Multi-attribute Partitioning of Power Networks Based on Electrical Distance

Eduardo Cotilla-Sanchez, *Member, IEEE*, Paul D. H. Hines, *Member, IEEE*, Clayton Barrows, *Member, IEEE*, Seth Blumsack, *Member, IEEE*, Mahendra Patel, *Senior Member, IEEE*

***Abstract*—Identifying coherent sub-graphs in networks is important in many applications. In power systems, large systems are divided into areas and zones to aid in planning and control applications. But not every partitioning is equally good for all applications; different applications have different goals, or attributes, against which solutions should be evaluated. This paper presents a hybrid method that combines a conventional graph partitioning algorithm with an evolutionary algorithm to partition a power network to optimize a multi-attribute objective function based on electrical distances, cluster sizes, the number of clusters, and cluster connectedness. Results for the IEEE RTS-96 show that clusters produced by this method can be used to identify buses with dynamically coherent voltage angles, without the need for dynamic simulation. Application of the method to the IEEE 118 bus and a 2383 bus case indicates that when a network is well partitioned into zones, intra-zone transactions have less impact on power flows outside of the zone; i.e., good partitioning reduces loop flows. This property is particularly useful for power system applications where ensuring deliverability is important, such as transmission planning or determination of synchronous reserve zones.**

***Index Terms*—Network clustering, power network partitioning, evolutionary algorithms, electrical distance.**

I. Introduction

THE electric power infrastructure of the United States and Canada is divided into four synchronous interconnections. Each of these is subsequently partitioned into Regional Transmission Organizations (RTO) or Balancing Authorities (BA). Balancing areas are frequently sub-divided into zones for particular planning or control applications. These divisions are used to reduce the computational and administrative complexity associated with many planning and operations applications. In most regions operational security analysis, resource adequacy assessments, zonal pricing, zone-based voltage control schemes, Area Control Error (ACE) calculations, reserves scheduling, and capacity obligation determinations all use areas or zones in one form or another. Some of these applications are reviewed in [1]–[4]. In some regions (PJM, for example), the existing zonal boundaries result from historical asset ownership, rather than the physical properties of the network. This simplifies the network partitioning problem, but raises questions regarding the quality of the planning, operational and reliability applications that make use of zones.

The partitioning of data and networks has a long history in the scientific literature. Kron and Happ [5]–[7] pioneered the study of diakoptics, or 'tearing' to reduce the computational requirements associated with analyzing large-scale systems. The clustering of data enables a variety of applications in statistics and artificial neural networks [8], [9]. Partitioning methods can also be used to reveal complex community structures in large-scale networks [10], [11].

Several approaches have been proposed for dividing power networks into clusters. Clustering methods have been used to define control areas for reactive power markets or zones for voltage security assessment [12]–[14]. Kamwa et al. [15], [16] employed clustering methods to evaluate the dynamic vulnerability of a real system (Hydro-Quèbec). Several studies used simulation methods to divide a power network, such as the slow coherency approach [17] or disturbance simulation approaches [15], [18], [19]. Li et al. [20] developed a hierarchical clustering method that takes into account the active and reactive power mismatch between areas. Recent studies have also shown that network partitioning can facilitate the integration of renewable sources [21], [22].

This paper describes an alternative approach, which defines zones within a power system as collections of buses such that buses within a zone are strongly connected and buses between zones are weakly connected. Strongly-connected buses could be treated as aggregated single buses for purposes of power system analysis. We formalize this notion of strong and weak connections using an electrical distance measure that relates network topology to active-power sensitivities, and use this measure in a multi-attribute network partitioning problem that seeks to minimize distances between nodes within a zone, and maximize distances between nodes in different zones. Our electrical distance measure is primarily based on information found in the system admittance (Y_{BUS}) matrix, thus avoiding the need for extensive use of simulation data. We derive a proper electrical distance measure from information contained in the Y_{BUS} matrix. Our definition of electrical distance and our multi-attribute partitioning approach are likely to be useful for a number of network analysis and security applications where identifying closely-tied buses is advantageous, including identification of locational load-shedding to maintain secu-

This work was funded in part by PJM Applied Solutions, NSF ECCS award #0848247, and US DOE award #DE-OE0000447.

E. Cotilla-Sanchez is with Oregon State U., School of Electrical Eng. & Computer Science, Corvallis, OR 97331, USA. E-mail: cotillaj@eecs.oregonstate.edu.

P. Hines is with the U. of Vermont, School of Eng., Burlington, VT 05405, USA. E-mail: paul.hines@uvm.edu.

C. Barrows is with the National Renewable Energy Laboratory, Golden, CO 80401, USA. E-mail: clayton.barrows@nrel.gov.

S. Blumsack is with the Pennsylvania State U., Leone Family Dept. of Energy and Mineral Eng., University Park, PA 16802, USA. E-mail: sethb@psu.edu.

M. Patel is with PJM Applied Solutions, Norristown, PA 19403, USA. E-mail: patelm3@pjm.com.

rity; wide-area monitoring problems such as synchrophasor placement; and the definition of reserve requirements (such as installed capacity or reactive reserve). The properties of our multi-attribute network partitions would also be useful for applications where ensuring deliverability is important, including transmission planning and reserves scheduling. One feature of our approach to clustering power networks is that the zonal definitions are dependent on the network topology and not on specific operating points. This has potential advantages (in making the clustering method useful in a variety of applications) but also has its limits - since the electrical distance measure is a function of network topology, our approach is probably not well-suited for applications where system dynamics are important.

This work improves substantially upon preliminary work by the authors [4] by using an improved measure of electrical distance (based on work in [23]), an improved representation of solutions for the evolutionary computational algorithm, a revised fitness function that considers both within- and between-cluster distances, and results showing that good partitioning results reduce extra-zonal power flows.

The remainder of this paper is organized as follows. Section II illustrates how to apply the concept of electrical distance to a power system clustering problem. Section III describes the individual metrics that compose the multi-attribute optimization function. Section IV details the evolutionary computational algorithm (EA) that we use in this work and Sec. V summarizes the results obtained for two example power networks. Finally, Sec. VI provides conclusions from this work.

II. Measuring electrical distance

The general graph partitioning problem can be stated as follows. Given a graph with n vertices V ($|V| = n$), a set of edges linking V and some measure of distance between all pairs of vertices, $d(a,b)$, $\forall (a,b) \in V$, find a way to divide V into approximately p sets $\{M_1, M_2, ..., M_p\}$ ($M_i \subset V$, $M_i \cap M_j = \varnothing$), such that the distances between the sets are maximized and the distances within the sets are minimized. Different methods measure distances differently, or put more relative emphasis on these two objectives, but all algorithms use distances in some form or another. In some methods (e.g., spectral clustering [24]), distances are inferred from the network topology. In others (e.g., K-means [32]) distances are used explicitly.

In this paper we explicitly evaluate the quality of solutions using electrical distances. While electrical distances have been used in a number of power systems problems [12], [13], [25], [26], only in [25] (and our own previous work [4]) were they explicitly used for network partitioning. Lagonotte et al. [25] showed that the logarithmic voltage magnitude sensitivity in a power grid can function as a proper distance metric[1], under some conditions.

[1]To qualify as a formal distance metric, each distance $d(a,b)$ must be non-negative, all self distances $d(a,a)$ must be zero, and any combination of three distances must satisfy the triangle inequality: $d(a,c) \leq d(a,b) + d(b,c)$.

The distance metric used in this paper captures the marginal impact of active-power transactions between nodes in a network on voltage phase-angle (θ) differences between the nodes. Specifically, our electrical distance metric $e_{a,b}$ estimates the incremental change in phase angles that would result from an incremental increase in active power flow from bus a to bus b. This is done by computing the "resistance distance" matrix [27] from the quadrant of the power flow Jacobian corresponding to real power injections and voltage phase angles ($\mathbf{J}_{P\theta}$). If we assume that voltages are held nearly constant by reactive power resources throughout a network ($\Delta V = \mathbf{0}$), then the incremental change in nodal power injections is given by:

$$\Delta P = \mathbf{J}_{P\theta} \Delta\theta.$$

Assuming that all branches in the network are symmetric (i.e., with nominal tap-changers and phase shifters), and neglecting shunt capacitance in transmission lines, $\mathbf{J}_{P\theta}$ functions as a symmetric Laplacian matrix describing the power network as a weighted graph. Under the DC power flow assumptions $\mathbf{J}_{P\theta}$ is the susceptance matrix $\mathbf{B} = \Im(Y_{BUS})$. If we let $\mathbf{J}^{+}_{P\theta}$ be the pseudo-inverse [28] of $\mathbf{J}_{P\theta}$, then each element ($e_{a,b}$) of the electrical "reactance" distance matrix (analogous to resistance distance) is:

$$e_{a,b} = (\mathbf{J}^{+}_{P\theta})_{a,a} - (\mathbf{J}^{+}_{P\theta})_{a,b} - (\mathbf{J}^{+}_{P\theta})_{b,a} + (\mathbf{J}^{+}_{P\theta})_{b,b}, \qquad (1)$$

which is conveniently independent of the reference bus chosen for the network. This results in a distance matrix ($\mathbf{E}$), for which each element $e_{a,b}$ measures the incremental change in phase angle difference between nodes a and b ($\Delta\theta_a - \Delta\theta_b$) given an incremental active power transaction between nodes a and b. Since, if we assume that angle differences are small and that voltages are nominal, incremental phase angle differences and incremental reactive power dissipation are the same, $\mathbf{E}$ also provides the reactive power dissipation that would result from orthonormal current excitations.

In [23] we found that $\mathbf{E}$, thus defined, satisfies the conditions for a proper distance metric so long as all series branch reactances are non-negative, and shunt capacitances are small. Note that this definition differs from the simpler metric used in [4], which was based on the simple inverse of the Y_{BUS} matrix. Since the metric in [4] did not have zeros on the diagonal, it can not be considered a formal distance metric; however, numerical comparisons indicated that the two metrics are strongly correlated.

It is important to note that electrical distances differ substantially from topological distances (d_{ij}, the number of branches that must be utilized to travel a topological path between Bus i and Bus j) in power grids. The power systems literature contains a number of examples where statistical clustering has been employed using topological distance metrics (e.g., [1], [2], [29]), but topological methods will not generally produce cohesive clusters from an electrical perspective [4]. The partitioning method developed in this paper, based on electrical distance, produces zones that are electrically cohesive, reduces loop flows (transaction leakage) between zones, and in at least one case identifies buses with dynamically coherent voltage angles (see Sec. V).

III. Measuring the quality of partitioning solutions

In power system applications, objectives beyond inter- and intra-cluster cohesiveness may be important. This section describes metrics used to evaluate partitioning solutions. The example metrics developed here were developed for problems such as transmission planning or zonal scheduling, but also illustrate how our approach can be used in combination with a variety of attributes. The first two quality metrics are based on electrical distances, the third and fourth measures relate to cluster sizes, and the fifth index measures connectedness. For consistency, each index is normalized to fit within the range [0,1] with 1 indicating highest quality.

A. Electrical Cohesiveness Index (ECI)

From the electrical distance matrix $\mathbf{E}$ it is straightforward to measure the total intra-cluster distance, for a given clustering solution C,

$$\hat{e}(C) = \sum_{a=1}^{n} \sum_{b \in M_a} e_{ab}. \tag{2}$$

where M_a is the set of buses that are in the same cluster as bus a. $\hat{e}(C)$ ranges between zero, when all nodes are in separate clusters, and the sum of all elements in $\mathbf{E}$ when all nodes are in a single, fully-connected cluster. For most applications, high-quality clusters will have lower electrical distances between nodes in each cluster (and thus high intra-cluster cohesiveness).

The Electrical Cohesiveness Index (ECI) uses $\hat{e}(C)$ to measure the extent to which the buses within each cluster are electrically proximate to other cluster members, as measured by the extent to which phase angles within a cluster will react in concert given a change in power injections within the cluster.

$$ECI = 1 - \frac{\hat{e}(C)}{\hat{e}_{\max}} = 1 - \frac{\sum_{a=1}^{n} \sum_{b \in M_a} e_{ab}}{\sum_{a=1}^{n} \sum_{b=1}^{n} e_{ab}}. \tag{3}$$

Equation (3) evaluates to one when all nodes are in separate clusters ($p = n$), because every node is "perfectly" connected to every other node in each cluster. Conversely, it evaluates to zero when all nodes are in one cluster, reflecting the fact that randomly chosen node pairs within very large clusters are, on average, electrically distant from one another.

B. Between-Cluster Connectedness Index (BCCI)

ECI incorporates the within-cluster distances for a clustering solution, but does not consider the connections across zonal boundaries. If a cluster boundary cuts through a low-impedance connection, such as a very short transmission line or a transformer, the effect on ECI would be insignificant. Therefore, we use (4) to measure the strength of connections between clusters.

$$h(C) = \sum_{a=1}^{n} \sum_{b \notin M_a} \frac{1}{e_{ab}} \tag{4}$$

The use of the inverse distance ($1/e_{ab}$) increases the contribution of low-impedance connections to h. This measure is similar to the graph "efficiency" measures proposed in [30] and adapted for power grids in [31].

The Between-Cluster Connectedness Index (BCCI) measures the extent to which buses in different clusters are loosely connected to one another, based on (4). Unlike ECI, BCCI evaluates to one when all nodes are in the same cluster, because there are no cross-cluster connections. For an atomistic solution ($p = n$), BCCI is zero because all nodes are strongly connected to nodes outside of their clusters.

$$BCCI = 1 - \frac{h(C)}{h_{\max}} = 1 - \frac{\sum_{a=1}^{n} \sum_{b \notin M_a} {}^{1}/_{e_{ab}}}{\sum_{a=1}^{n} \sum_{\substack{b=1 \\ a \neq b}}^{n} {}^{1}/_{e_{ab}}} \tag{5}$$

In (5), the numerator, $h(C)$, is the sum of connection strengths between clusters, while the denominator is the maximum possible $h(C)$.

C. Cluster Count Index (CCI)

The Cluster Count Index (CCI) measures the proximity of the number of clusters in a given clustering solution, p, to a predetermined ideal number of clusters, p_*. p_* is a user-defined parameter, based on the assumption that the user of the method has some preference for the number of clusters that result. We define CCI using the shape of the log-normal probability density function with its mode set at p_*, as follows:

$$CCI = e^{\frac{-(\ln p - \ln p_*)^2}{2\sigma^2}}, \tag{6}$$

where $\sigma = w \ln(n)$. This shape is desirable because it gives CCI$= 1.0$ when $p = p_*$, and approaches zero as $p \to n$. The parameter w sets the width of the fitness function relative to n; that is, w is effectively a penalty factor, with larger values for w increasing the penalty for p being far from p_*. The results in this paper use $w = 0.05$.

D. Cluster Size Index (CSI)

The Cluster Size Index (CSI) evaluates the extent to which the cluster sizes deviate from the ideal cluster sizes of $s_* = n/p_*$. To obtain CSI we measure the size of each cluster, and then obtain a weighted average of cluster sizes:

$$\bar{s} = \sum_{i=1}^{n} s_i / n \tag{7}$$

where s_i is the size of the cluster that node i resides in. Note that by summing over each node, rather than over each cluster, the result is a weighted average, rather than a simple average, of cluster sizes. As with CCI, CSI follows the shape of the log-normal distribution with the width parameter $\sigma = w \ln n$:

$$CSI = e^{-\frac{(\ln \bar{s} - \ln s_*)^2}{2\sigma^2}}. \tag{8}$$

E. Cluster Connectedness (CC)

By definition, a cluster is a set of nodes that are physically linked to one another. All buses in a cluster, therefore, should be reachable by traversing links within that cluster. To enforce this definition we define Cluster Connectedness (CC) as a binary measure that evaluates to zero when any cluster is not fully connected, and one if all clusters are fully connected.

F. Aggregate clustering fitness

To evaluate the aggregate quality (or fitness, f) of a given clustering solution, we use a multiplicative aggregate fitness function, calculated as the weighted product of the five quality measures above:

$$f = ECI^{\alpha} \cdot BCCI^{\beta} \cdot CCI^{\gamma} \cdot CSI^{\zeta} \cdot CC \qquad (9)$$

where $\{\alpha,\ \beta,\ \gamma,\ \zeta\} \in [0, 1]$ are user-defined scalars that define the relative importance of ECI, $BCCI$, CSI, CCI respectively. Eq. (9) uses the product of the five measures for three related reasons. First, we seek an aggregate fitness function satisfying "preferential independence" of the five individual quality metrics. Preferential independence is defined by the condition $\frac{\partial^2 f}{\partial a \partial b} \neq 0$ for any two values of distinct individual quality metrics a and b; thus any multiplicative fitness function would satisfy preferential independence (additive fitness metrics, for example, will not satisfy preferential independence). Using a fitness function that does not satisfy preferential independence may produce clustering solutions with high total fitness, but low scores for one or more quality metrics. Second, a multiplicative form of the fitness function gives $f = 0$ whenever any cluster is not fully connected. Third, the interaction of the five individual clustering quality metrics prevents our evolutionary algorithm (Sec. IV) from pre-maturely converging on trivial solutions, such as creating clusters with only one node.

IV. A hybrid K-means/Evolutionary Algorithm for multi-attribute network partitioning

Conventional partitioning algorithms such as spectral and K-means approaches are computationally efficient, but are not easily adapted to produce solutions that are optimal with respect to objectives beyond those of maximizing between-cluster distances or minimizing within-cluster distances. This section presents a hybrid of the K-means algorithm [32] and an evolutionary computational algorithm (EA), which can be used to optimize with respect to a multi-attribute objective function. The K-means algorithm is used to generate an initial set of candidate solutions and the EA is used to improve the initial solutions according to the fitness function (9).

A. K-Means algorithm implementation

K-means clustering uses a top-down, or divisive, approach that begins with a complete network, divides the network into clusters, and finally adjusts those clusters based upon some criteria. The aim of the K-means algorithm is to divide the n nodes in the network into K clusters so that the within-cluster distances are minimized [32]. The algorithm starts by randomly choosing K nodes within the network as centroids for new clusters. The remaining nodes are subsequently assigned to the closest of the K initial centroids. The K centroids are then relocated to the node in each cluster that minimizes the mean distance between the centroid and other nodes with the cluster. Each node then calculates the distance between itself and the K points, and reassigns itself to the cluster associated with the nearest centroid. The method iterates until the movement of the K centroid points falls below some minimum threshold and a stable set of clusters is obtained.

In our implementation electrical distances were used for the distance metric, and we chose $K = p*$ as the (exogenously determined) optimal number of clusters.

B. Evolutionary algorithm implementation

Genetic algorithms, a type of evolutionary algorithm (EA), can be very effective at solving non-convex optimization problems, particularly when solutions can be represented as strings of numbers, and the quality of solutions can be represented using a single objective [33]. In this paper we adapted the standard genetic algorithm (henceforth referred to as a evolutionary algorithm, EA) to the problem of finding clustering solutions that maximize the fitness function in Eq. (9). The following sections describe the methods used to represent solutions in the EA, to generate the initial population of solutions, and to implement selection, crossover and mutation.

1) Initial population: The initial populations for the EA were produced using a combination of random clusters and clusters generated from the K-means algorithm. Random clusters were generated by selecting random nodes as cluster centroids and iteratively expanding the clusters to encompass neighboring nodes until each node is assigned to exactly one cluster. Both the random and K-means initial solutions were selected to have a CSI score greater than 0.9 in order to produce a balanced set of initial conditions with respect to cluster sizes. The random solutions provided the EA with a wide variety of different solutions, whereas the K-means solutions provided a group of good solutions to improve upon.

2) Representation of solutions: In a standard genetic algorithm (a type of evolutionary algorithm) each solution is represented as a string of (typically binary) numbers. This string is known as the genotype, since it encodes the actual solution (the phenotype) into an abstract representation. In our clustering EA, each solution is represented as a string of $n_g = n$ integers ($\mathbf{g} = [g_1 \dots g_n]$), where n is the number of nodes in the network (see Fig. 1). In this representation, each g_i is an index between 0 and the number of topological neighbors for Node i (m_i, which is equivalent to the number of buses adjacent to Bus i). When $g_i = 1$, Node i is located in the same cluster as its first neighbor; when $g_i = 2$, it is located in the same cluster as its second topological neighbor; etc. When $g_i = 0$, Node i is not necessarily placed in the same cluster as any other node, however it may end up clustered with one or more of its neighbors (e.g., Node k with connection g_k) if g_k indicates a connection to Node i. Every solution within the bounds $0 \leq g_i \leq m_i\ \forall i$ is a valid solution to the cluster problem (with a CC score of 1, Sec. III-E). A change to a single byte can divide a cluster into two clusters, or merge it with a neighboring cluster. This flexibility increases the power of mutation and crossover operations, which are the primary methods for search in a genetic algorithm (see Sec. IV-B3).

In order to evaluate the benefits of encoding the solutions with the integer genotype described above, we designed an experiment that compares the integer approach with the binary representation that was used in [4]. The binary genotype

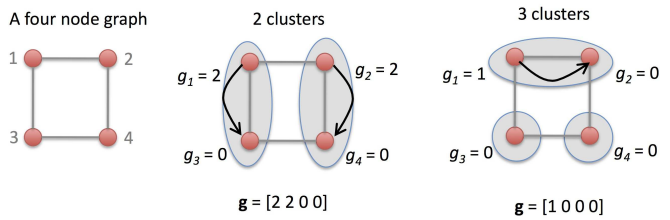

Figure 1. Illustration of three solutions (phenotypes) to a four-node clustering problem, using the integer genotype described in Sec. IV-B2. Each byte g_i in the genotype indicates which of Node i's neighbors are in the same cluster as Node i. Each Node numbers its neighbors sequentially, such that in the above, Node 1's first neighbor ($g_i = 1$) is Node 2 and its second neighbor ($g_i = 2$) is Node 3.

consisted of a string of $n_g = r$ bits ($\mathbf{g} = [g_1 \dots g_r]$), where r is the number of branches in the network. For a given branch (i), the state $g_i = 1$ indicates that the endpoint buses for branch i are located in the same cluster. The state $g_i = 0$ indicates the location of a potential boundary. The experiment consisted of applying the clustering algorithm to the test system in Sec. V-D with a flat start (all the initial solutions are random) and a predetermined maximum number of generations. We then measured the fitness of the solutions obtained when using the binary genotype versus using the integer genotype, while keeping the rest of parameters constant for both runs. Figure 2 illustrates the fitness evolution for both representations given an identical set of initial solutions.

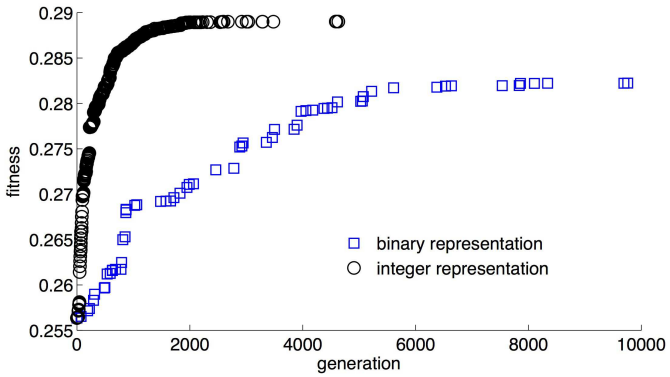

Figure 2. The fitness of the EA-produced partitioning solutions, by generation, for two alternate representation methods.

The results clearly show that the EA with the integer representation achieves substantially better solutions in far fewer generations. This is the case because, in the integer representation, a smaller number of mutations are required to produce substantially different solutions. In the binary representation, in order to separate a given subgraph into two subgraphs, every bit along a cut-set of that subgraph would need to mutate from one to zero. In the integer representation a single integer mutation can separate a subgraph into two subgraphs. In the integer representation, distinct solutions are thus a smaller distance from one another, which dramatically improves the power of crossover and mutation.

3) Selection, recombination and mutation: Our EA selected individuals to combine (crossover) and preserve between generations (elitism) based on fitness scores from (9). Parents for crossover operations are selected using the standard tournament selection method without replacement [34], [35], in which the selection of each individual depends upon a fitness comparison with a given number (tournament size) of other fitness scores from different individuals. In each generation the EA generates a set of new individuals equal to 80% of the population through crossover. The new individuals replace existing individuals probabilistically, using the roulette wheel method. The crossover process is summarized as follows:

1) Select two parents using a tournament algorithm without replacement that compares the fitness scores.
2) Choose a single random point in the genotype pair at which both parents are split (single-point crossover).
3) Create a new individual with the vector head of one parent and the vector tail of the second.
4) Replace an individual in the population with the new individual, with replacement probabilities proportional to fitness.

In addition, the top three individuals are retained without modification at each generation (elitism). For the rest of the population, mutation occurs at the end of each generation. When a byte mutates it is randomly reset to an integer in the feasible set $g_{i,new} \in \{0, 1, \dots, m_i\}$. We use a mutation probability of $1/n$, so that we get approximately one mutation per individual per generation.

C. Choosing weights for multiple objectives

The fitness function (9) uses a multiplicative form in order to ensure that the EA penalizes solutions that have low scores in any of the four dimensions. However, in some applications some of the objectives are less important than others. In order to adjust the weights in (9), we specifically tested a variety of solutions for a specific application that requires zones (reducing loop flows, for example, Sec. V-A), and used multiple-regression to determine the relative importance of the four objectives to the desired outcome. Sec. V-D discusses the application of this approach to a relatively large system.

V. Results

This section illustrates the proposed partitioning algorithm using three different test systems, the relatively small IEEE-RTS-96 and IEEE-118 test systems and a larger model that represents the power grid in Poland.

A. Measuring transaction leakage (loop flow)

A common application for zonal analysis in power system planning and operations is monitoring the impact of transactions between distant locations. In transmission planning applications, for example, it is desirable that intra-zonal transactions do not significantly affect currents, voltages or power flows outside of the zone. In reserves scheduling, the importance of deliverability of reserves has recently been emphasized, i.e. reserves themselves should not face curtailment because of transmission constraints and the dispatch of reserves should not cause additional congestion elsewhere in the system [36].

In this section we test the hypothesis that when zones are designed to have higher BCCI and ECI scores, while controlling for the number and size of clusters (CCI, CSI), intra-zonal real power transactions result in less current flow changes on branches outside of the zone. If B_z is the set of buses in zone z, Eq. (10) provides a measure of average impact of transactions within zone z on currents in branch m in the network.

$$T_z(m) = \frac{\sum_{\forall a,b \in B_z} \max(I_m(a \to b) - I_m(0), 0)}{|B_z|(|B_z| - 1)} \qquad (10)$$

where $I_m(0)$ is the current magnitude (in per unit) on branch m before the transaction, and $I_m(a \to b)$ is the (p.u.) current on the same branch after adding an additional 1 MW (0.01 p.u.) of demand at bus b and allowing a "slack bus" at bus a to meet this additional demand. We use the term T_z to refer to the average $T_z(m)$ over all branches m that are outsize of zone z, and $\overline{T_z}$ to be the average T_z over all zones in a given solution to the clustering problem. The currents $I_m(0)$ and $I_m(a \to b)$ were calculated using a standard Newton-Raphson ac power flow. Transaction leakage thus measures the average amount that currents (in per unit) change outside of a zone, given transactions within a zone. Note that because we averaged over a large number of branches, many of which are far from a particular zone, and because the numbers are expressed in per unit, the numbers are small (on the order of 10^{-5} to 10^{-4} per unit). Because transaction leakage is in some contexts referred to as "loop flow", this paper uses the two terms interchangeably.

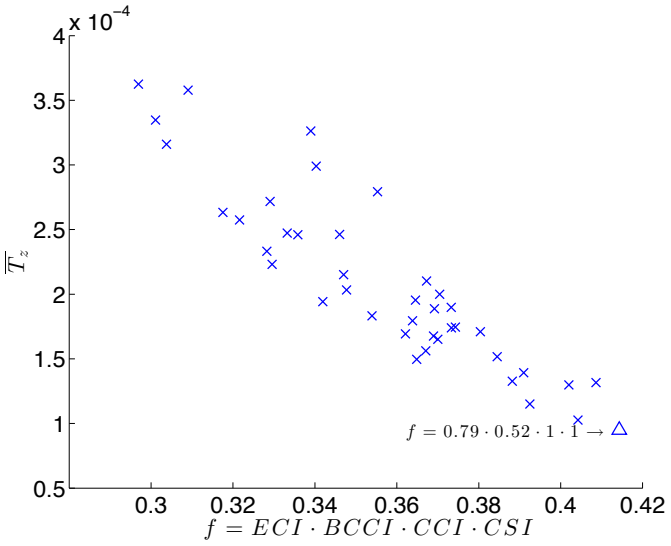

Figure 3. Scatter plot showing clustering quality (f) and the amount of loop flow ($\overline{T_z}$) for random clustering solutions ($\times$) and a solution generated by the EA ($\triangle$) for the IEEE-RTS-96 network and 3 clusters.

Table I
LINEAR REGRESSION RESULTS FOR $\overline{T_z}$ AS A FUNCTION OF ECI, $BCCI$ AND CSI IN THE RTS-96 CASE.

	B_k[1]	T stat.	P-val
intercept	-12.3	-42.6	$< 10^{-3}$
ECI	-6.7	-3.59	$< 10^{-3}$
$BCCI$	-2.6	-2.82	$< 10^{-2}$
CSI	0.5	0.33	0.74678

1 The coefficients B_k correspond to: $\ln(\overline{T_k}) = \ln(\text{intercept}) + B_1 \ln(ECI) + B_2 \ln(BCCI) + B_3 \ln(CSI)$.

B. Test system: IEEE-RTS-96

The IEEE reliability test system 1996 constitutes one of the most commonly utilized benchmarks for power system analysis [37]. It is composed by 73 buses, 120 transmission lines and 99 generators.

To illustrate how the use of our clustering method (fitness function and evolutionary algorithm) could improve the performance of planning applications, we partition the IEEE-RTS-96 bus test case into three zones, and compare the quantity of transaction leakage to the fitness scores f. We make this comparison for 100 randomly generated divisions of the IEEE-RTS-96 case into three zones (using the random centroid method) and the best solution that resulted from our evolutionary algorithm. To avoid distortions that might result from unbalanced cluster sizes, we compared only random solutions with a CSI score of 0.9 or higher.

The results shown in Fig. 3 indicate a strong, and statistically significant, negative correlation between f and $\overline{T_z}$ (see Fig. 3). We take this as strong evidence in support of our hypothesis that defining network partitions with high electrical cohesiveness reduces transaction leakage. Note that the values of $\overline{T_z}$ computed are small, largely because the size of the initial transaction was small (0.01 p.u.) and because the results show an average over all transmission lines outsize of the zone (many of which are quite far from the zone, and thus will have very small numbers for $T_z(m)$ from Eq. (10).

Additionally, the solution achieved by the EA (marked with a triangle in Fig. 3) yields identical clusters to those found by the dynamic simulation method proposed in [15]. This provides evidence that using electrical distances for partitioning can identify generators that are likely to be dynamically coherent.

C. Test system: IEEE-118

The IEEE-118 bus test case consists of 118 buses and 186 branches and comes from a reduced model of the Midwestern US power grid in 1962 [38]. We follow the same procedure described in Sec. V-B to initialize the EA and obtained a 3-cluster partition of the IEEE-118 test case that substantially improves the multi-objective goal with respect to the random solutions. Figure 4 and Table II further support the hypothesis that partitions with high electrical cohesiveness reduce loop flows.

In order to test the hypothesis that our new distance metric from (1) better correlates with $\overline{T_z}$ than the metric in [4], we generated 100 random divisions of the 118 bus case into 3 clusters, and evaluated the four metrics from (9) (neglecting CC, since all solutions had $CC = 1$) using the old and new distance metrics. Linear regression was used (as in Table II) to predict $\overline{T_z}$ from the four metrics. The model using distance as defined in (1) had an R^2 value of 0.75, vs. $R^2 = 0.61$ for the old distance measure. Note that the predictiveness would further decrease if $BCCI$ were not included in the regression, as was the case in [4]. We interpret this as evidence supporting

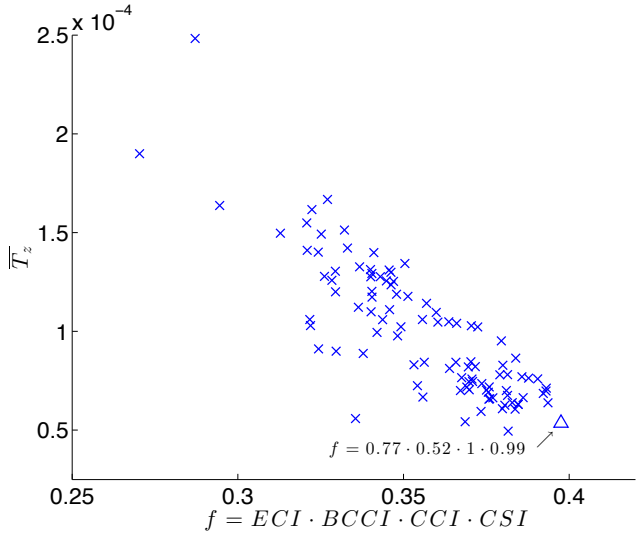

Figure 4. Scatter plot showing clustering quality (f) and the amount of loop flow ($\overline{T_z}$) for random clustering solutions ($\times$) and a solution generated by the EA ($\triangle$) for the IEEE-118 network and 3 clusters.

Table II
LINEAR REGRESSION RESULTS FOR $\overline{T_z}$ AS A FUNCTION OF ECI, $BCCI$ AND CSI IN THE IEEE-118 CASE.

	B_k[1]	T stat.	P-val
intercept	-13.5	-52.3	$< 10^{-3}$
ECI	1.01	0.6	0.5506
$BCCI$	-6.08	-8.54	$< 10^{-3}$
CSI	-5.94	-4.37	$< 10^{-3}$

1 The coefficients B_k correspond to: $\ln(\overline{T_k}) = \ln(\text{intercept}) + B_1 \ln(ECI) + B_2 \ln(BCCI) + B_3 \ln(CSI)$.

the hypothesis that the new measure is superior in terms of predicting transaction leakage.

D. Test system: Poland, case2383wp

To test this algorithm on a realistic, relatively large power system, we use data from the Poland grid, corresponding to a snapshot of the grid for a winter peak load profile [39]. After collapsing leaf-nodes[2] (which will cluster within the same set as their immediate upstream neighbors) the network comprises 1733 buses and 2240 transmission lines. The following algorithm summarizes the application of the method proposed in Sec. IV to this specific case and illustrates how to manage multiple objectives, to produce partitioning solutions for a large system.

1) *Merge the leaf-nodes (nodes with exactly 1 other connection) with their immediately connected neighbor.*
2) *Generate an initial population of random and K-means clustering solutions for calibration*. The total number of calibration solutions in this example is 40 (20 random + 20 K-means).
3) *Choose the fitness function coefficients according to the relative importance of the metrics with respect to the specific application*. In this case, we fit a linear model for $\overline{T_k}$ and decide that given the strong (anti-)correlation with $BCCI$, an appropriate weighting would be $f = ECI^{0.8} \cdot BCCI^{1} \cdot CCI^{1} \cdot CSI^{0.8}$ (see Table III). Because the total number of clusters is an external constraint usually given by a client (e.g. a regional transmission organization), this combination of coefficients also maintains a strong selection preference toward CCI.
4) *Generate the EA initial population.* In this case we construct an initial population of 400 individuals, combining the 40 calibration solutions from Step 2 and 360 additional random solutions.
5) *Run the EA*. We evolved the set of solutions according to the parameters described above and saved a set of 273 improved solutions over 6,000 generations.
6) *Calculate the Pareto set of solutions from all the EA solutions*. We further refined the results by choosing the 33 non-dominated solutions with respect to the individual metrics (represented by triangle markers in Fig. 5).
7) *Calculate $\overline{T_k}$ on the reduced set of solutions and select a final partition that is optimal with respect to the weighted multiple objectives*. (See Fig. 5, the red-shadowed symbols compose the Pareto optimal set from within the full aggregate set of 73 solutions)

[2]Leaf-nodes are those that only have one connection.

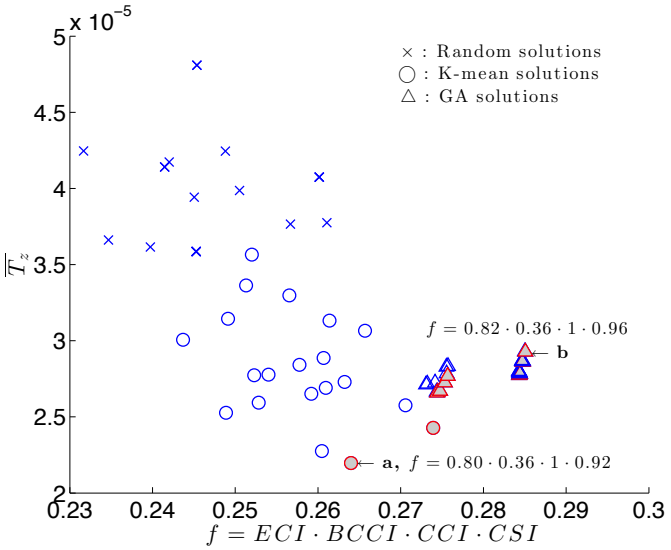

Figure 5. Scatter plot showing clustering quality f and the amount of loop flow ($\overline{T_z}$) for the Polish power grid, with $p* = 5$. The red-shadowed symbols are Pareto optimal with respect to $\overline{T_z}$ and f. Two of these optimal solutions $\{a, b\}$ are further illustrated in Fig. 6.

Table III
LINEAR REGRESSION RESULTS FOR $\overline{T_z}$ AS A FUNCTION OF ECI, $BCCI$ AND CSI FOR THE POLISH CASE (INITIAL SOLUTIONS).

	B_k[1]	T stat.	P-val
intercept	-15.8	-27.4	$< 10^{-4}$
ECI	11.9	4.88	$< 10^{-4}$
$BCCI$	-7.1	-10.55	$< 10^{-4}$
CSI	-8.7	-6.22	$< 10^{-4}$

1 The coefficients B_k correspond to: $\ln(\overline{T_k}) = \ln(\text{intercept}) + B_1 \ln(ECI) + B_2 \ln(BCCI) + B_3 \ln(CSI)$.

The top panel of Fig. 6 illustrates solution '**a**' in Fig. 5, which was generated from the K-means algorithm and has small average loop flow ($\overline{T_z}$). However, the balance of cluster

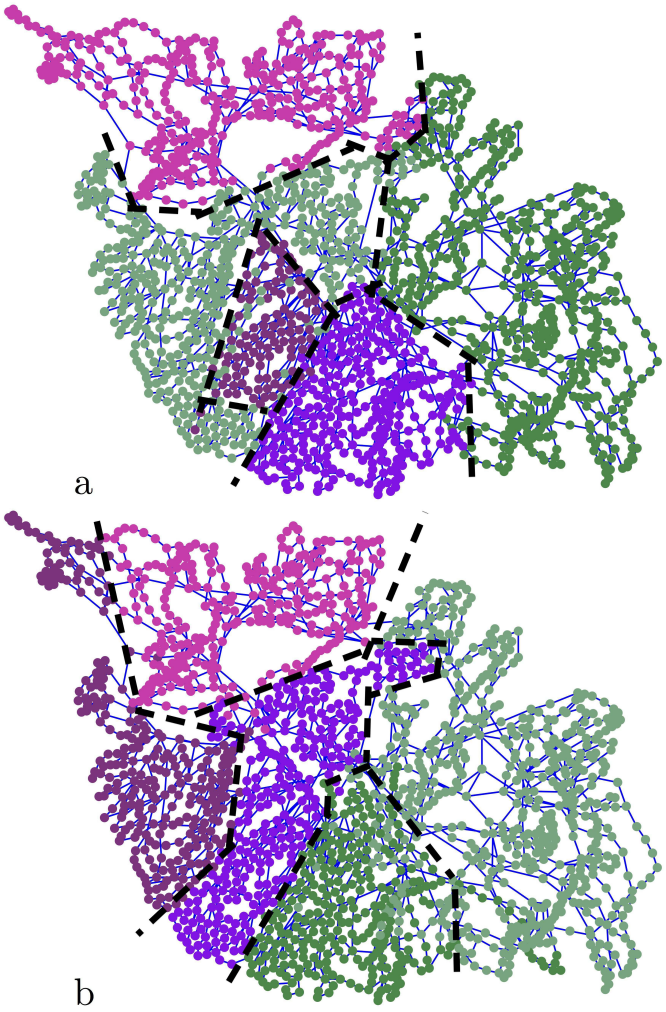

Figure 6. Two partitions of the Polish power grid that are Pareto optimal with respect to f and $\overline{T_z}$. The fitness scores of these solutions are $f_a = 0.80 \cdot 0.36 \cdot 1 \cdot 0.92$ and $f_b = 0.82 \cdot 0.36 \cdot 1 \cdot 0.96$ for $f = ECI \cdot BCCI \cdot CCI \cdot CSI$.

sizes (CSI) in solution 'a' is substantially worse than that of the EA solution ('b' in Fig. 5). For a small increase in T_z, the most fit EA solution (solution 'b') is optimal in a Pareto sense, and substantially superior in terms of total fitness.

In order to compare our method with spectral clustering approaches [40] we generated and evaluated a clustering solution based on the Fiedler Vector [41] for this test case. The loop flow evaluation for this solution yielded $\overline{T_z} = 2.26 \times 10^{-5}$ which makes it comparable to a good K-means solution. The overall fitness score of this solution, from Eq. (9), was $f = 0.259$. While this solution is competitive with the K-means approach, it would not qualify for the Pareto front in Fig. 5.

VI. Conclusions

This paper presents a multi-attribute, hybrid method for dividing a power system into electrically coherent partitions (zones), using electrical distances. Using a K-means algorithm in combination with an evolutionary computational algorithm produces solutions that balance several measures of quality (attributes). Our EA includes a novel integer representation that allows for efficient exploration of the search space. The method is not designed to optimize zonal partitions for a specific application. Rather, this paper presents a general approach to the power network partitioning problem whose solutions have demonstrably advantageous properties for some classes of power system applications, and whose formulation could be tailored to specific operational or planning problems. An application of the method to the 76 bus IEEE Reliability Test Case produced a partitioning solution that groups buses that respond coherently to dynamic disturbances. A study of the extent to which intra-zonal transactions impact extra-zonal currents (loop flows) in two test cases (the IEEE 118 bus and the Polish network), showed that clustering based on electrical distances can reduce unwanted loop flows. This property of our clustering solutions seems advantageous for system security applications, such as location-specific load-shedding. The localized (intra-cluster) response to active-power perturbations also suggests that electrical-distance clustering could be utilized in wide-area monitoring schemes, or cascading failure analysis [42].

The general results in this paper suggest opportunities for future research on electrical-distance based partitioning schemes for particular applications. For example, transmission topology reconfiguration problems, such as transmission expansion planning and optimal transmission switching, are often intractable on large power systems [43], [44]. By using zonal definitions that minimize loop flows, the inter-zonal effects of topology modifications are limited and topology reconfiguration problems may be applied to distinct zones/areas with greater accuracy.

References

[1] V. H. Quintana and N. Muller, "Partitioning of power networks and applications to security control," *IEE Proceedings C*, vol. 138, no. 6, pp. 535–545, 1991.

[2] A. O. M. Saleh and M. A. Laughton, "Cluster analysis of power-system networks for array processing solutions," *IEE Proceedings C*, vol. 132, no. 4, pp. 172–178, 1985.

[3] S. Walton and R. D. Tabors, "Zonal transmission pricing: Methodology and preliminary results from the WSCC," *The Electricity Journal*, vol. 9, no. 9, pp. 34–41, 1996.

[4] S. Blumsack, P. Hines, M. Patel, C. Barrows, and E. Cotilla-Sanchez, "Defining power network zones from measures of electrical distance," *Proc. of the IEEE Power and Energy Society General Meeting*, 2009.

[5] G. Kron, "A set of principles to interconnect the solutions of physical systems," *Journal of Applied Physics*, vol. 24, no. 8, pp. 965–980, 1953.

[6] H. Happ, "The operation and control of large interconnected power systems," *IEEE Trans. Circuit Theory*, vol. 20, no. 3, pp. 212–222, 1973.

[7] H. Happ, "Diakoptics, the solution of system problems by tearing," *Proceedings of the IEEE*, vol. 62, no. 7, pp. 930–940, 1974.

[8] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, vol. 5. Prentice Hall Upper Saddle River, NJ, 2002.

[9] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, 2002.

[10] M. E. J. Newman, "Communities, modules and large-scale structure in networks," *Nature Physics*, vol. 8, no. 1, pp. 25–31, 2012.

[11] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[12] L. Hang, A. Bose, and V. Venkatasubramanian, "A fast voltage security assessment method using adaptive bounding," *IEEE Trans. Power Syst.*, vol. 15, no. 3, pp. 1137–1141, 2000.

[13] J. Zhong, E. Nobile, A. Bose, and K. Bhattacharya, "Localized reactive power markets using the concept of voltage control areas," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1555–1561, 2004.

[14] Y. Wang, F. Li, Q. Wan, and H. Chen, "Reactive power planning based on fuzzy clustering, gray code, and simulated annealing," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2246–2255, 2011.

[15] I. Kamwa, A. Pradhan, and G. Joos, "Automatic segmentation of large power systems into fuzzy coherent areas for dynamic vulnerability assessment," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1974–1985, 2007.

[16] I. Kamwa, A. Pradhan, G. Joos, and S. Samantaray, "Fuzzy partitioning of a real power system for dynamic vulnerability assessment," *IEEE Trans. Power Syst.*, vol. 24, no. 3, pp. 1356–1365, 2009.

[17] H. You, V. Vittal, and X. Wang, "Slow coherency-based islanding," *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 483–491, 2004.

[18] M. Wang and H. Chang, "Novel clustering method for coherency identification using an artificial neural network," *IEEE Trans. Power Syst.*, vol. 9, no. 4, pp. 2056–2062, 1994.

[19] C. Juarez, A. Messina, R. Castellanos, and G. Espinosa-Pérez, "Characterization of multimachine system behavior using a hierarchical trajectory cluster analysis," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 972–981, 2011.

[20] J. Li, C. Liu, and K. Schneider, "Controlled partitioning of a power network considering real and reactive power balance," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 261–269, 2010.

[21] M. Ramezani, C. Singh, and M. Haghifam, "Role of clustering in the probabilistic evaluation of TTC in power systems including wind power generation," *IEEE Trans. Power Syst.*, vol. 24, no. 2, pp. 849–858, 2009.

[22] F. Vallee, G. Brunieau, M. Pirlot, O. Deblecker, and J. Lobry, "Optimal wind clustering methodology for adequacy evaluation in system generation studies using nonsequential Monte Carlo simulation," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2173–2184, 2011.

[23] E. Cotilla-Sanchez, P. Hines, C. Barrows, and S. Blumsack, "Comparing the topological and electrical structure of the North American electric power infrastructure," *IEEE Syst. J.*, 2012.

[24] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[25] P. Lagonotte, J. Sabonnadiere, J.-Y. Leost, and J.-P. Paul, "Structural analysis of the electrical system: application to secondary voltage control in France," *IEEE Trans. Power Syst.*, vol. 4, no. 2, pp. 479–486, 1989.

[26] Q. Lu and S. Brammer, "A new formulation of generator penalty factors," *IEEE Trans. Power Syst.*, vol. 10, no. 2, pp. 990–994, 1995.

[27] D. Klein and M. Randic, "Resistance distance," *Journal of Mathematical Chemistry*, vol. 12, pp. 81–95, 1993.

[28] T. N. E. Greville, "Some applications of the pseudoinverse of a matrix," *SIAM Review*, vol. 2, no. 1, pp. 15–22, 1960.

[29] Y. Jiaxi, Y. Lin, L. Ruiye, and G. Zhizhong, "Research on time process-oriented power system static security analysis," in *3rd Int. Conf. on Electric Utility Deregulation and Restructuring and Power Technologies*, pp. 1516–1521, IEEE, 2008.

[30] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical Review Letters*, vol. 87, no. 19, 2001.

[31] S. Arianos, E. Bompard, A. Carbone, and F. Xue, "Power grids vulnerability: a complex network approach," *Chaos*, vol. 19, 2009.

[32] J. Hartigan and M. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Journal of the Royal Statistical Society C*, vol. 28, no. 1, pp. 100–108, 1979.

[33] A. Eiben and J. Smith, *Introduction to evolutionary computing*. Springer Verlag, 2003.

[34] D. Goldberg, "A note on Boltzmann tournament selection for genetic algorithms and population-oriented simulated annealing," *Complex Systems*, vol. 4, no. 4, pp. 445–460, 1990.

[35] D. Goldberg, K. Deb, and J. Clark, "Genetic algorithms, noise, and the sizing of populations," *Urbana*, vol. 51, p. 61801, 1991.

[36] Y. Chen and P. Gribik, "Determining reserve zone requirements from the energy and ancillary service co-optimization," in *INFORMS Annual Meeting, Phoenix AZ*, October 2012.

[37] C. Grigg *et al.*, "The IEEE Reliability Test System-1996.," *IEEE Trans. Power Syst.*, vol. 14, no. 3, pp. 1010–1020, 1999.

[38] "Power Systems Test Case Archive," University of Washington, Electrical Engineering. online: http://www.ee.washington.edu/research/pstca/.

[39] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning and analysis tools for power systems research and education," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 12–19, 2011.

[40] B. Lesieutre, S. Roy, V. Donde, and A. Pinar, "Power system extreme event screening using graph partitioning," in *Proceedings of the 38th North American Power Symposium*, 2006.

[41] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its applications to graph theory," *Czechoslovak Mathematical Journal*, vol. 25, no. 100, pp. 619–633, 1975.

[42] M. J. Eppstein and P. D. H. Hines, "A 'Random Chemistry' algorithm for identifying collections of multiple contingencies that initiate cascading failure," *IEEE Trans. Power Syst.*, vol. 27, no. 3, 2012.

[43] C. Barrows and S. Blumsack, "Transmission switching in the RTS-96 test system," *IEEE Trans. Power Syst.*, vol. 276, no. 2, pp. 1134–1135, 2012.

[44] R. Bent, G. L. Toole, and A. Berscheid, "Transmission network expansion planning with complex power flow models," *IEEE Trans. Power Syst.*, vol. 27, no. 2, pp. 904–912, 2012.

Eduardo Cotilla-Sanchez (S'08, M'12) received the M.S. and Ph.D. degrees in electrical engineering from the University of Vermont, Burlington, in 2009 and 2012, respectively.

He is currently an Assistant Professor in the School of Electrical Engineering and Computer Science at Oregon State University, Corvallis. His primary field of research is the vulnerability of electrical infrastructure, in particular, the study of cascading outages.

Paul D. H. Hines (S'96, M'07) received the Ph.D. in Engineering and Public Policy from Carnegie Mellon University in 2007 and M.S. (2001) and B.S. (1997) degrees in Electrical Engineering from the University of Washington and Seattle Pacific University, respectively.

He is currently an Assistant Professor in the School of Engineering, and the Dept. of Computer Science, at the University of Vermont, and a member of the adjunct research faculty at the Carnegie Mellon Electricity Industry Center. He received an NSF CAREER Award in 2013, and is an associate editor of the IEEE Transactions on Smart Grid.

Clayton Barrows (S'07) received the B.S. degree in electrical engineering from the University of Wyoming, Laramie, in 2005, and the Ph.D. in energy management and policy from the John and Willie Leone Family Department of Energy and Mineral Engineering, Pennsylvania State University.

He is currently a postdoctoral researcher in the Energy Forecasting and Modeling Group at the National Renewable Energy Laboratory. His research interests include complex networks, large system optimization and energy policy, especially with respect to electrical power systems.

Seth Blumsack (M'06) received the M.S. degree in Economics from Carnegie Mellon in 2003 and the Ph.D. in Engineering and Public Policy from Carnegie Mellon in 2006.

He is currently Assistant Professor of Energy Policy and Economics at The Pennsylvania State University, University Park PA. His main interests are in the regulation and deregulation of network industries, environmental management and performance, transmission planning and pricing, and complex networks and systems, especially for energy and electric power.

Mahendra Patel (M'74, SM'81) received the B.E. (Electrical) from S. P. University, India, M.S.E.E. from West Virginia University and the M.B.A. from University of Pittsburgh.

He is currently a Sr. Business Solutions Engineer in the Applied Solutions Department at PJM Interconnection. He has 40 years of experience in Electric Power Industry, mainly in the areas of Transmission Planning and System Reliability, Transmission Technologies, System Dynamics, SynchroPhasor Technology, Switching Surges and Insulation Coordination, Voltage Stability, System Protection and Power Quality.