Hindawi Publishing Corporation
Journal of Control Science and Engineering
Volume 2012, Article ID 478373, 10 pages
doi:10.1155/2012/478373

Review Article

Progress in Root Cause and Fault Propagation Analysis of Large-Scale Industrial Processes

Fan Yang[1,2] and Deyun Xiao[1]

[1] *Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing 100084, China*
[2] *Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB, Canada T6G 2V4*

Correspondence should be addressed to Fan Yang, yangfan@tsinghua.edu.cn

Received 15 September 2011; Revised 17 January 2012; Accepted 1 February 2012

Academic Editor: Onur Toker

In large-scale industrial processes, a fault can easily propagate between process units due to the interconnections of material and information flows. Thus the problem of fault detection and isolation for these processes is more concerned about the root cause and fault propagation before applying quantitative methods in local models. Process topology and causality, as the key features of the process description, need to be captured from process knowledge and process data. The modelling methods from these two aspects are overviewed in this paper. From process knowledge, structural equation modelling, various causal graphs, rule-based models, and ontological models are summarized. From process data, cross-correlation analysis, Granger causality and its extensions, frequency domain methods, information-theoretical methods, and Bayesian nets are introduced. Based on these models, inference methods are discussed to find root causes and fault propagation paths under abnormal situations. Some future work is proposed in the end.

1. Introduction

In a large-scale industrial process, process units are connected; thus a fault can easily propagate from one unit to another along material or information flow paths. Therefore, the problem of fault detection and isolation cannot be limited in a local unit, but should be laid in a large scale, which leads to a set of new problems that have attracted many researchers.

First of all, the large-scale problem is featured by causality. Causality is a physical phenomenon based on cause-effect relationship between different variables [1]. When one focuses on the interconnections of the process units, the first step is to recognize the causality between variables and that is what an engineer is interested in because one should find the root cause and the fault propagation paths in a faulty mode [2, 3] before analyzing the accurate dynamics based on first-principle or mathematical models.

The main research topics are modelling methods from process knowledge and process data and inference methods based on the model. Initially, the signed directed graph (SDG) is established by representing the process variables as graph nodes and representing causal relationships as directed arcs [4, 5]. An arc from node A to node B implies that the deviation in A may cause a deviation in B. Positive or negative influence between nodes is assigned to the arc. This is a qualitative description of process knowledge. When a fault occurs, the fault propagates along consistent paths forming a set of nodes whose values are beyond the normal range. This set of variables with signs is called a symptom. Different symptoms reveal different fault types. In real-time supervision, symptoms are obtained by sensor readings. As soon as a symptom is triggered, operators should identify the possible cause(s) and take appropriate actions immediately for remedial action. The SDG model has its obvious disadvantages due to its qualitative features; thus we should explore other established methods, including quantitative models and take into account latest and effective formal techniques. In Section 2, the model description methods are summarized; process knowledge is also the main resource for modeling.

Another resource for modelling is process data because process knowledge is not always available. Even if it is available, a lot of insignificant information may easily

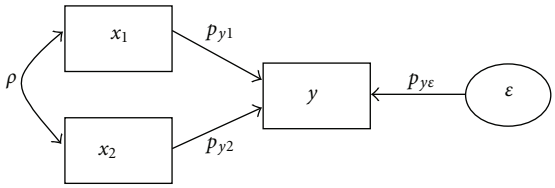

Figure 1: Path diagram of a structural model.

disturb the modelling procedure and make it too complex. Process data can effectively complement the information requirement and simplify the procedure; moreover, it can screen the nuisance information and improve the accuracy of the models. Here, the pairwise causality capture methods are developed to identify cause and effect. In a real process that is usually multivariate, a topology should be constructed based on pairwise analysis results. Several sets of method are introduced in Section 3.

Based on the models, diagnosis applications consist in finding the root cause whose abnormity accounts for all the abnormities detected in other parts [6]. Thus the purpose of the model-based inference is to interpret the symptom detected by finding the root cause and fault propagation paths. The most common algorithm for searching for root cause(s) is "depth-first traversal on the graph" [4, 7]. However, since there are various models, corresponding inference methods are needed. They are overviewed in Section 4.

2. Model Description Based on Process Knowledge

Based on a priori process knowledge, including first-principle and mathematical models, models can be built to describe process topology. Here, the term model has a broad meaning, not limited to equations.

2.1. Structural Equation Models. Structural equation modeling (SEM) is a statistical technique for testing and estimating causal relations [1, 8]. A structural model shows potential causal dependencies between endogenous/output and exogenous/input variables, and the measurement model shows relations between latent variables and their indicators. For example, if endogenous variable y is influenced by exogenous variables x_1 and x_2 (assume that all variables are normalized to be zero mean and unit variance), a regression model can be built as

$$y = p_{y1}x_1 + p_{y2}x_2 + p_{y\varepsilon}\varepsilon \tag{1}$$

and thus be depicted as a path diagram in Figure 1, where each parameter p is called a path coefficient, and ε represents the residual, that is, collective effect of all unmeasured variables that could influence y. The directed arrows represent the influence of the exogenous variables and the residual on the output variable, and the bidirectional arrow represents the correlation between exogenous variables.

Since the exogenous variables are not independent, there is some ambiguity about the real or dominant path. Based on the statistical analysis, components of direct and indirect relations can be evaluated via variance decomposition [2]; this gives some indication of the model structure. Typically, factor analysis, path analysis, and regression, as special cases of SEM, are widely used in exploratory factor analysis, such as psychometric design. IBM SPSS Amos (Analysis of Moment Structures) provides an easy-to-use program for visual SEM.

The limitations of this modeling approach are as follows (1) exogenous and endogenous variables should be selected in advance as a hypothesis and the result highly depends on this partition; (2) the causal relations are static relations; (3) only linear regression is considered. To overcome the last two limitations, dynamic causal modeling embraces nonlinear and dynamic nature [9]. In total, this approach is more suitable for confirmatory modeling than exploratory modeling to construct a network topology and suffers from large number of variables.

In recent years, some novel models have been developed such as undirected or directed graphs or networks [10], data models in databases [11], and production rules in expert systems [12]. Following is introduction to some typical cases and applications of these models.

2.2. Causal Graphs. We have seen that a graphical model provides an intuitive way to show causality. There are quite a few causal graphs that are dedicated to this description.

2.2.1. Signed Directed Graphs. Signed directed graphs (SDGs) are established by representing the process variables as graph nodes and representing causal relations as directed arcs. An arc from node A to node B implies that the deviation of A may cause the deviation of B. For convenience, "+", "−", or "0" is assigned to the nodes in comparison with normal operating value thresholds to denote higher than, lower than, or within the normal region, respectively. Positive or negative influence between nodes is distinguished by the sign "+" (promotion) or "−" (suppression), assigned to the arc [4, 5, 13, 14].

Take a bitank system as an example, as shown in Figure 2. Two tanks are connected by a pipe; both tanks have outlet pipes, and Tank 1 has a feed flow. This system can be described by the following set of differential and algebraic equations:

$$\begin{aligned} C_1\frac{de_2}{dt} &= f_1 - f_3 - f_5, \\ C_2\frac{de_7}{dt} &= f_5 - f_8, \\ f_3 &= \frac{1}{R_{b1}}\sqrt{l_2}, \\ f_5 &= \frac{1}{R_{12}}\left(\sqrt{l_2} - \sqrt{l_7}\right), \\ f_8 &= \frac{1}{R_{b2}}\sqrt{l_7}, \end{aligned} \tag{2}$$

where l_2 and l_7 are the levels in Tanks 1 and 2, f_1, f_3, f_5, and f_8 are flowrates, and R_{12}, R_{b1}, and R_{b2} are the resistances of the pipes between Tanks 1 and 2 and the two outlet pipes, respectively. Since l_i ($i = 2$ or 7) appears as the square root

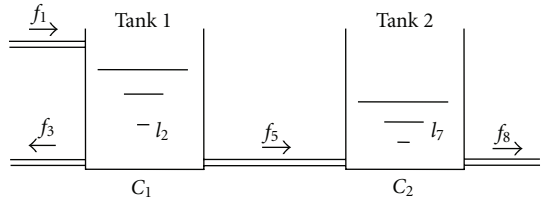

Figure 2: Schematic of a bitank system. C_1 and C_2 are cross-sectional areas of the two tanks, respectively.

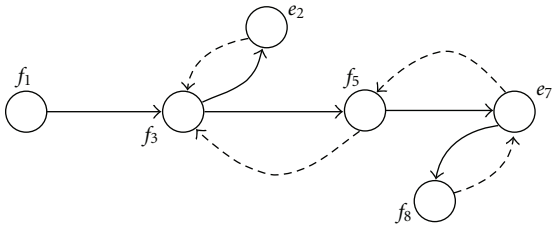

Figure 3: SDG of the bitank system. e_2 and e_7 are the square roots of levels in the two tanks respectively.

form, we use e_i to denote it. One can convert these equations to nodes and arcs to form an SDG, as shown in Figure 3, where solid lines denote positive influences and broken lines denote negative influences. Although no control is taken, there are still some recycles based on the principles.

An SDG can be built manually from first principles and mathematical models and more practically from process knowledge including flowsheets [15, 16].

2.2.2. Other Causal Graphs. Graphical models are commonly used to describe large systems, and yet they have different forms with different meanings. Bond graphs [17] and their extension, temporal causal graphs [18], use different symbols to further describe dynamic characteristics. More precisely, qualitative transfer functions [19], differential equations [20], and trend analysis [21, 22] have been integrated into causal graphs, and complex algorithms are introduced to improve their correctness [23]. Similar or improved approaches were investigated by many researchers [24–27].

The bond graph of the bitank system is shown as Figure 4, and the temporal causal graph is shown as Figure 5. In the bond graph, there are two types of junctions—common effort (0−) junction and common flow (1−) junction. It is obvious that the bond graph describes the exchange of physical energy by bonds. A bond graph can be used to derive the steady-state model automatically; this property is similar with signal flow graphs, which, as another graphical model, can be used for derivation of transfer functions. The temporal causal graph converts the junctions and bonds in the bond graph into nodes and arcs and imposes labels on arcs to describe detailed temporal effects such as integration and rate of change.

Compared to the SDG in Figure 3, the temporal graph provides more detailed information and forms a quantitative

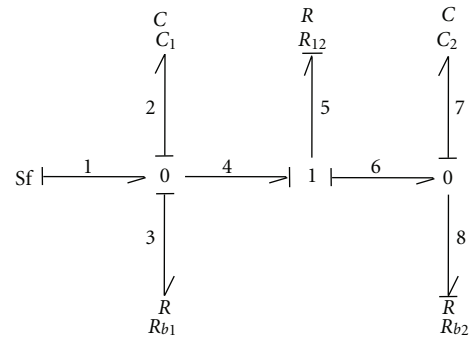

Figure 4: Bond graph of a bitank system.

model, while the SDG is only concerned with the qualitative trends. Since the exact model is often difficult to obtain for industrial processes, the SDG model is more widely used for its simplicity because it can be validated by process data [28].

2.3. Rule-Based Models. Kramer and Palowitch [29] used rules to describe SDG arcs and thus expert systems can be employed as a tool in this problem. Each arc can be described by a rule using logical functions p, m, and z

$$\begin{aligned} (pAB) &\Longleftrightarrow A \longrightarrow B \text{ (positive relation)} \\ (mAB) &\Longleftrightarrow A \dashrightarrow B \text{ (negative relation)} \\ (zAB) &\Longleftrightarrow A \quad B \text{ (zero relation).} \end{aligned} \tag{3}$$

Therefore, an SDG can be converted into a set of rules. These rules can be expressed as IF-THEN forms to make reasoning by rule reduction. Since only qualitative information is included, there may exist a lot of illusive results. To prevent this disadvantage, some quantitative information, such as steady-state gain, is taken into account to find dominant propagation paths [30].

2.4. Ontological Models. In order to standardize the conversion procedure from process knowledge to ontology, the semantic web has been developed, the architecture of which includes a series of languages produced by World Wide Web Consortium (W3C, http://www.w3.org/), for example, XML, RDF, RDFS, and OWL. Extensive markup language (XML) is the basic and widely accepted open standard for the representation of arbitrary data structure in a text document, especially web services. XML gives the user sufficient freedom to further define and apply in their respective areas, but for the purpose of semantic description, we need a more uniform way to define the process units (considered as resources). Resource Description Framework (RDF) provides a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats (http://www.w3.org/RDF/). To structure these RDF resources, RDFS (RDF Schema) is used. It provides an XML vocabulary to express classes with relationships (taxonomies) and define properties associated with classes, which facilitates the inferencing on the data [31]. RDFS is an ontological

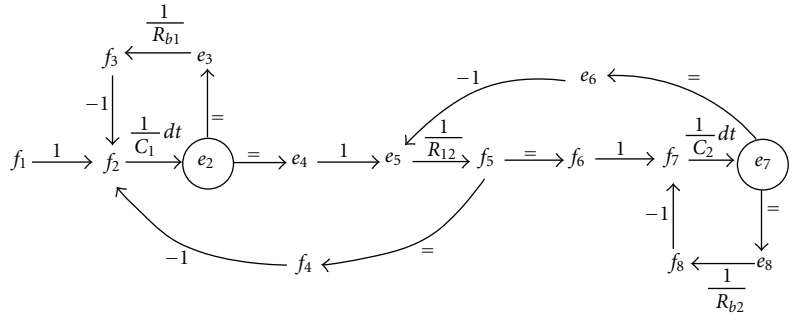

Figure 5: Temporal causal graph of a bitank system.

primitive, upon which Web Ontology Language (OWL) released in 2004 adds extensive features and becomes a more expressive language. An ontology is stored and referred to a unique name space to be retrieved easily. As an improvement over XML, RDF/OWL describes the semantics that is interchangeable between different programs and is convenient for inference. It is the trend of representation of process knowledge in the future. Matrikon's new software platform, Intuition, is built on RDF/OWL standards and incorporates the semantics to enable all people, processes, and applications to work in concert. Several software tools are available for editing RDF/OWL files, such as TopBraid Composer and Protégé-OWL.

In RDF/OWL standards, a data model is described by a collection of triples of subject, predicate, and object expressed as XML syntax, where the subject denotes the resource, the predicate denotes a property of this resource (can be multiple), and the object denotes the value of this property (should be unique, and can be literal or another resource). By this way, not only the inclusive relationship between resources is defined by the taxonomy of classes and subclasses, but also the directed logic relationship or linkage between instances is described by properties.

Apart from datatype and annotation properties, we define the following object properties to describe the physical and information linkage;

(i) UncontrolledElement.measuringElement: linkage from an uncontrolled element to a measuring element, for example, the level of a tank measured by a sensor.

(ii) UncontrolledElementOutlet.uncontrolledElementInlet: linkage from an uncontrolled element to another uncontrolled element, for example, a tank connected to a pipe as an outlet.

(iii) UncontrolledElementOutlet.controllingElementInlet: linkage from an uncontrolled element to a controlling element, for example, a pipe connected to a control valve.

(iv) ControllingElementOutlet.uncontrolledElementInlet: linkage from a controlling element to an uncontrolled element, for example, a valve connected to a pipe.

(v) Computer.computer: linkage from a computer to another computer, for example, a controller connected to a signal line (information connecting element).

The domain and range of the properties should be defined as appropriate resources.

3. Topology Capturing from Process Data

The cause-effect relationship can be explained from several different viewpoints. First, the propagation needs time, so the cause precedes the effect; this property can be tested by cross-correlation with an assumed lag or fitting the input-output data into dynamic models. Second, cause-effect relationship means information transfer; thus the measure of transfer entropy in information theory can also be employed. Third, causal relationship shows probabilistic properties; thus Bayesian nets are introduced to describe these relationships.

3.1. Cross-Correlation Analysis. Assume that x and y are normalized time series of n observations, then the cross-correlation function (CCF) with an assumed lag k is [32]

$$\phi_{xy}(k) = E[x_i y_{i+k}], \quad k = -n+1, \ldots, n-1. \tag{4}$$

A value of the CCF is obtained by assuming a certain time delay for one of the time series. Thus the absolute maximum value can be regarded as the real cross-correlation and the corresponding lag as the estimated time delay between these two variables. For mathematical description, one can compute the maximum and minimum values $\phi^{\max} = \max_k\{\phi_{xy}(k), 0\} \geq 0$ and $\phi^{\min} = \min_k\{\phi_{xy}(k), 0\} \leq 0$, and the corresponding arguments $k^{\max}$ and $k^{\min}$. Then the time delay from x to y is

$$\lambda = \begin{cases} k^{\max}, & \phi^{\max} \geq -\phi^{\min} \\ k^{\min}, & \phi^{\max} < -\phi^{\min} \end{cases} \tag{5}$$

(corresponding to the maximum absolute value) and the actual time delayed cross-correlation is $\rho = \phi_{xy}(\lambda)$ (between -1 and 1). If λ is less than zero, then it means that the actual delay is from y to x. Thus the sign of λ provides the directionality information between x and y. The sign of ρ corresponds to the sign of the arc in the signed directed graph meaning whether the correlation is positive or negative; this sign provides more information than the causality.

Although this method is practical and easy for computation, it has many shortages, some of which are explained below.

(i) Nonlinear causal relationship does not necessarily show up in correlation analysis. For example, if y equals the square of x with the time delay of one sampling time, then based on the time-delayed cross-correlation, this obvious causality cannot be found because all the values are small relative to a threshold. This can be explained because the true correlation should be zero.

(ii) Correlation simply gives us an estimate of the time delay. And the sign of the delay is an estimate of the directionality of the signal flow path. The time delay obtained, however, is only an estimate. In addition, the trend in a time series is ignored, and values at different time instances are regarded as samples of the same random event. Thus the causality obtained by this measure is purely the time delay based on the estimate of the covariance.

3.2. Granger Causality and Its Extensions. Regression is a natural way to test the relationship between variables. By taking into account dynamics, the lags in the models reflect the causality. A regression of a variable on lagged values of itself is compared with the regression augmented with lagged values of the other variable. If the augmentation is helpful for better regression, then one can conclude that this variable is Granger-caused by the other variable. Some tests are used, such as the t-test and the F-test.

Aiming at time series y and x, to test if there is a Granger causality from x to y, a univariate autoregression of y is obtained first:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_m y_{t-m} + \text{residual}_t. \quad (6)$$

Next, lagged values of x are included to obtain another regression:

$$\begin{aligned} y_t = {} & a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots a_m y_{t-m} \\ & + b_p x_{t-p} + \cdots + b_q x_{t-q} + \text{residual}_t. \end{aligned} \quad (7)$$

If the result is significantly better than the previous one, then a Granger causality is detected.

The multivariate version is available for this method based on the vector regression model and thus the conditioning is performed to exclude the influence of the intermediate variables.

This method needs a regression model; thus the following disadvantages are obvious. First, a linear relation between x and y is assumed, which is very strict. Second, the model accuracy affects the result, especially the predefined model order. There are some extensions of the basic Granger causality concept, such as variants of the Wiener-Granger causality [33], to describe more general forms.

3.3. Frequency Domain Methods. A process can also be described in the frequency domain where the energy transfer at every frequency can be shown. Based on this idea, several methods have been developed, such as the directed transfer function (DTF) [34] and the partial directed coherence (PDC) [35]. These quantities DTF and PDC are normalized measures of the total and direct influence, respectively, between two variables in a multivariate process. Conditioning is conducted to exclude the influence of the confounding variables [36]; this is very important under the multivariate framework [37].

Gigi and Tangirala [36] did quantitative analysis on the strength and proved that the total effect, in fact, consists of three components, namely, direct, indirect, and an interference term. The total effect can be quantified by the DTF, whilst the direct effect is hard to quantify. Anyway, the analysis can be performed with the visualization of a curve matrix.

The frequency domain methods have the similar advantages as the corresponding time domain methods (Granger causality methods). However, they provide a better vision for the energy transfer description at different frequencies.

3.4. Information-Theoretical Methods: Transfer Entropy. According to information theory, the transfer entropy from x to y is defined as [38]

$$t(y \mid x) = \sum_{y_{i+h}, \mathbf{y}_i, \mathbf{x}_j} p\left(y_{i+h}, \mathbf{y}_i, \mathbf{x}_j\right) \cdot \log \frac{p\left(y_{i+h} \mid \mathbf{y}_i, \mathbf{x}_j\right)}{p\left(y_{i+h} \mid \mathbf{y}_i\right)}, \quad (8)$$

where p means the complete or conditional probability density function (PDF), $\mathbf{x}_j = [x_j, x_{j-\tau}, \ldots, x_{j-(k-1)\tau}]$, $\mathbf{y}_i = [y_i, y_{i-\tau}, \ldots, y_{i-(l-1)\tau}]$, τ is the sampling period, and h is the prediction horizon. The transfer entropy is a measure of information transfer from x to y by measuring the reduction of uncertainty while assuming predictability. It is defined as the difference between the information about a future observation of x obtained from the simultaneous observation of past values of both x and y, and the information about the future of x obtained from the past values of x alone. It gives a good sense of the causality information without having to require the delay information. Several parameters, especially τ and h, should be tried. If the transfer entropies in two directions are considered, then $t(x \rightarrow y) = t(y \mid x) - t(x \mid y)$ is used as a measure to decide the quantity and direction of information transfer, that is, causality. In (2), the PDF can be estimated by the kernel method [28, 39], to fit any shape of the distributions.

Transfer entropy is a model-free method. However, it has the following main shortcomings. First, it is highly dependent on the estimation of PDFs (although it can have any non-Gaussian forms); thus the computational burden is very high. Second, the time delay cannot be estimated, and the arc signs in SDGs cannot be obtained. Third, the assumption that the time series is stationary does not hold and thus the noise (may be nonstationary) is often greater than expected; these problems affect the computational results.

3.5. Bayesian Nets. Random phenomenon is everywhere in the real world, including industrial processes. Due to the

existence of random noises, there are stochastic factors that can be described. The Bayesian net [40] provides a graph with probabilities, where nodes denote fault modes as well as process variables, and arcs denote conditional probabilities. Although the structure remains the same as an ordinary causal graph, both nodes and arcs mean probabilities. The causality from x to y is described by a conditional probability $p(y \mid x)$ [41].

This model is also a general model, although the meaning is different from the previous one. It is to be noted that, in industrial processes, dynamics, or time factors, should be included, which is a key feature to capture causality. The traditional Bayesian net has a fatal limitation that it should be a directed acyclic graph. In a logical system with no time factor, this assumption makes sense, but in a dynamic process, cycles are very common. A cyclic causal discovery algorithm has been developed [42] to allow the existence of cycles.

The major limitations of the application of Bayesian nets are as follows the physical explanation of probabilities is not straightforward, which is sometimes unacceptable by engineers; and the data requirement is hard to meet because one needs the data in all modes to build the model.

3.6. Other Methods and Comments. In addition to the above methods, there are more alternative methods to capture causality between time series. For example, predictability improvement [43, 44] is another general method but without the shortcoming of requiring a large data set. It computes the reduction of uncertainty of one variable with the help of the other variable. Smith et al. and Lungarella et al. have summarized and compared many methods to capture causality for bivariate series [45] and in a network [46], respectively. Each of all these methods has its own advantages and limitations; they complement each other and no one method is powerful enough to replace the others. Hence we should try different methods to obtain reasonable results. In real applications, one may mainly choose one method but sometimes use other methods to gain additional insights or make validations.

Most of the above data-based methods (except model-based methods in Sections 3.2 and 3.3) cannot capture the true causality because they are pairwise methods. If both x and y are driven by a common third variable, sometimes with different lags, one might still find some causality. In fact, there is no causality between these two variables and neither of them can have influence on the other if the third variable does not change. Thus one needs to test all the pairs of variables to obtain their causality measures and then construct the structure. The structure should be a mixture of the typical serial structure and parallel structure. Indeed, the topology determination needs additional information beyond pairwise tests.

4. Model-Based Inference

Based on the models, inference should be made to find the fault propagation paths and thus the root cause. Following are some typical approaches.

4.1. Graph Traversal. The most common algorithm for searching the fault origin is depth-first traversal on the graph [4, 5], which is a kind of efficient fault inference for both the single and multiple fault origin cases [47]. Its theoretical basis is nodal balance [48]. A depth-first traversal algorithm constructs a path by moving each time to an adjacent node until no further arcs can be found that have not yet been visited, the implementation of which is a recursive procedure.

For the purpose of fault propagation analysis, forward traversal is applied from the assumed origin to predict all the variables based on consistency, which is deductive reasoning [49, 50]. For the fault detection purpose, backward traversal is applied within the causal-effect graph to find the maximal strongly connected component [4], which is abductive reasoning. Actually, the whole procedure includes two steps.

Step 1. Trace the possible fault origins back along the arcs.

Step 2. Make forward inference from these nodes to screen the candidates to choose which one is the real or most probable fault origin.

Loops should be treated specially, which is very common in control systems because of control loops [51].

The time complexity of a traversal search is $O(n^2)$ in which n denotes the node number in the graph. When the system scale increases, the time for a traversal is too long to meet the demands of fault detection. Thus the model structure should be transformed from a single-level one to a hierarchical one [52–54]. By this way, the search is first performed in the higher level to restrict the fault origin in a subsystem. Then the search is performed in the subgraph of this subsystem.

For the hierarchical model, hierarchical inference from top to bottom is obtained naturally. The graph traversal is performed firstly in the higher level finding the possible super-node that includes the fault origin. Next perform the graph traversal in the lower level to restrict the possible location of the root cause. Assume the subsystem contains m control systems, and each control system contains k variables, then the time complexity of a traversal in a single-level model is $O(m^2k^2)$, and the time complexity in a 2-level model is $O(m^2 + k^2) \ll O(m^2k^2)$. Thus the fault analysis in a hierarchical model has much higher efficiency.

Take the case of a boiler system in a power plant [55] for example. There are about 40 key variables that are measured or manipulated, and several control loops are maintaining the steady operation. Of course, this process can be simulated by a large set of equations according to sufficient process knowledge. For fault analysis in an abnormal situation, however, we are more concerned about localizing the root cause before estimating the accurate value. For instance, if the coal quality is changed suddenly, many variables in different subsystems and control systems appear abnormal. If we look at the single-level model, it is not easy to focus on the faulty part. However, if we have a 2-level model, in which the high level describes the relationship between process units

and the low level describes the detailed relationship between process variables, the traversal at the high level can help us find that the root cause is located within the overheated steam pressure control system. Then by digging into this system, we can more easily find the real problem is the change of the coal quality. In this case, the search efficiency is highly improved.

Here the number of fault origins is assumed to be only one, that is, the reason that leads to the fault is only one [4]. This is reasonable because multiple faults seldom appear at the same time [56]. For multiple fault origin cases, minimal cut sets diagnosis algorithm was presented [57], where all possible combinations of overall bottom events should be input into the computer to explore and those which make the top events appear are the cut sets. This algorithm has the distinct disadvantage of low efficiency because of exponential explosion.

In order to utilize the system information more sufficiently, Han et al. [58] used fuzzy set to improve the existing models and methods, but their method is not so convenient for online inference and is not applicable for dynamical systems. Some scholars introduced temporal evolution information such as transfer delay [59, 60] and other kind of information into SDG for dynamic description.

4.2. Inference Based on Expert Systems. Rule-based inference [29] is applicable when an expert system is available. This method can be used to improve the inference accuracy with the appropriate rule description and operation. Rough set theory provides an idea of handling vague information and can be used to data reduction; thus it can be introduced to the fault isolation problem (a kind of decision problems) to optimize the decision rules. The decision algorithm is proposed by Yang and Xiao [61], in which the generation and reduction methods of the rules are related to the structure of the SDG model.

The main steps are listed as follow.

(1) List all the possible rules as *Table A* (as shown in Table 1), with each row denoting a rule $\varphi \rightarrow \psi$, where φ denotes the values of the condition attributes are assumed and ψ denotes the decision to be obtained. For convenience, we can give each attribute value a notion.

(2) Try to delete each condition attribute in turn and test the consistency of the formula and obtain the reducts and the core. Delete all the elements except the cores and get *Table B*. There are several methods to test the consistency. For example;

   (a) Each condition class $E \in X|IND(\mathbf{C})$ has the same decision value.

   (b) For each object x, the condition class covering x is contained in the decision class covering x.

   (c) For every two decision rules $\varphi \rightarrow \psi$ and $\varphi' \rightarrow \psi'$, we have $\varphi = \varphi' \rightarrow \psi = \psi'$.

(3) Calculate the reducts of each rule by use of *Table B* and obtain *Table C*.

Table 1: Framework of a decision table.

	Attribute **Q**	
Objects **X**	Condition attributes **C**	Decision attributes **D**

(4) Delete redundant rules and thus obtain *Table D*.

(5) Deduce the rules and the decision algorithm according to *Table D*.

The authors combine the algebraic and logical expression ways to achieve the purpose. Moreover, due to the convenience of expressing granularity, the decision algorithm is still applicable when the types of the faults of concern are changed or reformed.

4.3. Inference Based on Bayesian Nets. In Bayesian nets, probability and conditional probability of fault events is used to describe causes and effects among variables. Hence the inference is in respect to the fault probability.

We can use Bayesian inference on the graph to calculate the probabilities; it is a direct method. Suppose that the node set of the probabilistic SDG is $V = E \cup F \cup H$, in which E is the subset of evidence nodes whose value or probabilities are known, F is the subset of query nodes whose probabilities are to be computed, and H is the subset of hidden nodes which is not cared about in the inference. The inference process is to compute the conditional probability of x_F given the known x_E

$$p(x_F \mid x_E) = \frac{p(x_E, x_F)}{p(x_E)}, \tag{9}$$

where

$$\begin{aligned} p(x_E, x_F) &= \sum_{x_H} p(x_E, x_F, x_H), \\ p(x_E) &= \sum_{x_F} p(x_E, x_F). \end{aligned} \tag{10}$$

To solve this problem, Bayesian formula and its chain rule should be used adequately, and also the junction tree algorithm can be used for multiple-fault origin cases. This method could be used where there is distinct random phenomenon, yet the cycles in SDGs should be handled. The algorithm is the combination of depth-first search and junction tree algorithm.

4.4. Query on Ontological Models. Similar to the query language SQL used in relational databases, query languages, SPARQL, RDQL, Versa, and so forth are used in ontology-based RDF/OWL files to capture useful information and conduct inference. Among them, SPARQL (SPARQL Protocol and RDF Query Language) is the predominant one and has been recommended by the W3C in 2008 (http://www.w3.org/TR/rdf-sparql-query/).

SPARQL uses query triples as expressions with logic operations such as conjunctions and disjunctions. It can perform inferences based on semantics.

The functions of SPARQL query can be summarized as follows.

(i) To perform query based on specific property constraints. For example, we can search all the outlet pipes of a tank by defining the subject and the predicate and constraining class of the resulting objects.

(ii) To test connectivity based on object properties. If we define a general object property and place all the other object properties meaning physical and information linkages under it, then the connectivity with specified steps can be obtained. In the matrix form, reachability matrix is defined as $\mathbf{R} = (\mathbf{X} + \mathbf{X}^2 + \cdots + \mathbf{X}^N)^{\#}$ where $\mathbf{X}$ is the adjacency (connectivity) matrix [62], N is the number of elements, and # is the Boolean operator [10]. But if we want to know the k-step propagation results from one element, then we should truncate the summation to the first k terms, from which each element can be obtained by a query. This truncated reachability reflects the precedence and strength of the propagation.

(iii) By defining the object property as transitive, reachability can be obtained directly to show the domain of influence triggered by a change in one object.

5. Conclusion and Future Directions

In this paper, various methods for the purpose of root cause and fault propagation analysis are introduced briefly, and the features and limitations are analyzed. For the fault detection and isolation of a large-scale industrial process, it is the first step to limit the scale of the problem by capturing the backbone and find the real problem before diagnosing the problem precisely.

We notice that there is no single method that can perfectly achieve our purpose. Therefore, a fusion of different methods is necessary. In real applications, one method, simple ones in most cases, can be used first; and then another method can be used as a validation or a comparison. To facilitate this procedure, a tool is under development by integrating various methods. The suggestions will also be given to the user when choosing appropriate methods.

There are also some theoretical problems that need attention. Instantaneous causality and bidirectional causality are possible in real cases; we need particular methods to deal with them. Most of the methods need some user-defined parameters; the choices of them should be studied to compromise between accuracy and computational complexity. Topology construction is still an open question; we should go beyond the pairwise analysis to study the multivariate analysis methods. Single-layer model is ineffective for large-scale systems; thus hierarchical models should be developed and the established various models should be extended. Under different abnormal situations, the model structure may be changed and thus the anomaly detection and model switch mechanism should be studied [63]. For the simulation study, the Tennessee Eastman process can be used as a benchmark [64].

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant nos. 60736026 and 60904044) and Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-Discipline Foundation. The authors would also acknowledge the guidance of Professors Sirish L. Shah and Tongwen Chen at the University of Alberta.

References

[1] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, UK, 2000.

[2] S. Y. Yim, H. G. Ananthakumar, L. Benabbas, A. Horch, R. Drath, and N. F. Thornhill, "Using process topology in plant-wide control loop performance assessment," *Computers and Chemical Engineering*, vol. 31, no. 2, pp. 86–99, 2006.

[3] A. Rahman and M. A. A. Shoukat Choudhury, "Detection of control loop interactions and prioritization of control loop maintenance," *Control Engineering Practice*, vol. 19, no. 7, pp. 723–731, 2011.

[4] M. Iri, K. Aoki, E. O'Shima, and H. Matsuyama, "An algorithm for diagnosis of system failures in the chemical process," *Computers and Chemical Engineering*, vol. 3, no. 1–4, pp. 489–493, 1979.

[5] M. Iri, K. Aoki, E. O'shima, and H. Matsuyama, "A graphical approach to the problem of locating the origin of the system failure," *Journal of Operations Research Society of Japan*, vol. 23, no. 4, pp. 295–311, 1980.

[6] N. F. Thornhill and A. Horch, "Advances and new directions in plant-wide disturbance detection and diagnosis," *Control Engineering Practice*, vol. 15, no. 10, pp. 1196–1206, 2007.

[7] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis part II: qualitative models and search strategies," *Computers and Chemical Engineering*, vol. 27, no. 3, pp. 313–326, 2003.

[8] S. Wright, "Correlation and causation," *Journal of Agricultural Research*, vol. 20, pp. 557–585, 1921.

[9] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling," *NeuroImage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[10] R. S. H. Mah, *Chemical Process Structures and Information Flows*, Butterworth Publishers, Boston, Mass, USA, 1990.

[11] G. Simson, *Data Modeling: Theory and Practice*, Technics Publications LLC, Denville, NJ, USA, 2007.

[12] OMG, *Business Semantics of Business Rules RFP*. br/03-06-03, 2003.

[13] F. Yang and D. Y. Xiao, "Review of SDG modeling and its application," *Control Theory and Applications*, vol. 22, no. 5, pp. 767–774, 2005.

[14] F. Yang, D. Xiao, and S. L. Shah, "Qualitative fault detection and hazard analysis based on signed directed graphs for large-scale complex systems," in *Fault Detection*, W. Zhang, Ed., pp. 15–50, IN-TECH, Vukovar, Croatia, 2010.

[15] M. R. Maurya, R. Rengaswamy, and V. Venkatasubramanian, "A systematic framework for the development and analysis of signed digraphs for chemical processes. 1. Algorithms and analysis," *Industrial and Engineering Chemistry Research*, vol. 42, no. 20, pp. 4789–4810, 2003.

[16] M. R. Maurya, R. Rengaswamy, and V. Venkatasubramanian, "A systematic framework for the development and analysis of signed digraphs for chemical processes. 2. Control loops and flowsheet analysis," *Industrial and Engineering Chemistry Research*, vol. 42, no. 20, pp. 4811–4827, 2003.

[17] H. M. Paynter, *Analysis and Design of Engineering Systems*, MIT Press, Cambridge, Mass, USA, 1960.

[18] P. J. Mosterman and G. Biswas, "Diagnosis of continuous valued systems in transient operating regions," *IEEE Transactions on Systems, Man, and Cybernetics Part A*, vol. 29, no. 6, pp. 554–565, 1999.

[19] L. Leyval, S. Gentil, and S. Feray-Beamont, "Model-based causal reasoning for process supervision," *Automatica*, vol. 30, no. 8, pp. 1295–1306, 1994.

[20] J. Montmain and S. Gentil, "Dynamic causal model diagnostic reasoning for online technical process supervision," *Automatica*, vol. 36, no. 8, pp. 1137–1152, 2000.

[21] M. R. Maurya, R. Rengaswamy, and V. Venkatasubramanian, "A signed directed graph and qualitative trend analysis-based framework for incipient fault diagnosis," *Chemical Engineering Research and Design*, vol. 85, no. 10, pp. 1407–1422, 2007.

[22] D. Gao, C. Wu, B. Zhang, and X. Ma, "Signed directed graph and qualitative trend analysis based fault diagnosis in chemical industry," *Chinese Journal of Chemical Engineering*, vol. 18, no. 2, pp. 265–276, 2010.

[23] H. Cheng, V.-M. Tikkala, A. Zakharov, T. Myller, and S. L. Jamsa-Jounela, "Application of the enhanced dynamic causal digraph method on a three-layer board machine," *IEEE Transactions on Control Systems Technology*, vol. 19, no. 3, pp. 644–655, 2011.

[24] C. J. Alonso, C. Llamas, J. A. Maestro, and B. Pulido, "Diagnosis of dynamic systems: a knowledge model that allows tracking the system during the diagnosis process," *Lecture Notes in Artificial Intelligence*, vol. 2718, pp. 208–218, 2003.

[25] J. Pastor, M. Lafon, L. Trave-Massuyes, J.-F. Demonet, B. Doyon, and P. Celsis, "Information processing in large-scale cerebral networks: the causal connectivity approach," *Biological Cybernetics*, vol. 82, no. 1, pp. 49–59, 2000.

[26] I. Fagarasan, S. Ploix, and S. Gentil, "Causal fault detection and isolation based on a set-membership approach," *Automatica*, vol. 40, no. 12, pp. 2099–2110, 2004.

[27] J. Aslund, J. Biteus, E. Frisk, M. Krysander, and L. Nielsen, "Safety analysis of autonomous systems by extended fault tree analysis," *International Journal of Adaptive Control and Signal Processing*, vol. 21, no. 2-3, pp. 287–298, 2007.

[28] F. Yang, S. L. Shah, and D. Xiao, "Signed directed graph based modeling and its validation from process knowledge and process data," *International Journal of Applied Mathematics and Computer Science*, vol. 22, no. 1, pp. 387–392, 2012.

[29] M. A. Kramer and B. L. Palowitch, "A rule-based approach to fault diagnosis using the signed directed graph," *AIChE Journal*, vol. 33, no. 7, pp. 1067–1078, 1987.

[30] C.-C. Chang and C.-C. Yu, "On-line fault diagnosis using the signed directed graph," *Industrial and Engineering Chemistry Research*, vol. 29, no. 7, pp. 1290–1299, 1990.

[31] J. Thambirajah, L. Benabbas, M. Bauer, and N. F. Thornhill, "Cause-and-effect analysis in chemical processes utilizing XML, plant connectivity and quantitative process history," *Computers and Chemical Engineering*, vol. 33, no. 2, pp. 503–512, 2009.

[32] M. Bauer and N. F. Thornhill, "A practical method for identifying the propagation path of plant-wide disturbances," *Journal of Process Control*, vol. 18, no. 7-8, pp. 707–719, 2008.

[33] S. L. Bressler and A. K. Seth, "Wiener-Granger causality: a well established methodology," *NeuroImage*, vol. 58, no. 2, pp. 323–329, 2011.

[34] M. J. Kaminski and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," *Biological Cybernetics*, vol. 65, no. 3, pp. 203–210, 1991.

[35] L. A. Baccala and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biological Cybernetics*, vol. 84, no. 6, pp. 463–474, 2001.

[36] S. Gigi and A. K. Tangirala, "Quantitative analysis of directional strengths in jointly stationary linear multivariate processes," *Biological Cybernetics*, vol. 103, no. 2, pp. 119–133, 2010.

[37] L. Faes, A. Porta, and G. Nollo, "Testing frequency-domain causality in multivariate time series," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 8, Article ID 5416292, pp. 1897–1906, 2010.

[38] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, pp. 461–464, 2000.

[39] M. Bauer, J. W. Cox, M. H. Caveness, J. J. Downs, and N. F. Thornhill, "Finding the direction of disturbance propagation in a chemical process using transfer entropy," *IEEE Transactions on Control Systems Technology*, vol. 15, no. 1, pp. 12–21, 2007.

[40] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Springer, New York, NY, USA, 1999.

[41] F. Yang and D. Xiao, "Model and fault inference with the framework of probabilistic SDG," in *Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision (ICARV '06)*, pp. 1023–1028, Singapore, 2006.

[42] T. Richardson and P. Spirtes, "Automated discovery of linear feedback models," in *Computation, Causation, and Causality*, C. Glymour and G. Cooper, Eds., MIT Press, 2001.

[43] U. Feldmann and J. Bhattacharya, "Predictability improvement as an asymmetrical measure of interdependence in bivariate time series," *International Journal of Bifurcation and Chaos*, vol. 14, no. 2, pp. 505–514, 2004.

[44] M. Bauer and N. F. Thornhill, "Measuring cause and effect between process variables," in *Proceedings of the the IEEE Advanced Process Control Applications for Industry Workshop*, Vancouver, Canada, May 2005.

[45] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi et al., "Network modelling methods for FMRI," *NeuroImage*, vol. 54, no. 2, pp. 875–891, 2011.

[46] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu, "Methods for quantifying the causal structure of bivariate time series," *International Journal of Bifurcation and Chaos*, vol. 17, no. 3, pp. 903–921, 2007.

[47] Z. Q. Zhang, C. G. Wu, B. K. Zhang, T. Xia, and A. F. Li, "SDG multiple fault diagnosis by real-time inverse inference," *Reliability Engineering and System Safety*, vol. 87, no. 2, pp. 173–189, 2005.

[48] O. O. Oyeleye and M. A. Kramer, "Qualitative simulation of chemical process systems: steady-state analysis," *AIChE Journal*, vol. 34, no. 9, pp. 1441–1454, 1988.

[49] V. Venkatasubramanian, J. Zhao, and S. Viswanathan, "Intelligent systems for HAZOP analysis of complex process plants," *Computers and Chemical Engineering*, vol. 24, no. 9-10, pp. 2291–2302, 2000.

[50] F. Yang and D. Xiao, "Probabilistic signed directed graph and its application in hazard assessment," in *Progress in Safety Science and Technology*, P. Huang, Y. Wang, and S. Li, Eds., vol. 6, pp. 111–115, Science Press, Beijing, China, 2006.

[51] F. Yang, S. Shah, and D. Xiao, "SDG model-based analysis of fault propagation in control systems," in *Proceedings of the 22nd Canadian Conference on Electrical and Computer Engineering (CCECE '09)*, pp. 1152–1157, May 2009.

[52] S. Gentil and J. Montmain, "Hierarchical representation of complex systems for supporting human decision making," *Advanced Engineering Informatics*, vol. 18, no. 3, pp. 143–159, 2004.

[53] F. Yang and D. Xiao, "Hierarchical description of SDG model and its fault inference," in *Proceedings of the Seminar on Production Safety and Control in Petrochemical Industry*, pp. 1–5, Beijing, China, October 2006.

[54] J. Chen and J. Howell, "A self-validating control system based approach to plant fault detection and diagnosis," *Computers and Chemical Engineering*, vol. 25, no. 2-3, pp. 337–358, 2001.

[55] F. Yang, *Research on dynamic description and inference approaches in SDG model-based fault analysis*, Ph.D. thesis, Tsinghua University, Beijing, China, 2008.

[56] J. Shiozaki, H. Matsuyama, E. O'Shima, and M. Iri, "An improved algorithm for diagnosis of system failures in the chemical process," *Computers and Chemical Engineering*, vol. 9, no. 3, pp. 285–293, 1985.

[57] H. Vedam and V. Venkatasubramanian, "Signed digraph based multiple fault diagnosis," *Computers and Chemical Engineering*, vol. 21, supplement 1, pp. S655–S660, 1997.

[58] C. C. Han, R. F. Shih, and L. S. Lee, "Quantifying signed directed graphs with the fuzzy set for fault diagnosis resolution improvement," *Industrial and Engineering Chemistry Research*, vol. 33, no. 8, pp. 1943–1954, 1994.

[59] K. Takeda, B. Shibata, Y. Tsuge, and H. Matsuyama, "Improvement of fault diagnostic system utilizing signed directed graph–the method using transfer delay of failure," *Transactions of the Society of Instrument and Control Engineers*, vol. 31, no. 1, pp. 98–107, 1995.

[60] F. Yang and D. Xiao, "Approach to fault diagnosis using SDG based on fault revealing time," in *Proceedings of the 6th World Congress on Intelligent Control and Automation (WCICA '06)*, pp. 5744–5747, Dalian, China, June 2006.

[61] F. Yang and D. Xiao, "SDG-based fault isolation for large-scale complex systems solved by rough set theory," in *Proceedings of the 17th IFAC World Congress*, pp. 7221–7226, Seoul, Korea, 2008.

[62] H. Jiang, R. Patwardhan, and S. L. Shah, "Root cause diagnosis of plant-wide oscillations using the concept of adjacency matrix," *Journal of Process Control*, vol. 19, no. 8, pp. 1347–1354, 2009.

[63] R. J. Doyle, S. A. Chien, U. M. Fayyad, and E. J. Wyatt, "Focused real-time systems monitoring based on multiple anomaly models," in *Proceedings of the 7th International Workshop on Qualitative Reasoning About Physical Systems*, pp. 75–82, Eastsound, WA, USA, May 1993.

[64] J. Chen and J. Howell, "Towards distributed diagnosis of the Tennessee Eastman process benchmark," *Control Engineering Practice*, vol. 10, no. 9, pp. 971–987, 2002.