unknown. Second, for some buses, the mapping is assumed to be wrong, so that the bus is assigned to a different substation's coordinates, potentially on the other side of the case. Third, for some substations, the coordinates are assumed to be slightly wrong, by adding a random error on the order of 1° latitude and longitude. Fourth, for some substations, the coordinates are assumed to be rounded to the nearest degree. Note that in all cases the algorithm does not know a priori which coordinates are correct or incorrect, but estimates this using GQA.

Table 3 and Fig. 2 show the results for scenarios in this case. Six cases were run, with varying levels of errors from 5% up to 30%. In each one, the selection of the buses to have errors and the type of error were assumed to be random. In the LCL algorithm, the parameters were the same as for the Hawaii40 cases in the prior subsection, except that c_i for buses with $q=1$, $q=2$, and $q=3$ were changed to 0.01, 0.1, and 1 respectively to allow more freedom for the algorithm to improve these coordinates. These parameters are always a trade-off in how trusted the coordinates are and how strongly the power flow data indicates the coordinates should be changed.

Figure 2 illustrates what the LCL algorithm is doing. The ground truth coordinates (gray) are modified to simulate errors. The errors cannot be shown in Fig. 2 without majorly cluttering the image, since lines appear to be crisscrossing the whole case, plus some substations are assumed to have no coordinates. However, the fixed coordinates (red) are shown and tend to estimate the original coordinates very well.

As shown in Table 3, adding the errors at various levels can be detected by GQA, with the number of buses with $q \leq 1$ approximately equal to the determined percentage of errors. Then with the coordinate estimation through LCL, the low-quality bus coordinates are greatly improved. Even with nearly 1/3 of buses incorrectly mapped, the LCL can find a solution with 94% of the buses having $q = 4$ or $q = 5$, and none with $q \leq 1$.

Table 3 Texas7K results of GQA with error correction

Scenario	Number of buses with quality flag $q=$					
	0	1	2	3	4	5
Base	0	0	5	5	199	6508
Errors 5%	72	162	75	18	1064	5326
after LCL	0	0	17	18	548	6134
Errors 10%	154	424	177	13	1612	4337
after LCL	0	0	54	66	717	5880
Errors 15%	230	630	342	25	2067	3423
after LCL	0	0	99	90	1046	5482
Errors 20%	361	880	437	31	2236	2772
after LCL	0	0	116	160	1316	5125
Errors 25%	465	1147	484	55	2431	2135
after LCL	0	0	129	135	1587	4866
Errors 30%	504	1440	617	41	2382	1733
after LCL	0	0	205	200	1696	4616
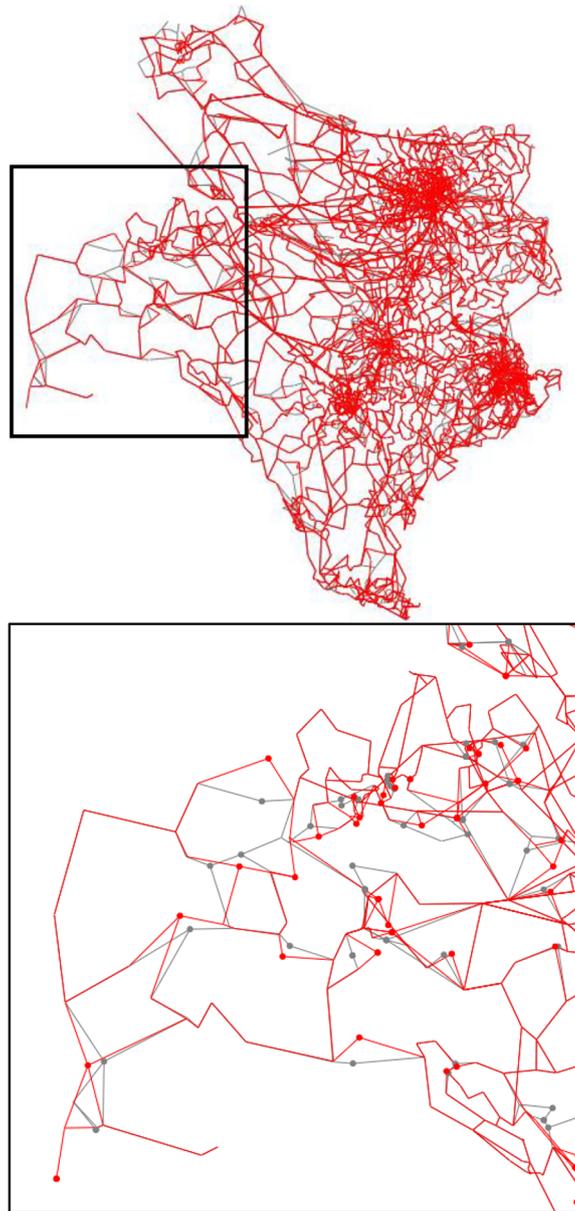
Fig. 2 Results of coordinate estimation with LCL in Texas7k, under the 15% error condition. Gray is ground truth, red is estimation. Bolder dots highlight the substations whose coordinates were assumed to have major errors. Top: full case view. Bottom: Zoomed-in view of the far western area

Building coordinate sets from sparse starting points   Next, six scenarios are considered to emulate the conditions in which very little is known about a case's geographic context. Texas 7 k is used for these as well. First, we look at the condition in which a single area is missing from the case. We look at the South Central Area (Austin and San Antonio region) being missing (NoSouthCent), and the South Area (Corpus Christi, Laredo, and Lower Rio Grande Valley region) (NoSouth). This is done so that both a central area being missing and an edge area being missing can both be tested. Second, we look at the condition in which only one area is known, in both of these cases as well (OnlySouthCent, OnlySouth).

We then look at the conditions in which the extra high voltage (EHV) network (345 kV in this case) is known but the rest of the case needs to be inferred (OnlyEHV). Finally, we consider the case where none of the coordinates are known at all (AllUnknown).

For these cases, in solving them with the LCL, the spread-out parameters are more important and are set to $\alpha = 1$ for all cases, and $\beta = 0.01$ for the area missing cases and $\beta = 0.3$ for the other cases. These were the slowest cases to run computationally, but still none of them took longer than 2 min.

The results are shown in Table 4 and in Fig. 3. The two cases with only one area missing are not too unlike the prior subsection cases with 10–15% of substation errors, except that the substations with unknown coordinates are all together in one region. Therefore there is more deviation as a whole from the ground truth coordinates, as the top panel of Fig. 3 shows. Nevertheless, the overall shape of the missing area tends to match the actual coordinates fairly well, and the GQA results (Table 4) show a very high level of correlation between the estimated coordinates and the line parameters.

The results for the cases with only one area known certainly have more deviation from the ground truth coordinates, since these cases involve over 80% of the coordinates for the case unknown. But, thanks to the spreading mechanism, the different unknown areas still tend to separate from one another and form reasonable structures, as shown in the bottom panel of Fig. 3. Similarly, the cases with only the 345 kV network known or with none of the case known involve the LCL algorithm having to estimate coordinates from scratch, but it does find coordinates that result in relatively high quality flags for most buses. The main application of these scenarios would be for quick data visualization on a case with no readily available coordinates.

Results for Grid3

This subsection presents some results from Grid3, an actual electric grid subsystem case located in North America with about 5000 buses. The base case coordinate set is relatively high quality but is not fully ground truth as there are some errors, missing coordinates, and some low-resolution coordinates. The GQA results are shown in Table 5 (note that exact values are not given but just a percentage), with the $q=0$ and $q=1$ buses mainly being the ones with major errors or missing data, and the larger group of $q=2$

Table 4 Texas7K results of GQA with sparse starts

Scenario	Number of buses with quality flag $q=$					
	0	1	2	3	4	5
NoSouthCent	1138	9	5	5	269	5291
after LCL	0	0	2	10	221	6484
NoSouth	630	5	5	5	236	5836
after LCL	0	0	3	9	245	6460
OnlySouthCent	5579	9	0	0	115	1014
after LCL	0	1	717	683	2225	3091
OnlySouth	6087	3	0	0	67	560
after LCL	0	2	942	738	2357	2678
OnlyEHV	5675	361	35	29	394	223
after LCL	0	2	892	818	2530	2475
AllUnknown	6717	0	0	0	0	0
after LCL	0	0	854	911	2516	2436
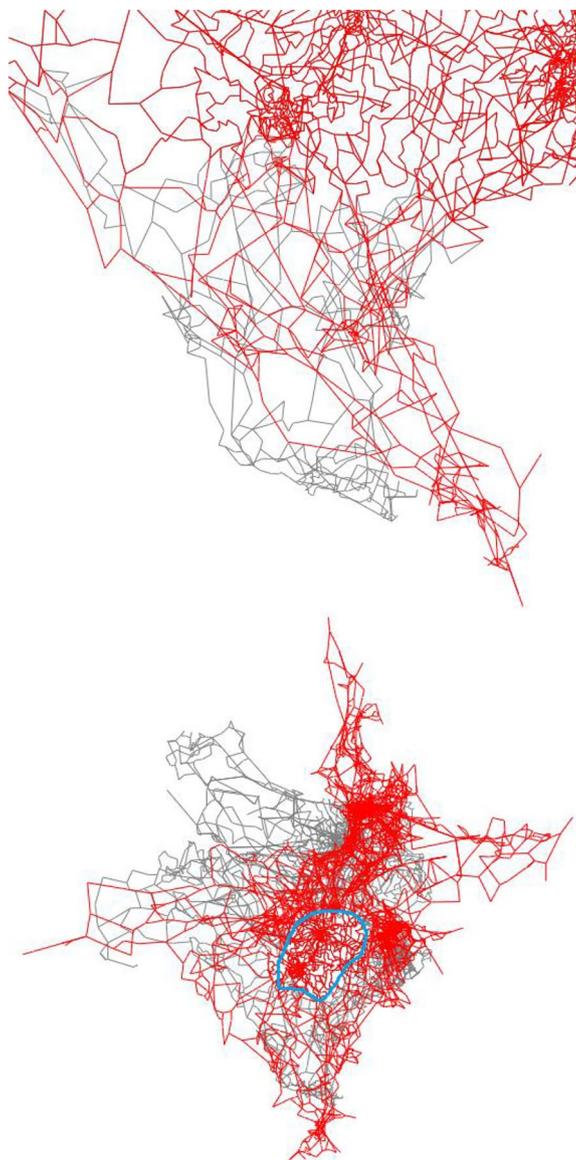
Fig. 3 Results of coordinate estimation with LCL in Texas7k, under the sparse start conditions. Gray is ground truth, red is estimation. Top: NoSouth case, zoomed in for detailed view. Bottom: OnlySouthCent case, with the known area circled in blue

Table 5 Grid3 results of GQA

Scenario	Percent of buses with quality flag $q =$					
	0	**1**	**2**	**3**	**4**	**5**
Base	0.1	0.6	32.3	7.7	25.5	33.9
with LCL	0	0	2.6	4.2	29.3	64.4
Add 10% errors	3.1	7.0	33.4	6.3	28.4	21.8
with LCL	0	0.3	5.9	5.2	33.6	55.0
All unknown	100	0	0	0	0	0
with LCL	0	0.1	11.5	10.1	37.0	41.3

buses (about 1/3 of all the buses) being the regions for which low-resolution rounded data were all that was available.

With LCL applied, the coordinates are greatly improved in their correspondence with the power flow data, with no buses in the $q \leq 1$ region and only very few in the $q \leq 3$ region.

For reference, two other scenarios were run on the Grid3 case. The first was introducing additional intentional errors, much as in the "Fixing coordinate mapping errors in Texas7k" section, at the 10% level. These can be seen in the additional 10% of buses with $q = 0$ or $q = 1$. The second scenario is with all coordinates unknown, starting from scratch, as in the "Building coordinate sets from sparse starting points" section. By comparing Table 5 to Tables 3 and 4, it is clear that the performance of the GLA and LCL on a real case is comparable to the results from the synthetic cases in Hawaii40 and Texas7k.

Conclusions

This paper addresses the problem that exists when a power system analysis task needs reasonable geographic coordinates for a grid and either (1) no such coordinates exist, or (2) some coordinates exist but others are missing, or (3) a set of coordinates exist but some are severely incorrect, or (4) a set of coordinates exist and are generally correct but the exact accuracy is not known. Given the importance of data visualization for large-scale power systems, the growing research efforts in preparing for potential GMD events, and the usefulness of applying other geo-mapped data in coordination with the electric grid, an automated process to create reasonable coordinates or fill in missing gaps can help to support further work in a number of areas. Of course, having actual known coordinates is always better if that is possible. But often such coordinates are not easy to obtain, or not easy to map to the buses in a given snapshot case, without extensive human labor. The work in this paper takes advantage of the fact that some geographic information is contained in the power flow data itself, particularly the branch impedance and susceptance parameters as indicators of branch length. By applying these constraints to a modified graph drawing algorithm, an optimization-based general approach can be made to estimate missing or incorrect coordinates. This algorithm is expected to perform well with respect to many practical cases since all actual cases are geographically embedded in reality. It is possible that in some unusual configurations or highly complex cases, there might be some challenges to the algorithm, particularly in tight downtown urban networks or fictitious test cases without a true geographic embedding. For practical implementation, the method is quite tolerant of data errors and does not require significant computational resources. The algorithm in this paper is scalable and fast and will result in reasonable coordinates that will respect the expected line lengths as well as keep any known accurate coordinates intact.

Methods

The purpose of the test cases described in the "Results and discussion" section was to evaluate the efficacy of the proposed geographic coordinate validation and length-constrained layout methods on large-scale, realistic power system network models. The study was designed with three test cases: Hawaii40, which has 40 buses, Texas7k, a

system with 6717 buses, and Grid3, with approximately 5000 buses. The study was performed in the setting of the Texas A&M University's campus. The only materials used for the study were the desktop computers used to run the analysis (laptop with an 11th Gen Intel i7 processor at 2.5 GHz clock speed and 64 GB RAM) along with the input data which consisted of synthetic electric grid models (Hawaii40 and Texas7k) and actual electric grid models (Grid3). The synthetic grid models used for this study are publicly available online (see [47]). The actual grid model is not available due to CEII restrictions. The non-linear optimization problem given in the "Length-constrained graph layout" section was formulated with the Pyomo platform and solved with the Interior Point Optimizer (IPOPT). The results interpretation did not involve a statistical analysis. There were no human or animal subjects involved in this research.

Abbreviations

CEII	Critical Energy Infrastructure Information
ERCOT	Electric Reliability Council of Texas
EMP	Electromagnetic pulse
EIA	Energy Information Administration
EHV	Extra high voltage
FERC	Federal Energy Regulatory Commission
GMD	Geomagnetic disturbance
GIC	Geomagnetically induced current
GQA	Geographic quality assessment
GIS	Geographic information systems
IPOPT	Interior Point Optimizer
LCL	Length-constrained graph layout
NERC	North American Electric Reliability Corporation
UTM	Universal Transverse Mercator

Declarations

Competing interests
The author declares no competing interests.

References

1. Birchfield AB, Overbye TJ (2018) Techniques for drawing geographic one-line diagrams: Substation spacing and line routing. IEEE Trans Power Syst 33(6):7269–7276
2. Belmudes F, Ernst D, Wehenkel L (2009) Pseudo-geographical representations of power system buses by multidimensional scaling. 2009 15th International Conference on Intelligent System Applications to Power Systems, Brazil, pp 1–6
3. Overbye TJ, Rantanen EM, Judd S, Electric power control center visualization using Geographic Data Views, (2007) iREP Symposium - Bulk Power System Dynamics and Control - VII. Revitalizing Operational Reliability, Charleston, SC, USA 2007:1–8
4. Cuffe P, Saiz Marin E, Keane A (2017) For power systems, geography doesn't matter, but electrical structure does. IEEE Potentials 36(2):42–46
5. TPL-007–4 (2020) Transmission system planned performance for geomagnetic disturbance events. NERC: Atlanta, GA

6. Overbye TJ, Hutchins TR, Shetye KS, Weber J, Dahman S (2012) Integration of geomagnetic disturbance modeling into the power flow: a methodology for large-scale system studies. 2012 North American Power Symposium, Champaign, IL

7. V.D. Albertson, J.G. Kappenman, N. Mohan, and G.A. Skarbakka (1981) Load-flow studies in the presence of geomagnetically-induced currents. IEEE Trans. Power App. Syst. 594–606.

8. Ahmed A, McFadden FJS, Rayudu R (2019) Weather-dependent power flow algorithm for accurate power system analysis under variable weather conditions. IEEE Trans Power Syst 34(4):2719–2729

9. Vallee F, Lobry J, Deblecker O (2007) Impact of the wind geographical correlation level for reliability studies. IEEE Trans Power Syst 22(4):2232–2239

10. Dong Z, Tian M (2021) Modeling and vulnerability analysis of spatially embedded heterogeneous cyber-physical systems with functional dependency. IEEE Transactions on Network Science and Engineering. 8(4):3404–3416

11. Zlotnik A, Roald L, Backhaus S, Chertkov M, Andersson G (2017) Coordinated scheduling for interdependent electric power and natural gas infrastructures. IEEE Trans Power Syst 32(1):600–610

12. Wert JL, Coupled infrastructure simulation of electric grid and transportation networks, et al (2021) IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). Washington, DC, USA 2021:1–5

13. Chini CM, Djehian LA, Lubega WN, Stillwell AS (2018) Virtual water transfers of the US electric grid. Nat Energy 3:1115–1123

14. Southwest Power Pool (SPP) Model Development Procedure Manual, Version 6.0, SPP.

15. MISO Reliability Planning Model Data Requirements and Reporting Procedures, Version 4.1, MISO, August 2022.

16. Tomlinson R (2007) Thinking About GIS, 3rd edn. ESRI Press, Redlands, California

17. Clarke KC (1986) Advances in geographic information systems. Comp Environ Urban Systems 10(3/4):175–184

18. Quiros-Tortos J, Valverde G, Arguello A, Ochoa LN (2017) Ochoa, Geo-information is power: Using geographical information systems to assess rooftop photovoltaics in Costa Rica. IEEE Power Energy Mag 15(2):48–56

19. Wang Y, Chen C, Wang J, Baldick R (2016) Research on resilience of power systems under natural disasters—a review. IEEE Trans Power Syst 31(2):1604–1613

20. L. Richaud, R. Pellerej, C. Benoit, and E. Ramos, Analysis of voltage patterns for topology identification and GIS correction, in *25th International Conference on Electricity Distribution*, Madrid, Spain, June 3–6, 2019.

21. Guzman A, Arguello A, Quiros-Tortos J, Valverde G (2019) Processing and correction of secondary system models in geographic information systems. IEEE Trans Industr Inf 15(6):3482–3491

22. K. Montano-Martinez, S. Ma, V. Vittal, and C. Rojas, Automated correction of GIS data for loads and distributed energy resources in secondary distribution networks, to appear in *IEEE Transactions on Power Systems*, 2023.

23. Shin J-H, Yi B-J, Kim Y-I, Yang I-K (2010) Development of power distribution facility map input system using automatic image recognition technology. IEEE Trans Power Delivery 25(1):231–238

24. Tollis IG, Di Battista G, Eades P, Tamassia R (1998) Graph Drawing: Algorithms for the Visualization of Graphs, 1st edn. Prentice-Hall, Englewood Cliffs, NJ, USA

25. H. C. Purchase, R. F. Cohen, and M. James, Validating graph drawing aesthetics, in *Proc. Symp. Graph Drawing*, Passau, Germany, 1995, pp. 435–446.

26. Huang W, Huang ML, Lin C (2016) Evaluating overall quality of graph visualizations based on aesthetics aggregation. Inf Sci 3300(10):444–454

27. Fruchterman TM, Reingold EM (1991) Graph drawing by force-directed placement. Softw Pract Exp 21(11):1129–1164

28. P. Eades, Q. Feng, and X. Lin, Straight-line drawing algorithms for hierarchical graphs and clustered graphs, in *Proc. Symp. Graph Drawing*, 1996, pp. 113–128.

29. Parker BJ, Chao RF, Sabiston JKM, Locke P (1991) An analytical technique to evaluate station one-line diagrams in a network context. IEEE Trans Power Del 6(4):1454–1461

30. Ong YS, Gooi HB, Chan CK (2000) Algorithms for automatic generation of one-line diagrams. IEE Proc Gener Transm Distrib 147(5):292–298

31. Moreira JC, Miguez E, Vilacha C, Otero A (2012) Large-scale network layout optimization for radial distribution networks by parallel computing: implementation and numerical results. IEEE Trans Power Del 27(3):1468–1467

32. Ravikumar G, Pradeep Y, Khaparde SA (2013) Graphics model for power systems using layouts and relative coordinates in CIM framework. IEEE Trans Power Syst 28(4):3906–3915

33. S. C. Teja and P. K. Yemula, Power network layout generation using force-directed graph technique, in *Proc. 18th Nat. Power Syst. Conf.*, Guwahati, India, 2014, pp. 1–6.

34. de AssisMota A, Mota LTM (2011) Drawing meshed one-line diagrams of electric power systems using a modified controlled spring embedder algorithm enhanced with geospatial data. J Comput Sci 7(2):234–241

35. P. Cuffe and A.Keane, Novel quality metrics for power system diagrams, in *Proc. IEEE Int. Energy Conf.*, Leuven, Belgium, 2016, pp. 1–5.

36. Cuffe P, Keane A (2017) Visualizing the electrical structure of power systems. IEEE Syst J 11(3):1810–1821

37. Overbye TJ, Wert J, Birchfield A, Weber JD, Wide-area electric grid visualization using pseudo-geographic mosaic displays, (2019) North American Power Symposium (NAPS). Wichita, KS, USA 2019:1–6

38. Birchfield AB, Overbye TJ (2020) Mosaic packing to visualize large-scale electric grid data. IEEE Open Access Journal of Power and Energy 7:212–221

39. Wang J, Chen J, Shi D, Duan X (2024) Automatic generation of topology diagrams for strongly-meshed power transmission systems. IEEE Trans Power Syst 39(1):1918–1931

40. Olauson J, Marin M, Söder L (2020) Creating power system network layouts: a fast parallel algorithm. IEEE Syst J 14(3):3687–3694

41. University of Washington. Power system test case archive. [Online]. Available: https://labs.ece.uw.edu/pstca/.

42. Barrows C et al (2020) The IEEE reliability test system: a proposed 2019 update. IEEE Trans Power Syst 35(1):119–127

43. U.S. Federal Energy Regulatory Commission (FERC) Order 683, April 2007; [Online] https://www.ferc.gov/sites/default/files/2020-04/OrderNo683-A.pdf.

44. U.S. Energy Information Administration (EIA) Form 860. https://www.eia.gov/electricity/data/eia860/.

45. Horton R, Boteler D, Overbye TJ, Pirjola R, Dugan RC (2012) A test case for the calculation of geomagnetically induced currents. IEEE Trans Power Delivery 27(4):2368–2373
46. Birchfield AB, Xu T, Gegner KM, Shetye KS, Overbye TJ (2017) Grid structural characteristics as validation criteria for synthetic networks. IEEE Trans Power Syst 32(4):3258–3265
47. Texas A&M Electric Grid Test Case Repository. [Online]. Available: electricgrids.engr.tamu.edu.
48. S. Taylor, A. Rangarajan, N. Rhodes, J. Snodgrass, B. Lesieutre, and L. A. Roald, California test system (CATS): A geographically accurate test system based on the California grid, pre-print available: https://arxiv.org/abs/2210.04351.
49. Overbye TJ, Shetye KS, Hutchins TR, Qiu Q, Weber JD (2013) Power grid sensitivity analysis of geomagnetically induced currents. IEEE Trans Power Syst 28(4):4821–4828
50. Defense Mapping Agency, *The Universal Grids: Universal Transverse Mercator (UTM) and Universal Polar Stereographic (UPS)*, AD-A266497, Fairfax, VA, Sept. 1989. https://apps.dtic.mil/sti/pdfs/ADA266497.pdf.
51. K. Kawase, Concise derivation of extensive coordinate conversion formulae in the Gauss-Kruger projection in *Bulletin of the Geospatial Information Authority of Japan*, vol. 60, Mar. 2013.
52. Birchfield AB, Schweitzer E, Athari MH, Xu T, Overbye TJ, Scaglione A, Wang Z (2017) A metric-based validation process to assess the realism of synthetic power grids. Energies 10(1233):1–14
53. Hart WE, Watson J-P, Woodruff DL (2011) Pyomo: modeling and solving mathematical programs in Python. Math Program Comput 3(3):219–260
54. Wachter A, Biegler LT (2006) Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math Prog Ser A 106:25–57
55. A. B. Birchfield and T. J. Overbye, Planning sensitivities for building contingency robustness and graph properties into large synthetic grids, *Hawaii International Conference on System Sciences (HICSS)*, Jan. 2020, pp. 1–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.