Feature Extraction and Visualization of Power System Transient Stability Results

Sudipta Dutta, *Member, IEEE*, and Thomas J. Overbye, *Fellow, IEEE*

***Abstract*—Transient stability simulation studies of large-scale power systems often generate large amounts of data. This can make it difficult for power system planners to understand the overall system response or to identify portions of the system with unusual signals. In this paper we present a novel approach that utilizes clustering to extract common features in the voltage magnitude and frequency response signals, and to identify outliers. The geographic visualization of the results is also discussed. Results are demonstrated using the IEEE 118-bus system and a 16 000-bus real-world system model.**

***Index Terms*—Clustering, feature extraction, model errors, power system visualization, spark lines, transient stability data.**

I. Introduction

TRANSIENT stability simulation tools are essential for enabling safe and reliable operations of power systems. In real life large scale systems with thousands of buses, transient stability studies often generate gigantic volumes of data. This poses a significant challenge to power system planners to analyze the overall system response and identify portions of the system which might be displaying "abnormal" behavior. An exhaustive analysis being prohibitively inefficient, it is of critical importance to develop methods to automatically extract such information from transient stability data. This paper addresses the above need by applying clustering techniques on transient stability data, more specifically, voltage and frequency response signals. Common features are extracted and outliers characterized by uncommon features are also identified. Furthermore, a visualization method is presented to infuse geographic information in the display of the results.

Information extraction from transient stability data is the first and most crucial part of this work. It will be shown how clustering [1] can be applied to identify a set of distinct signals that characterize the overall system behavior. Groups of coherent generators [2]–[5] can also be identified in a "model-free" approach. In a prior work [6] we have developed a clustering based methodology to identify groups of generators in a 2400-generator system with similar frequency responses to a large fault. While common features of the frequency signals were identified, outliers characterized by abnormal signals were intelligently flagged down in two of the generators in the same process. These abnormalities caused by errors in these generators' exciter models had previously been undetected and were subsequently corrected. This provided promise of the methodology as an operational tool. However, in this earlier work, only frequencies were considered which in a stable system vary within a narrow range both pre and post disturbance. Feature extraction from other power system quantities such as bus voltages poses greater challenges and has been addressed here. In the previous work, the choice of parameters for applying clustering techniques was not discussed. These have been addressed in this current paper.

The second part of this work focuses on visualization of the information extracted from transient stability data, which is time-varying in nature. Traditionally strip-charts have been used for this purpose. Such an approach can be quite effective provided one is only interested in showing a few signals. However, displaying multiple signals on a single plot using different colors becomes ineffective as the number of colors surpasses ten [7, p. 125]. In addition, such plots cannot show geographically distributed information. Wide area frequency visualization using animated event replays has been presented in [8] and [9]. Data collected from frequency disturbance recorders in a wide area FNET system are displayed with colored contours for every time step and played in the form of a movie. While certainly useful for some situations, it requires time to display the animation loop and cannot provide results at a glance. Again, it is extremely important to display time-varying data over a reasonable period as opposed to just showing the present value and say one time step prior, in order to convey the variation of signal value over time [10]. This visualization challenge has also been addressed in this paper.

The paper is organized as follows. A motivating example of the problem addressed in this paper is presented in Section II. The proposed analysis procedure is explained in Section III. The case study and discussions are presented in Section IV. Finally the conclusions are provided in Section V.

II. Motivating Example

Fig. 1(a) and (b) demonstrates how transient stability voltage variation is often shown for a large system with 16 000 buses. Fig. 1(a) shows the voltage signals at all of the buses over 20 s with a large generation loss contingency occurring at 2 s. The individual buses are not identified due to data confidentiality

Manuscript received June 14, 2013; revised August 20, 2013; accepted September 21, 2013. Date of publication October 04, 2013; date of current version February 14, 2014. This work was supported in part by the Stanford University Global Climate and Energy Project (GCEP) and in part by the U.S. Department of Energy (DOE) Consortium for Electric Reliability Technology Solutions (CERTS). Paper no. TPWRS-00777-2013.

The authors are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: sudiptamoonray@gmail.com; overbye@illinois.edu).

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

Digital Object Identifier 10.1109/TPWRS.2013.2283276

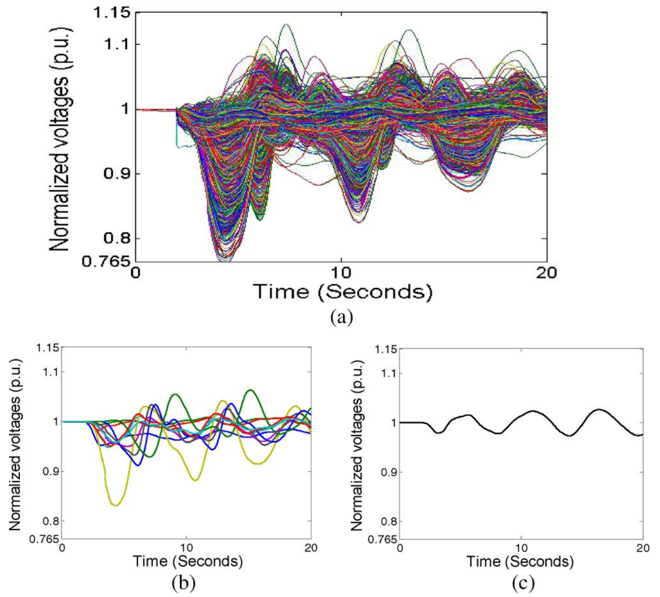


Fig. 1. (a) 16 000 transient bus voltage signals from a large system. (b) 11 randomly selected voltage signals from (a). (c) Outlier signal present in (a) but absent in (b), hence potentially missed out in system analyses.

concerns. Obviously with 16 000 buses the purpose of Fig. 1(a) is not to show the response at any individual location, but rather to give bounds on the overall system response. Power system planners often study the system by looking at a small number of pre-selected buses. While these locations are chosen with great care, they only represent a small sample of the overall voltage signals. To illustrate, Fig. 1(b) shows the voltage signals at eleven selected buses spread throughout the system. Note that by doing so some of the distinct features of the overall system response are indeed captured. However, this selection is not guaranteed to include all of the unique features, especially when these signals may vary depending on the assumed contingency. This failure is illustrated in Fig. 1(c) which shows an abnormal or outlier signal that is absent in Fig. 1(b) and stays hidden in Fig. 1(a). This abnormality could potentially indicate errors or condition warranting attention.

Another issue is with the visualization shown in Fig. 1(a)–(c). Just by studying the plots, it is not possible to locate where a certain set of signals originates within the geographic footprint of the system. This information can provide important insights into the behavior of the system such as identifying weak areas of the system prone to stability issues etc.

The clustering based proposed analysis procedure addresses these challenges. Important capabilities of this analysis procedure include extracting unique signals from the range of signals exhibited by the complete system based on common features. Similarities between signals are quantified followed by grouping together similar signals using clustering techniques. Finally a novel visualization technique is presented which embeds locational information with the transient stability signals.

III. Analysis Procedure

The analysis procedure involves *Feature Extraction* and *Visualization*. Feature extraction requires quantifying similarities

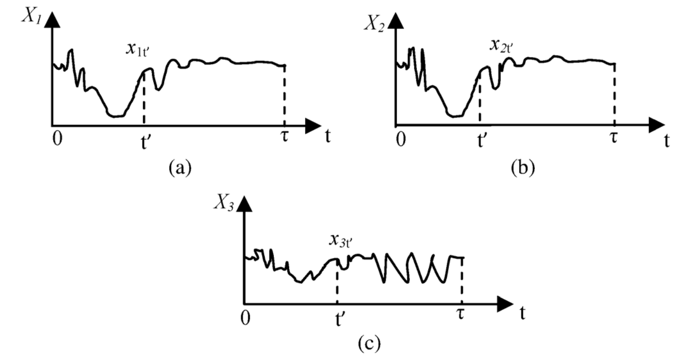


Fig. 2. Three τ-dimensional signals are shown in (a), (b), and (c). Visual inspection determines that (a) and (b) are similar and different from (c). So (a) and (b) should be clustered as one group.

between signals for which the *Similarity Measures* have been introduced. This is followed by *Clustering* similar signals which warrant a discussion on the clustering techniques used, namely Quality Threshold Clustering and K-Means Clustering.

The transient stability data described in this work is an $\mathrm{N}\times\tau$ volume of data where number of nodes (buses or generators whose signals are being studied) is N, and the number of time points considered is τ.

A. Feature Extraction

This can be described as the process of disentangling jumbled up transient stability signals into groups based on common features. Our studies reveal that certain transient stability data, such as voltage signals with varying ranges between buses, requires preprocessing for proper clustering. Preprocessing involves normalizing voltage signals with respect to pre-fault values. Frequency signals however vary within a similar range in stable systems both pre and post fault condition and can be analyzed without any preprocessing.

A similarity measure is required to compare multiple signals following this step.

Similarity Measures: Consider Fig. 2(a)–(c) which shows three signals in which the X-axis represents time and the Y-axis represents the signal value. By visual inspection the signals in Fig. 2(a) and (b) appear similar but the one in Fig. 2(c) does not. There are several different measures available to quantify signal similarity [11]. For example some measures are time shift invariant, some are amplitude shift invariant, some are time scale invariant, some are amplitude scale invariant, etc. With the focus on transient stability signals that are all correctly time stamped, the measures used here are invariant to amplitude shifts and scaling.

Mathematically, let X_p and X_q be τ-dimensional vectors, where

$$X_n = [x_{n_1}, x_{n_2}, \ldots, x_{n_\tau}]^T. \tag{1}$$

Then an amplitude scale invariant similarity measure will determine these similar if

$$X_p(\mathrm{t}) = X_q(\mathrm{t}) + \alpha. \tag{2}$$

An amplitude scale invariant similarity measure will determine the same two vectors similar if

$$X_p(\mathrm{t}) = (1+\alpha)X_q(\mathrm{t}). \tag{3}$$

In (2) and (3), α is a scalar term.

Two common categories of similarity measures are the distance measure and the proximity measure. One of the most common types of distance measure is the Mikowski distance defined as

$$d_M = \sqrt[r]{\sum_{t=1}^{\tau} (x_{p_t} - x_{q_t})^r}. \tag{4}$$

The popular Euclidean distance is a special case of Mikowski distance with $\mathrm{r} = 2$ in (4):

$$d_E = \sqrt{\sum_{t=1}^{\tau} (x_{p_t} - x_{q_t})^2}. \tag{5}$$

The normalized Euclidean distance can accept both amplitude shift and amplitude scaling.

Another distance measure that is very useful is the root mean square (RMS) distance given by

$$d_{RMS} = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau} (x_{p_t} - x_{q_t})^2}. \tag{6}$$

Using the RMS distance is preferable to using the Euclidean distance since in the former the choice of cluster diameter that controls cluster quality is invariant with length of the vector, τ.

Among proximity measures, a common method is use of correlation coefficients between different time series. This similarity measure is given by

$$d_{CC} = \frac{\sum_{t=1}^{\tau} (x_{p_t} - \overline{x_p})(x_{q_t} - \overline{x_q})}{\sqrt{\sum_{t=1}^{\tau} (x_{p_t} - \overline{x_p})^2}\sqrt{\sum_{t=1}^{\tau} (x_{q_t} - \overline{x_q})^2}}. \tag{7}$$

Terms with bars represent average values. Correlation measures are known to accept amplitude shifts such as DC offsets and amplitude scales.

Another proximity measure is the angle-based cosine measure:

$$d_{\cos} = \frac{\sum_{t=1}^{\tau} x_{p_i} x_{q_i}}{\sqrt{\sum_{t=1}^{\tau} x_{p_t}^2}\sqrt{\sum_{t=1}^{\tau} x_{q_t}^2}}. \tag{8}$$

Cosine similarity is a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. This measure accepts amplitude scale but not amplitude shifts.

Clustering: Transient stability voltage and frequency signals are clustered together based on feature similarities. Data mining literature provides several different types of clustering algorithms and proper algorithm selection is the key to effective separation of groups of signals. Clusters formed should be of high quality. Based on our prior work in this area and a survey of available clustering techniques, the algorithm we have applied is the Quality Threshold clustering algorithm [12].

1) Quality-Threshold (QT) Clustering: The goal of QT clustering is to form large clusters with similar patterns and to ensure a quality guarantee for each cluster. Cluster quality is defined by a pre-defined threshold diameter specified in terms of a similarity measure. This algorithm proceeds by forming candidate clusters around every τ-dimensional signal by adding signals which are within the range of the threshold diameter around it. The candidate cluster with largest cardinality is retained as the first QT cluster and signals grouped in the cluster removed. The procedure is iteratively repeated with the reduced set of signals until all signals are clustered. Further details of this algorithm can be found in available literature [13]. Typical QT algorithms have a restriction on minimal cardinality required to form a cluster. In this work, this restriction is removed allowing formation of even single response clusters, which often indicate "outliers" in the data. Another useful feature of QT algorithm is that it returns the same result when run several times.

The choice of quality threshold diameter dictates what value assigned to the similarity measure can classify two signals in one cluster or different clusters. There are no correct or incorrect choices and needs some operator experience. In the results section of this paper, we have provided a sensitivity analysis of the choice of this diameter to the number of clusters formed in a large 16 000-bus system. As the results will reveal, the choice of threshold diameter in this work is adequate. Note that lower the cluster threshold diameter, "tighter" the clusters are formed. Often this is tantamount to generation of many more clusters. Of course if all the signals are identical, no matter how small the threshold diameter is, only one cluster will be formed.

One of the major problems with QT clustering is computation complexity which manifests itself in the execution time. Complexity of the QT clustering algorithm could be as high as $O(\mathrm{N}^5)$ [14] or $O(\mathrm{N}^2)$ [15] depending on the implementation. This can lead to prohibitively large execution time with large value of N, which is the situation faced in this work with thousands of buses. Hence, our approach is to form preliminary clusters with a fast algorithm such as the well-known K-means clustering algorithm and then clustering median signals of each of these using the more accurate but computationally intensive QT algorithm to define the final good quality clusters.

2) K-Means Clustering: The application of K-means clustering can be considered as a preprocessing step to reduce data volume. Depending on the size of the system/data set, this step can be skipped. Issues of large data volumes have led to several research works [16]–[18]. Data clustering for reducing data volume has been proposed in the literature [19], [20]. Along these lines, in this paper we apply K-means clustering to generate the preliminary clusters.

K-means is a well-known clustering algorithm with low running time [21]. Its solution relies on an iterative scheme, which starts with arbitrarily chosen initial cluster memberships or centers. The distribution of signals among clusters and the updating of cluster centers are the two main steps of K-means algorithm. The algorithm alternates between these two steps until the value of the objective function (9) cannot be reduced anymore. It requires k, the number of clusters to be formed to be pre-specified. The choice of k is based on the total number of signals to be clustered. In the results section of this paper, we have provided a sensitivity analysis of the choice of k to the number of clusters formed in a large 16 000-bus system. As the results will reveal, the choice of k in this work is adequate.

Mathematically, the K-means algorithm determines a set of k τ-dimensional cluster centers $\{V_i \mid i = 1, 2, \ldots, k\}$, where

$$V_i = [v_{i_1}, v_{i_2}, \ldots, v_{i_\tau}]^T. \tag{9}$$

Given a set of N τ-dimensional signals $\{X_n \mid n = 1, 2, \ldots, N\}$, the objective function to be minimized is

$$J(U, V) = \sum_{i=1}^{k} \sum_{n=1}^{N} u_{in} \|X_n - V_i\| \tag{10}$$

s.t.

$$u_{in} \in \{0, 1\} \forall i, n \tag{11}$$

$$\sum_{i=1}^{k} u_{in} = 1 \forall n \tag{12}$$

where $\|.\|$ represents the similarity measure.

At the end of the *Feature Extraction* stage with two-stage clustering approach, the final groups of transient signals are obtained. The set of median signals from these final clusters capture the range of signals and the complete transient behavior of the system. This provides comprehensive understanding of the system dynamics.

B. Visualization

Here the visualization challenge of representing transient stability data is addressed. The simplest method of representation is with the help of plots and for multiple time series, with multi-color plots. But with increase of the number of series, the signals cannot be distinguished. Traditionally strip charts have been used for this purpose. But strip charts do not provide location information. An approach is to integrate small strip charts onto existing one-lines next to the field of interest [22]. This of course requires strip charts to be quite small. In our current work, the visualization technique used is "spark-lines" [23] on geographic overlays similar to the approach in [6].

As defined in [23], spark-lines are "data-intense, design-simple, word-sized graphics". They present a historical trend of data in the space of a typical word. A spark-line is a graph without axis labels and numbers. Obviously there is a tradeoff between display space and the amount of information shown. Spark-lines can only show data with several significant digits, but "the idea is to be approximately right rather than exactly wrong" [23, p. 50]. Unique signals found in the previous Feature Extraction process are displayed as spark-lines overlaid on one-line diagram or the actual latitude-longitude map, thus infusing location information. The x-axis time-scale could be common for all spark-lines, e.g., 20 s in a transient stability run. The y-axis could also be implicit based on the type of value, for example between 59.8 to 60.2 Hz for frequencies.

Spark-lines are automatically generated and laid out on the geographical map of the power system. It should be made sure that spark-lines do not overlap. This was achieved by developing overlap correction algorithms which automatically detect occurrence of overlaps and relocate spark-lines on the available display space using force based layout methods [24].

IV. Case Study and Discussions

The proposed analyses have been applied on transient stability data from two systems, namely the IEEE 118-bus system and a large real-world wide-area power system consisting of 16 000 buses. The analytic methods have been implemented in MATLAB with the transient stability data acquired by running different simulations using PowerWorld Simulator [25]. Note that the case studies are the heart of this application-centric work since they bring out the effectiveness and usefulness of the proposed analyses. The *Feature Extraction* studies have been organized into transient frequency and transient voltage studies. While both voltage and frequency response results are presented for the 16 000-bus case, only voltage response results are presented for the 118-bus system. This is followed by the *Visualization* results and a sensitivity study pertaining to choice of clustering parameters.

A. Feature Extraction

Frequency Transient Stability in 16 000-Bus System Case: A generator outage fault is applied at time 2 s which triggers a transient response in the system. Fig. 3(a) shows transient frequency signals of all 16 000 buses displayed in one plot over simulation time of 20 s. With the data stored every 0.05 s (i.e., every 12th value using a 1/4 cycle time step), the number of time points considered for a 20-s simulation is $\tau = 400$.

After applying the process of Feature Extraction, using $k = 600$ for K-means clustering and QT threshold diameter of 0.03 RMS distance, these signals are disentangled to identify eight distinct signals as shown in Fig. 3(b). Fig. 4(a)–(i) shows each of the eight clusters with the cluster representative median response in black and cyan lines showing range of signals in corresponding cluster. The cardinality is respectively 8754, 3983, 1920, 1281, 348, 43, 10, and 44 buses.

Greater the number of clusters formed, greater is the extent to which the patterns get sieved out. But, as can be seen, just eight clusters effectively captures all of the 16 000 frequency signals.

Voltage Transient Stability Analysis in 16 000-Bus System Case: The transient voltage signals of all 16 000 buses for the same fault as the previous case at each of 400 time points are analyzed here. These have been previously displayed in Fig. 1(a).

After applying the Feature Extraction process 60 distinct signals are identified shown in Fig. 5. It is well-known that voltage

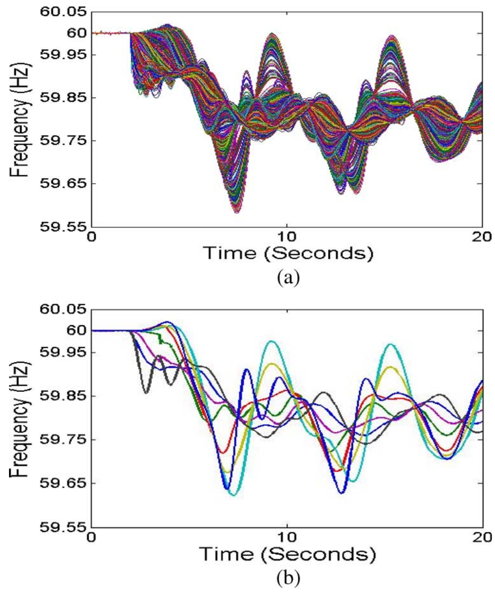


Fig. 3. (a) Plot of frequency signals at 16 000 buses. (b) Plot of 8 frequency signals identified as distinct patterns.

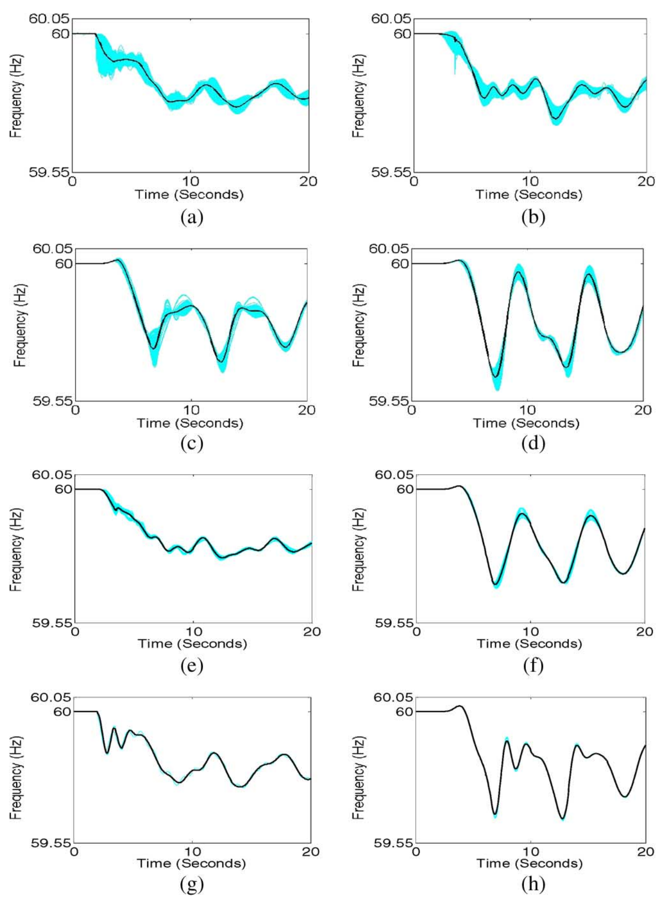


Fig. 4. Eight frequency clusters with respective cardinality. (a) Cluster 1 (8754 buses); (b) cluster 2 (3983 buses); (c) cluster 3 (1920 buses); (d) cluster 4 (1281 buses); (e) cluster 5 (348 buses); (f) cluster 6 (43 buses); (g) cluster 7 (10 buses); (h) cluster 8 (44 buses).

signals vary significantly between different buses which is reflected in more clusters obtained for voltage signals compared to frequencies. The clusters are obtained using 600 K-means clusters and QT threshold diameter 0.012 in RMS similarity measure.

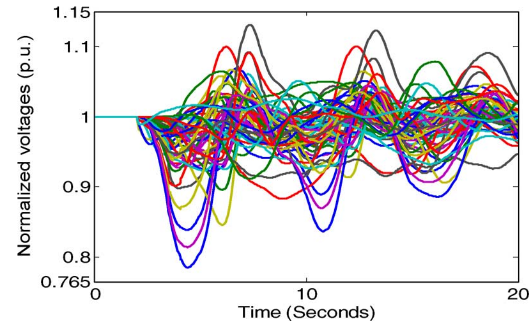


Fig. 5. Plot of 60 distinct voltage signals obtained by proposed analysis.

To limit the size of the paper, only ten of these 60 signals are shown in Fig. 6(a)–(j) in separate larger plots with each showing the range of voltages included in its cluster (colored in cyan). The number of buses included in each cluster is also mentioned. Note here that the voltages are normalized with respect to pre-fault values. Some of the "interesting" signals isolated as clusters with cardinality unity or very low can be seen.

It is interesting to note that the proposed method is able to isolate out unique and "unexpected" signals, which often reveal underlying issues with the system model and parameters etc. For example, analysis of the two signals in Fig. 6(d) determined that the generators connected to these two buses have positive eigen values in respective SMIB studies indicating model errors in them. Fig. 6(f) is also an outlier in that it is a generator with zero MW output directly connected to a 500-kV bus with compensation. This outlier detection ability proves the power of this method as an analysis tool.

Voltage Transient Stability Analysis in IEEE 118-Bus System Case: All 19 generators of the IEEE 118-bus system are modeled as GENCLS, i.e., classic generator model. In this model a synchronous generator is modeled as a Thevenin voltage source to play back known voltage/frequency signal. A 3-phase solid fault is simulated in the line between buses 23 and 25 at 1 s and cleared by opening the line at 1.12 s. This contingency triggers a disturbance in the system. The terminal p.u. voltages are recorded for a total simulation time of 5 s with time step of 0.001 s, i.e., for $\tau = 5000$ time points. The terminal p.u. voltages without any preprocessing are shown in Fig. 7. Patterns can be observed, but even using multi-color plots it is difficult to distinguish the signals. Also there seem to be groups of generators with common features in corresponding signals.

The terminal voltage signals are separated out and grouped by QT clustering. K-means clustering is not required here. Using RMS distance similarity measure and QT cluster threshold diameter of 0.03, six clusters are obtained as shown in Fig. 8(a)–(f). Cardinality of each cluster is also mentioned. Terminal voltages are preprocessed to be normalized with respect to pre-fault voltages for the purpose of clustering. In the figures however, the actual voltage values are depicted once the clusters are obtained. Running time of the algorithm is $\sim$0.03 s.

Clustering shows that there is one major group of generators based on voltage response with 13 generators (grey dots on one-line diagram in Fig. 9). The second cluster consists of two

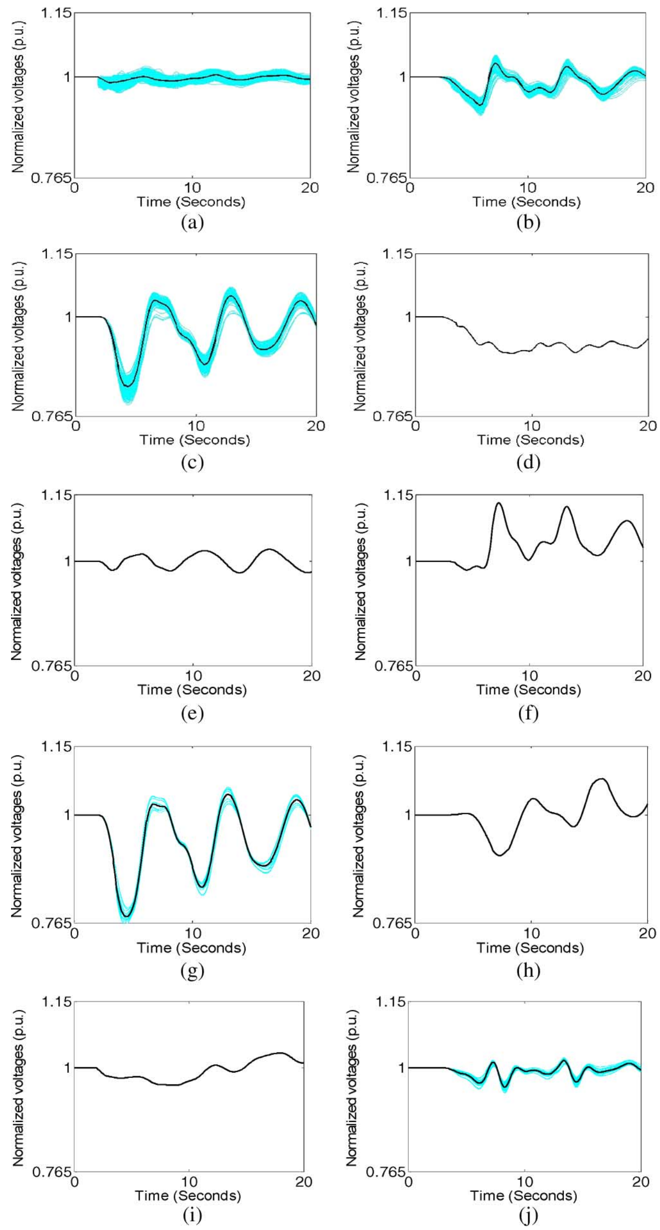


Fig. 6. Ten of sixty voltage clusters with respective cardinality. (a) Cluster 1 (5236 buses); (b) cluster 2 (408 buses); (c) cluster 3 (626 buses); (d) cluster 4 (2 buses); (e) cluster 5 (2 buses); (f) cluster 6 (1 bus); (g) cluster 7 (16 buses); (h) cluster 8 (1 bus); (i) cluster 9 (1 bus); (j) cluster 10 (125 buses).

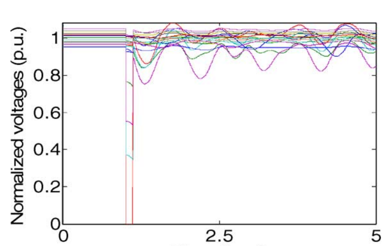


Fig. 7. Plot of terminal p.u. voltages of all 19 generators.

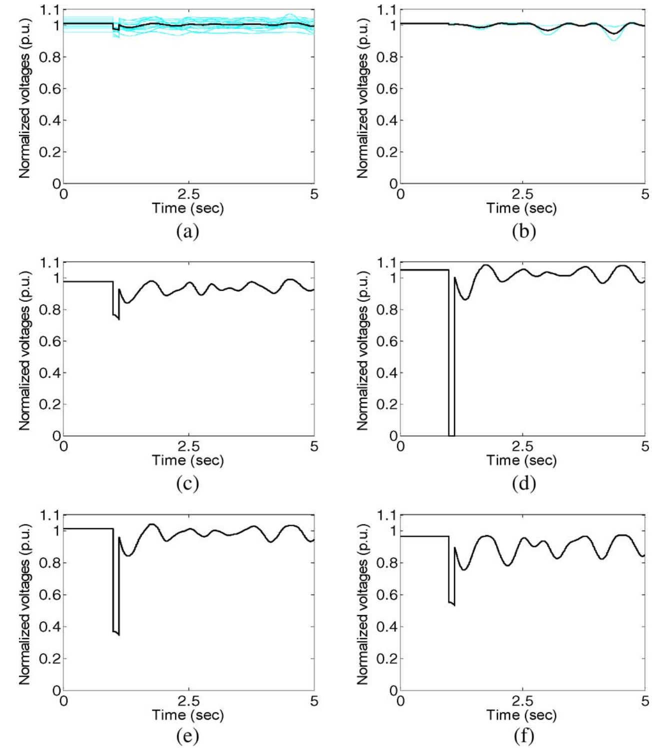


Fig. 8. Generator signals separated out (blue lines show individual signals, black line shows the average response in a cluster). (a) Cluster 1 (13 gens in grey dots); (b) cluster 2 (2 gens in blue dots); (c) cluster 3 (1 gen in brown dot); (d) cluster 4 (1 gen in yellow dot); (e) cluster 5 (1 gen in orange dot); (f) cluster 6 (1 gen in green dot).

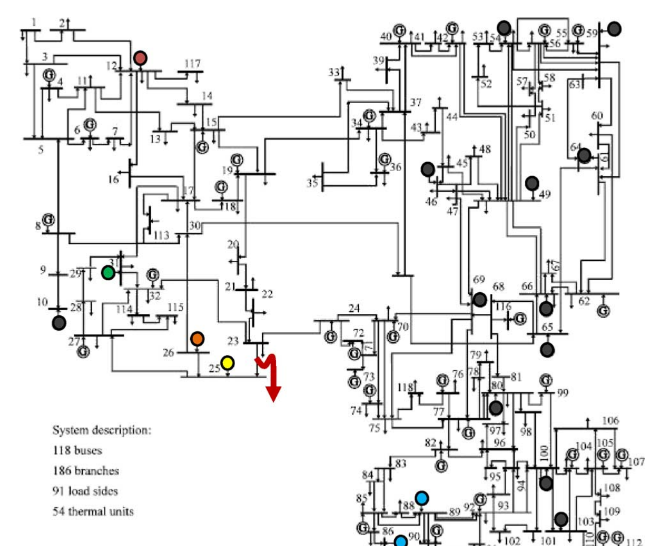


Fig. 9. IEEE 118-bus system with color-coded generators whose dynamic signals are recorded and faulted line (red arrow).

generators (blue dots in Fig. 9). The remaining four clusters each consist of a single generator as can be expected observing the group of signals in Fig. 7 and the isolated ones in Fig. 8(c)–(f). These are shown by brown, yellow, orange, and green dots in Fig. 9. The generator in Cluster 4 [signal shown in Fig. 8(d)] is connected to bus 25, which is one end of the line where fault occurs.

From the perspective of a power system operator, the IEEE 118-bus system is a very small system compared to real life large systems, but note how even in a system of this size there is the

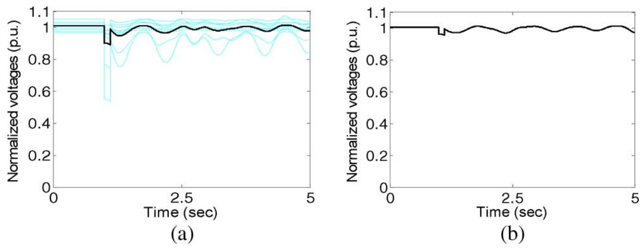


Fig. 10. Two out of six groups of generator signals when correlation is used for clustering (cyan lines are individual signals, black line is the average).

potential benefit of finding out more about the system by simply analyzing the data aggregated.

With a cluster threshold diameter of 0.16 and using correlation distance subtracted from unity as the similarity measure, six clusters are also formed. However, this measure fails to generate appropriate clusters as shown by two of these clusters in Fig. 10(a) and (b) which include nine generators and one generator, respectively.

Given the physical meaning of cosine similarity described earlier and noticing the voltage patterns, it is expected that the cosine distances between pairs of signals vary between 0.9 and 1 and there is not much scope of differentiating out groups of similar signals by orientation. Hence, the cosine distance measure fails to generate the proper clusters irrespective of different choices of cluster diameter.

B. Visualization

The concept of visualization of transient signals on geographic overlays with spark-lines is demonstrated on the real life 16 000-bus system.

In the interest of size limitations of this paper, only the transient stability visualization of frequency signals are shown. In Fig. 11, all 16 000 buses are displayed at respective latitude-longitude coordinates on a geographical map of the system with different color codes corresponding to different clusters. Thus there are eight colors visible corresponding to the eight clusters. The exact latitude and longitude values have been intentionally removed for confidentiality concerns. The X-axis corresponds to an East-West direction and the Y-axis to a North-South direction. The transient stability data corresponding to each cluster is displayed by spark-lines showing the trend and the current value. In addition, lines drawn from a spark-line plot to the corresponding bus indicates the cluster to which a spark-line belongs.

C. Sensitivity to Choice of Clustering Parameters

Both K-means and QT clustering algorithms need parameters to be specified. For K-means algorithm, the number of clusters k has to be specified. Since this step forms crude clusters very quickly to be fed into the more computationally complex QT algorithm, a sufficiently large number needs to be chosen for k. However, there is a tradeoff involved since too high k will slow down the QT algorithm significantly. The main parameter for the QT algorithm is d, i.e. the threshold distance. In Tables I and II, the variation of the number of clusters for different values of d and for different values of k is shown for clustering voltage and frequency signals respectively in the 16 000-bus system.

As can be seen, for any k, the number of clusters converges rapidly with increase in d. For very large d, all nodes converge

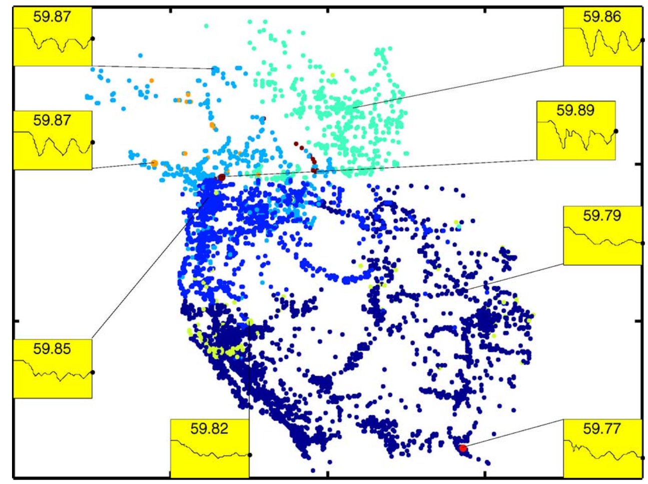


Fig. 11. Frequency visualization at 20 s.

TABLE I
SENSITIVITY OF VOLTAGE CLUSTERS TO CLUSTER PARAMETERS (16 000-BUS SYSTEM)

	k				*k*		
d	**200**	**400**	**600**	**d**	**200**	**400**	**600**
0	200	400	600	**0.015**	37	50	47
0.001	199	387	353	**0.016**	32	44	44
0.002	184	343	297	**0.017**	28	39	40
0.003	160	278	232	**0.018**	24	37	37
0.004	139	216	187	**0.019**	24	33	33
0.005	121	185	158	**0.02**	22	30	32
0.006	102	154	135	**0.03**	13	16	15
0.007	84	128	113	**0.04**	8	10	12
0.008	73	109	101	**0.05**	7	8	9
0.009	63	91	88	**0.06**	5	6	8
0.01	59	82	76	**0.07**	3	4	4
0.011	55	74	68	**0.08**	3	3	3
0.012	49	67	60	**0.09**	3	3	3
0.013	45	59	55	**0.1**	2	3	3
0.014	41	56	51				

TABLE II
SENSITIVITY OF FREQUENCY CLUSTERS TO CLUSTER PARAMETERS (16 000-BUS SYSTEM)

	k				*k*		
d	**200**	**400**	**600**	**d**	**200**	**400**	**600**
0	200	400	600	**0.015**	16	17	21
0.001	170	272	361	**0.016**	14	17	19
0.002	124	172	207	**0.017**	14	18	18
0.003	94	118	139	**0.018**	15	15	16
0.004	76	90	107	**0.019**	14	15	14
0.005	58	66	80	**0.02**	12	12	12
0.006	48	55	62	**0.03**	7	8	9
0.007	43	48	54	**0.04**	4	4	5
0.008	35	39	46	**0.05**	4	4	4
0.009	29	33	35	**0.06**	3	3	3
0.01	23	31	31	**0.07**	3	3	3
0.011	21	25	32	**0.08**	3	3	3
0.012	21	22	26	**0.09**	2	3	2
0.013	19	21	22	**0.1**	2	2	2
0.014	16	20	21				

to form 1 large cluster. For analyzing the frequency transient signals of the 16 000-bus system, the value of k is chosen as 600 and d as 0.035 which results in formation of eight clusters, where d is specified as RMS similarity measure. For voltage signals, k is chosen as 600 and d as 0.012 which results in formation of 60 clusters.

V. Conclusion

This work presents a clustering based method for processing power system transient stability data for important information and its visualization. The overall goal is to develop better understanding of the underlying system. Transient frequency and voltage signals from IEEE 118-bus and a real-world large 16 000-bus system have been considered for this work. Clustering algorithms identify portions of the system displaying similar transient signals and isolates the subset of distinct signals which characterizes the overall system response. Spark-lines are used for visualizing time-varying transient stability information on a geographic map of the system showing results at a glance. The most useful aspect of the method however, is in the ability to detect outliers in the transient signals which could potentially indicate model errors or other condition requiring further attention as evident from the case studies. The algorithms are extremely fast even when run on thousands of data points and hence can be used for real-time analysis in tracking mode with PMU measurements for example. These applications will be addressed in a future work.

References

[1] R. Xu and D. Wunsch, II, *Clustering, IEEE Press Series on Computational Intelligence*. Hoboken, NJ, USA: Wiley, 2009.

[2] R. Podmore, "Identification of coherent generators for dynamic equivalents," *IEEE Trans. Power App. Syst.*, vol. PAS-97, no. 4, pp. 1344–1354, Jul. 1978.

[3] J. R. Winkelman, J. H. Chow, B. C. Bowler, B. Avramovic, and P. V. Kokotovic, "An analysis of interarea dynamics of multi-machine systems," *IEEE Trans. Power App. Syst.*, vol. PAS-100, no. 2, pp. 754–763, Feb. 1981.

[4] B. Yang, V. Vittal, and G. T. Heydt, "Slow coherency based controlled islanding—A demonstration of the approach on the August 14, 2003 blackout scenario," *IEEE Trans. Power Syst.*, vol. 21, no. 4, pp. 1840–1847, Nov. 2006.

[5] L. Wang, M. Klein, S. Yirga, and P. Kundur, "Dynamic reduction of large power systems for stability studies," *IEEE Trans. Power Syst.*, vol. 12, no. 2, pp. 889–895, May 1997.

[6] S. Dutta and T. J. Overbye, "Information processing and visualization of power system time varying data," in *Proc. 2013 IEEE Symp. Series Computational Intelligence (SSCI)*, Singapore, Apr. 16–19, 2013.

[7] C. Ware, *Information Visualization: Perception for Design*, 2nd ed. Boston, MA, USA: Morgan Kaufmann, 2004.

[8] Y. Zhang, L. Chen, Y. Ye, P. Markham, J. Bank, J. Dong, Z. Yuan, Z. Lin, and Y. Liu, "Visualization of wide area measurement information from the FNET system," in *Proc. IEEE PES General Meeting*, Jul. 2011.

[9] Z. Zhong, C. Xu, B. Billian, L. Zhang, S.-J. S. Tsai, R. W. Conners, V. A. Centeno, A. G. Phadke, and Y. Liu, "Power system frequency monitoring network (FNET) implementation," *IEEE Trans. Power Syst.*, vol. 20, no. 4, pp. 1914–1921, Nov. 2005.

[10] *Technology Assessment of Power System Visualization*. Palo Alto, CA, USA: EPRI, 2009, 1017795.

[11] T. W. Liao, "Clustering of time series data—A survey," *Pattern Recognit.*, vol. 38, pp. 1857–1874, 2005.

[12] L. J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring expression data: Identification and analysis of coexpressed genes," *Genome Res.*, vol. 9, pp. 1106–1115, Nov. 1999.

[13] QT (Quality Threshold) Clustering, sig_support@agilent.com, 2005. [Online]. Available: http://www.chem.agilent.com/cag/bsp/products/gsgx/Downloads/pdf/qt_clustering.pdf.

[14] A. Danalis, C. McCurdy, and J. S. Vetter, "Efficient quality threshold clustering for parallel architectures," in *Proc. IEEE 26th Int. Parallel and Distributed Processing Symp.*, 2012.

[15] O. Dan and H. Mocian, "Scalable web mining with newistic," in *Proc. 13th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining*, 2009.

[16] P. J. Labuschagne, "Automatic clustering with application to time dependent fault detection in chemical processes," M.Eng. thesis, Univ. Pretoria, Pretoria, South Africa, 2008.

[17] G. Heydt and E. Gunther, "Post-measurement processing of electric power quality data," *IEEE Trans. Power Del.*, vol. 11, no. 4, pp. 1853–1859, Oct. 1996.

[18] E. Y. Hamid and Z.-I. Kawasaki, "Wavelet-based data compression of power system disturbances using the minimum description length criterion," *IEEE Trans. Power Del.*, vol. 17, no. 2, pp. 460–466, Apr. 2002.

[19] Y. Hu, C. Li, J. Li, Y.-H. Han, X.-W. Li, W. Wang, H.-W. Li, L.-T. Wang, and X.-Q. Wen, "Test data compression based on clustered random access scan," in *Proc. Asian Test Symp.*, 2006, pp. 231–236.

[20] V. Arya, J. Hazra, P. Kodeswaran, D. Seetharam, N. Banerjee, and S. Kalyanaraman, "CPS-Net: In-network aggregation for synchrophasor applications," in *Proc. 3rd Int. Conf. Communication Systems and Networks (COMSNETS)*, 2011.

[21] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. and Prob.*, 1967, vol. 1, pp. 281–297, Univ. Calif. Press.

[22] R. Klump, R. E. Wilson, and K. E. Martin, "Visualizing real-time security threats using hybrid SCADA./PMU measurement displays," in *Proc. 38th Hawaii Int. Conf. System Sciences*, Poipu, HI, USA, Jan. 2005.

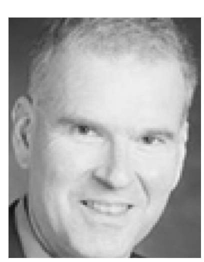
[23] E. Tufte, *Beautiful Evidence*. Cheshire, CT, USA: Graphics Press, 2006.

[24] T. Fruchterman and E. Reingold, "Graph drawing by force-directed placement," *Softw.—Practice & Exper.*, vol. 21, no. 11, pp. 1129–1164, Nov. 1991.

[25] PowerWorld Simulator. [Online]. Available: http://www.powerworld.com/.

Sudipta Dutta (S'08–M'12) received the B.E., M.Tech., and Ph.D. degrees in electrical engineering from Jadavpur University, India, Institute of Technology, Banaras Hindu University, India, and University of Illinois at Urbana-Champaign, respectively.

She is currently a Post-Doctoral Research Associate at University of Illinois at Urbana-Champaign. Her research interests include wind power forecasting and storage integration, wind farm collector system design, graph theoretic and data mining applications on power systems, and power system visualization.



Thomas J. Overbye (S'87–M'92–SM'96–F'05) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Wisconsin-Madison.

He is currently the Fox Family Professor of Electrical and Computer Engineering at the University of Illinois Urbana-Champaign. His research interests include power system visualization, power system analysis, and computer applications in power systems.