Metagraph Transformations and Workflow Management

Amit Basu and Robert Blanning
Owen Graduate School of Management
Vanderbilt University
Nashville, TN 37203
{basua, blanirw}@ctrvax.vanderbilt.edu

Abstract

Although workflow and process management have become important areas of research and development, there are no methods available yet for formal modeling and analysis of workflow systems. Available systems are largely limited to graphical visualization of workflows. We have developed an approach to formal analysis of workflows represented using a graph theoretic construct called a metagraph. In this paper, we show how certain graph-theoretic transformations on a workflow metagraph can be used to derive orthogonal metagraph representations that facilitate analysis of various dependencies between workflow components.

1 Introduction

Although many “legacy” information systems were built as isolated artifacts for use in specific narrow functions, most information systems developed today in business organizations do not fit this model. Rather, there is a growing realization that information systems are important resources throughout a variety of business processes, and their use spans not only different process activities but also different functional areas. Furthermore, most business processes are characterized by complex, non-sequential and overlapping workflows, and thus there is growing interest today in both the academic and corporate environments, in the area of workflow analysis and the use of information systems in workflow analysis.

Two streams of research on workflow analysis have evolved, each out of a distinct tradition; the first is work based on information systems analysis and design, while the second is based on database management research. In the first area, there have been a number of efforts to build workflow analysis tools similar to CASE tools, such as IBM’s Flowmark product. Also, a variety of methodological work has evolved [Fischer, 1995; Georgakopoulos, Hornick, and Sheth, 1995]. In the second area, most of the work is based on the notion that the activities comprising workflows can be viewed as very long transactions on distributed databases. Thus, much of the work has focused on the examination of transaction management issues such as scheduling, concurrency control and recovery as applied to the workflow environment.

What has been lacking in workflow analysis research so far is a formalization of workflows that supports effective visualization for control purposes, as well as formal procedures for performance analysis and planning of workflows. In order to address this, we have developed an approach based on the use of a graph-theoretic construct called a metagraph [Basu and Blanning, 1990(b)]. We have shown that metagraphs can be used to represent the important components of a workflow, such as information elements, documents, activities and resources, and can be used to formally analyze workflow structure. In this paper, we demonstrate that certain useful relationships between workflow components can be extracted using certain graph-theoretic transformations on workflow metagraphs. In particular, we develop the notion of the dual and pseudodual of a metagraph, and apply these to obtain useful views of workflow metagraphs.

The paper is organized in five sections. In section 2, we describe the essential features of workflows and processes. Then in section 3, we briefly overview metagraphs and show how they can be used to represent workflows. In section 4, we introduce the metagraph transformations and describe how they provide useful views of workflow metagraphs. Finally, section 5 presents some areas for future research.

2 Workflows

The notion of a workflow is generally used to describe the flow of information and work through one or more
organizational entities involved in business processes. Within any one business process, the workflow may span a set of activities conducted by one or more groups of resources (people, machines), with the activities occurring in some meaningful schedule. As an example, consider the processing of a loan application by a bank. The process starts with an applicant submitting a completed loan application, and completes with either the signing and recording of the approved loan or the recording of the rejected loan. In between, various activities take place, involving a variety of resources such as loan officers, credit analysts, branch managers, loan evaluation systems, fax machines, etc. While some activities depend upon other activities, and thus have to be processed sequentially, other activities can occur concurrently. Understanding the relationships between the various activities within a process and managing their scheduling, execution and control are important aspects of workflow analysis.

Workflows can be based on routine processes, such as the order fulfillment process or the manufacturing process, or ad-hoc processes, such as the process of dealing with crises and market changes. However, all workflows can be described in terms of a common set of components, which include the following:

1. An information element is an item or variable of interest to the process, such as sales volume, quantity on hand, purchase date, customer number, etc. or a proposition describing a state (or an assumption) such as items are out of stock, inflation is low, etc.

2. A report (or more generally, a document) is a collection of information elements that are handled together in one or more activities.

3. An activity is anything that takes an input set of information elements into some transformation that yields an output document (it may also have some side effects on the organization or its environment beyond the effects recorded in the outputs). In general, there is a many-to-many mapping between information elements and activities.

4. A resource is an entity that is essential for the execution of one or more activities. Thus, each activity has an associated set of resources (in general there is a many-to-many mapping between activities and resources).

5. A process is a collection of activities that together transform a collection of predetermined inputs into a set of relevant outputs. The input set of elements is called the source of the process, and the output set is known as its target.

6. A workflow system is an organization of the activities, resources and information elements needed for one or more processes within an organization (or within a related group of organizations).

Issues that are relevant to workflow analysis then include:

1. What activities, resources and information elements does a particular activity depend upon?

2. More generally, what are the relevant interactions between different activities, information elements and resources?

3. What is the impact of the removal of a component of the workflow upon the relevant processes?

4. What is a viable structure for a specific process (under a set of known conditions)?

5. How, in complex workflow systems, can specific relationships of interest be isolated and analyzed effectively?

Another important factor to consider is that the workflow within an organization often spans multiple processes. In other words, each organization has multiple business processes, and these processes often overlap, in the sense that they have certain activities in common. Similarly, multiple processes or even multiple activities within processes may share resources. For example, in the bank context, the task of assessing the impact of assets such as loans upon the bank's risk exposure is part of not just the loan evaluation process, but also the process of managing the bank's mortgage portfolio. In order to effectively manage the workflow within the bank, it is therefore important to recognize and exploit such commonalities between processes.

Our approach to representing and analyzing workflow systems is based on the use of a construct called a metagraph. In the next section, we briefly review the existing perspectives on workflow systems, and then in the following section, describe metagraphs and discuss their use in workflow representation.

## 3 Workflow Analysis Tools

There are two problems with workflow systems. The first is that they are often large and complex, thus requiring formal methodologies for their representation
and management [Khosafian and Buckiewicz, 1995; Mars hak, 1995]. The second is they are subject to change. The changes may be marginal, or they may be substantial changes resulting from a reengineering of the business processes that give rise to the workflows [Davenport, 1993]. In the latter case it is even more important that there be an explicit model of the workflow system, because changes made to one part of the system may have unforeseen repercussions in other parts of the system (i.e., a ripple effect). These repercussions may not be intuitively apparent, but they might be revealed by an examination or analysis of a formal workflow model.

To address this, a number of companies have begun to develop and market workflow software [Silver, 1995; Georgakopoulos, Hornick, and Sheth, 1995]. IBM offers two products - FlowMark, a workflow management system, and MQSeries, a messaging system - and it is continuing to do research in this area with its Exotica project [Mohan, et al., 1995]. In addition, some companies have developed their own proprietary systems [Morschlueuser, Raufer, and Wargitsch, 1996].

Although the response of the information technology industry to the need for workflow analysis tools is beneficial, there are two problems. The first is that most of the commercially available tools are primarily tools for visualization, not for analysis. That is, the tools allow their users (i.e., the designers of workflow systems) to construct graphical representations of workflow processes and to inspect the representations visually in an attempt to identify anomalies. Thus, they are primarily visualization tools and are not based on analytical (algebraic or logical) models of workflow processes which could be analyzed to detect potential or actual problems in workflow design.

The second problem is that the tools are not standardized. Thus, there is a need for a set of specifications of workflow processes that will allow for the integration of processes and for an interface between them and other information systems. A number of companies have joined together to form the Workflow Management Coalition (WFMC), which has developed a set of basic definitions of workflow systems and their components. But these are based on visualization and not analysis [Swenson, 1995]. Thus, there is still a need for an analytical approach to workflows.

4 Metagraphs and Workflow Representation

A metagraph is a graphical structure that represents directed relationships between sets of elements1. More formally,

**Definition I:** Given a finite generating set \( X = \{x_i, i = 1 \ldots I\} \), a metagraph is an ordered pair \( S = (X, E) \), in which \( E \) is a set of edges \( E = \{e_k, k = 1 \ldots K\} \). Each edge is an ordered pair \( e_k = (V_k, W_k) \), in which \( V_k \subseteq X \) is the vertex of the edge \( e_k \) and \( W_k \subseteq X \) is the output. The coinput of any \( x \in V_k \) is \( V_k \setminus \{x\} \) and the cooutput of any \( x \in W_k \) is \( W_k \setminus \{x\} \). Also \( V_k \cup W_k \neq \emptyset \) for all \( k \) [Basu and Blanning, 1994 (a,b)].

**Definition II:** Given a metagraph \( S = (X, E) \), a simple path from a source \( x \in X \) to a target \( x' \in X \) is a sequence of edges \( h(x, x') = \{e_1, e_2, \ldots, e_L\} \) such that \( x \in V_1, x' \in W_L \), and \( W_i \cap V_{i+1} \neq \emptyset \), \( \forall i = 1, \ldots, L - 1 \). The coinput of \( x \) is

\[
\left( \bigcup_{l=1}^{L} V_l \right) \setminus \{x\}
\]

and the cooutput of \( x' \) is

\[
\bigcup_{l=1}^{L} W_l \setminus \{x'\}
\]

The length of a simple path is the number of edges in the path; thus, the length of \( h(x, x') \) is \( L \). An edge is a simple path of length one. A cycle is a simple path from an element to itself; thus it is of the form \( h(x, x) \) for some \( x \in X \).

**Definition III:** Given a metagraph \( S = (X, E) \) a metapath from a source \( B \subseteq X \) to a target \( C \subseteq X \) is a set of edges \( M(B, C) = \{e_l, l = 1 \ldots L\} \) such that2:

1. There is a set of simple paths \( \{h_m(x_m, x'_m), m = 1 \ldots M\} \) with \( x_m \in B, x'_m \in C \) \( \forall m \), such that \( M(B, C) = \bigcup_{m=1}^{M} \text{Set}(h_m(x_m, x'_m)) \)

2. \( \left( \bigcup_{l=1}^{L} V_l \setminus \bigcup_{l=1}^{L} W_l \right) \subseteq B \)

3. \( C \subseteq \bigcup_{l=1}^{L} W_l \)

---

1 This material is taken from [Basu and Blanning, 1994 (a,d), 1996(a)], where it is presented in greater detail.

2 A set is denoted by \( \{a, b, c, \ldots\} \) and a sequence by \( \langle a, b, c, \ldots\rangle \). Also, the function \( \text{Set}(\langle a, b, c, \ldots\rangle) \) is a set for which \( \text{Set}(\langle a, b, c, \ldots\rangle) = \{a, b, c, \ldots\} \).
A useful notion for comparing metagraphs is that of dominance. A metagraph $M(B,C)$ is said to be dominant if there is no metapath to $C$ from any proper subset of $B$ and no subset $M' \subseteq M$ which is also a metapath from $B$ to $C$ [Basu and Blanning, 1994 (a,b)].

The algebraic analysis of metagraphs is based on a multiplication operator for adjacency matrices, which can be used to calculate the powers of an adjacency matrix [Basu and Blanning, 1994(a)]. Each member of $A^n$, $a_{ij}^n$, is a set of zero or more triples, one for each simple path of length $n$ connecting $x_i$ to $x_j$. The first component of the triple is the input of $x_i$ in the path, the second component is the output of $x_j$, and the third component is the path.

The closure $A^* = A + A^2 + A^3 + A^4 + \ldots$ of the adjacency matrix, discloses all paths of any length connecting any two elements, and it can also be used to identify metapaths. Since a metapath is the union of the edges in a set of simple paths, metapaths can be formed by combining triples in $a_{ij}^n$ for all $x_i \in B$ and $x_j \in C$ in such a way that the conditions of the definition of a metapath are met [Basu and Blanning, 1994 (b), 1996(a)].

We now describe some useful views of a metagraph. However, we first define a conditional metagraph.

**Definition IV:** ([Basu and Blanning, 1994 (c,d)]) A conditional metagraph is a metagraph $S = (X, E)$ in which:

1. $\forall e' \in E, V' \cup W' \neq \emptyset$ and $V' \cap W' = \emptyset$
2. $X = X_s \cup X_p$ with $X_s \cap X_p = \emptyset$ such that $\forall p \in X_p$
   (a) $\forall e' \in E$, if $p \in W'$, then $W' = \{p\}$
   (b) $\forall e' \in E$ and $e'' \in E$, if $p \in W'$ and $p \in W''$, then $e' = e''$
   (c) $\forall e' \in E$ and $e'' \in E$, if $p \in V'$ and $W'' = \{p\}$, then $W' \cap V'' = \emptyset$

Thus, in a conditional metagraph, the generating set is partitioned into a set of variables, denoted $X_s$, and a set of proportional statements, denoted $X_p$. Each $x \in X_s$ represents a variable such as revenue, production level or inflation rate. Each $p \in X_p$ represents a proposition, such as “The inflation rate is five percent or less,” which can be either true or false. A variable that appears in the inverse of an edge represents an input to the model represented by the edge. However, a proposition that appears in the inverse of an edge does not represent an input but rather an assumption that must be true for the model to be valid. A conditional metagraph is in effect a generalization of a simple metagraph, in that simple metagraphs can be viewed as conditional metagraphs in which $X_p = \emptyset$.

**Definition V:** Given a conditional metagraph $S = (X, X_p, E)$, a source $B \subseteq X_s$ and a target $C \subseteq X_s$, a conditional metapath is a set of edges $CM(B, C) = \{e'_i, i = 1, \ldots, L\}$ forming a metapath from $B \cup X_p$ to $C$ [Basu and Blanning, 1994 (c,d)]

In situations where the available base of modules is quite large, it is useful to extract relevant information using simplified views. In metagraphs such views can be defined in two ways: as a context and as a projection. A context of a conditional metagraph is defined by partitioning the assumptions in the metagraph into three sets: those known to be true (denoted $P$), those known to be false (denoted $Q$), and those whose truth values are unknown ($X_p \setminus (P \cup Q)$). Then, the conditional metagraph is simplified so that only the latter (unknown) set of assumptions is present. The following definition of a context is a constructive definition of the simplification process.

**Definition VI:** ([Basu and Blanning, 1994 (c,d)]) Given a conditional metagraph $S = (X, X_p, E)$, a set of assumptions $P \subseteq X_p$ that are known to be true and a set of assumptions $Q \subseteq X_p$ that are known to be false, we define a context, $K(P, Q, S)$ as a conditional metagraph derived from $S$ as follows:

1. for any edge $e' \in E$ containing an assumption $p \in P$ simplify the edge by deleting $p$; if the resulting edge has a null in- or out-vertex, delete the edge.
2. for any edge $e' \in E$ containing an assumption $q \in Q$, delete the edge.

Another type of higher-level view is a projection. In projecting a metagraph, a user specifies a subset of the generating set, and the projection is a simplified metagraph that contains only the relevant elements and edges.

**Definition VII:** ([Basu and Blanning, 1996 (a)]) Given a conditional metagraph $S = (X, X_p, E)$, a projection of $S$ over the set $X' \subseteq X_s$ is a conditional metagraph $N(X', S) = (X' \cup X_p, E')$ such that $e' \in E'$ if there is a dominant metapath from $V_{e'}$ to $W_{e'}$ in $S$.

Finally, we note that these two views of context and projection, are commutative - that is, $N(X', K(P, Q, S)) = K(P, Q, N(X', S))$ [Basu and Blanning, 1994(c)].

Having described metagraphs, we can now show how they can be used to represent the various elements of a workflow system. Each information element in a workflow can be represented as an element of the generating set $X_s$, or more specifically, of $X_p$. A collection
of information elements comprising a report can then be represented as a vertex. This presumes that each report is either the input or output of some activity; however, this is not a restrictive assumption, and characterizes most real workflows. Each activity is itself represented as an edge in the metagraph. This assumes that both the input and output of each activity is a report. As before, this assumption is reasonable, since the report comprising each activity’s input can be composed of elements from one or more reports (and/or manual inputs from some resource). Thus, a metagraph can be used to represent the activities comprising a workflow system.

A process in a workflow can be represented in the metagraph representation as a metapath from a set of information elements comprising a source to another set comprising the target. Specific assumptions made in each activity are represented in the metagraph by including the relevant propositions in the vertices of the activity edges. Similarly, resources needed for each activity are represented as additional inputs to the corresponding edge. The primary motivation for the separation of the generating set into the three component sets $X_a$, $X_p$, $X_r$ is that the evaluation of elements from each set is different. While information elements can have any value from their particular domain, propositions evaluate to either “true” or “false” (with an activity becoming viable only if all its assumptions evaluate to “true”) and resources evaluating to either “available” or “unavailable” (with an activity becoming viable only if all its resources are available). From a visualization perspective, the assumptions underlying an activity and the resources it needs can be presented to the user as labels on the edge itself, rather than as inverx assumptions; the former may be more intuitive.

As an example, consider the metagraph in Figure 1, which describes a highly simplified workflow for financial analysis for a new product.

The components of the workflow are as follows:

**Elements**
- $C.H$: Cost History
- $R.H$: Revenue History
- $M.C$: Market Conditions
- $C.E$: Cost Estimate
- $R.E$: Revenue Estimate
- $P.E$: Profit Estimate

**Activities**
- $rpe$: revenue projection estimation
- $rme$: revenue market estimation
- $che$: cost history estimation
- $pfe$: profit estimation

The metagraph identifies the relationships between the different information elements through the relevant activities. It is also possible to assign resources to activities by labelling the edges of the metagraph with the names of the relevant resources. A resource may be a particular person, program or piece of equipment, or it may be a class of individual resources that serve in the same role (e.g., financial analysts, word processing packages and workstations). We can thus assign roles as resources for activities, with each role being a named set of resources that are equivalent with respect to the requirements of that activity. In the case where a unique resource (e.g., John Doe) is needed, that resource can be viewed as belonging to a role containing only one member. This approach is used in [Basu and Blanning, 1996(b)]; in the current paper, we do not consider resources or roles in the metagraph representation.

In summary, the metagraph representation can be used to analyze dependencies between elements, the composition of processes and the role of specific activities and elements. It should be evident, however, that while this representation is comprehensive and facilitates the visualization and analysis of element dependencies, the analysis of activities is not as easy. In the next section, we show how certain dual transformations on metagraphs solve this problem.

## 5 Metagraph Transformations as Workflow Views

In this section, we introduce two transformations of metagraphs, the dual and the pseudodual, and show how they facilitate workflow analysis.

**Definition VIII:** Given a metagraph $S = <X, E>$, $X = \{x_i, i = 1, ... I\}$, and $E = \{e_j, j = 1, ... J\}$, its dual is a metagraph $S' = <X', E'>$ with $X' = \{\alpha, \beta, x'_j, j = 1, ... J\}$ and $E' = \{e'_i, i = 1, ... I\}$ such that:

1. For each primal element $x_i \notin \cup_j W_j$, there is a dual edge $e'_i = \langle \{\alpha\}, \{x'_i\}\rangle$.

2. For each primal element $x_i \notin \cup_j V_j$, there is a dual edge $e'_i = \langle \{x'_i\}, \{\beta\}\rangle$.
3. For each primal element \( x_i \in (\cup_j V_j) \cap (\cup_j W_j) \), there is a dual edge \( e'_i \) such that:

(a) for each primal edge \( e_j \) with \( x_i \in W_j, x'_j \in V'_j \), and there are no other dual elements in \( V'_j \);

(b) for each primal edge \( e_j \) with \( x_i \in V'_j, x'_j \in W'_j \), and there are no other dual elements in \( W'_j \);

4. There are no other dual edges.

The dual of the example workflow metagraph in Figure 1 is shown in Figure 2. Note that in the dual, the elements correspond to activities, while the edges correspond to information elements (the two special dual elements \( \alpha \) and \( \beta \) represent external input and output respectively). Thus, the dual identifies the interactions between different activities, and furthermore, identifies the information flows between activities. Thus, the dual of the workflow metagraph can be viewed as a form of Data Flow Diagram [Conger, 1994; Hoffer et al., 1996], a useful construct in information systems analysis and design.

Although the dual is a metagraph, its semantics are different. In the workflow context, the interpretation of a primal edge is that the activity represented by the edge takes all the input elements, and generates all the output elements. In other words, the primal can be interpreted as a conjunctive normal form (CNF) representation of element relationships. On the other hand, the dual is a disjunctive normal form (DNF) representation of the activity relationships. That is, each output activity for a dual edge requires the information element corresponding to the dual edge from any one of the activities in the edge’s invertex. This is not a very useful representation for resource planning, and furthermore, it complicates the algebraic analysis of the dual, since the algebraic operators on the adjacency matrix have to be modified.

To address this, we construct a related representation, called the pseudodual metagraph, which captures activity relationships in a CNF representation.

**Definition IX:** Given a metagraph \( S = \langle X, E \rangle \), \( X = \{x_i, i = 1, \ldots, J\} \), and \( E = \{e_j, j = 1, \ldots, J\} \), and its dual \( S' = \langle X', E' \rangle \), the corresponding pseudodual is a metagraph \( S'' = \langle X'', E'' \rangle \) with \( X'' = \{\alpha, \beta, x'_j, j = 1, \ldots, J\} \), \( E'' = \{e''_k, k = 1, \ldots, K\} \) and each edge \( e''_k \) constructed as follows:

1. For each dual edge \( e'_i \) of the form \( \langle \{\alpha\}, \{x'_i\} \rangle \), there is a pseudodual edge \( e'' = e'_i \).

2. For each dual edge \( e'_i \) of the form \( \langle \{\alpha\}, \{\beta\} \rangle \), there is a pseudodual edge \( e'' = e'_i \).

3. For each dual edge \( e'_i \) in which \( \{\alpha, \beta\} \notin (V'_j \cup W'_j) \), there are pseudodual edges constructed as follows:

   (a) Identify all dual elements \( x'_j \) such that \( x'_j \notin \{\alpha, \beta\} \) and there is no \( e'_j = \langle \alpha, x'_j \rangle \);

   (b) For each such \( x'_j \), identify all \( e'_j \) such that \( x'_j \in W'_j \). Let \( \theta_j \) be the set of all such \( e'_j \);

   (c) Construct the Cartesian product of all invertices \( V'' \) such that \( e_i \in \theta_j \); let this product be \( \gamma_j = \{\gamma_{jn}, n = 1, \ldots, N_j\} \);

   (d) Construct edges \( e'' = \{\gamma_{jn}, x_{jn}\} \) for \( n = 1, \ldots, N_j \).

4. There are no other pseudodual edges.

The pseudodual metagraph for the example metagraph of Figure 1 is shown in Figure 3. It should be evident that this represents activity interactions with a CNF interpretation for the edges. The edges now correspond to one or more information elements that are obtained from the respective activities in the edge invertices. Furthermore, the analogy to data flow diagrams still holds. This is a useful construct, not only for visualization, but also because all metagraph operations, such as the identification and evaluation of metapaths, identification of bridges and cycles, and the construction of higher level views such as projections and contexts, can be applied to the dual and pseudodual as well as for the primal.

6 Conclusion and Research Directions

In this paper, we have demonstrated that (1) metagraphs are a viable construct for representing workflows, and (2) the use of certain metagraph transformations such as the dual and pseudodual yield useful and intuitive views of workflows that offer both the analytical and visualization powers of metagraphs.

Clearly, this is just an initial step towards comprehensive computer-based support of workflow analysis. There are a number of directions that can be pursued towards this end. First, we are examining the use of metagraph views of resources, similar to the activity-oriented pseudodual. Also, once that is achieved, the representation of resource interactions can be exploited to deal with resource allocation and management issues. This is a promising goal, since to
date there has been very little research on how business process redesign impacts resource (and especially human resource) management and organizational design. Another promising direction is the integration of dependency based metagraph analysis with transaction management analysis of workflow systems currently being pursued by the database research community.

7 REFERENCES


Figure 1: An Example Workflow Metagraph (Primal)

Figure 2: The Dual Metagraph

Figure 3: The Pseudodual Metagraph

\[ x' : \text{RE (from rpe)} \land \text{CE} \]
\[ x'' : \text{RE (from rme)} \land \text{CE} \]