Software Models and Representations
Part 2
Data Flow Graphs

Approaches that are Problematic
- Natural Language
- Disciplined Natural Language
- Code

They have their advantages, but serious disadvantages too

Using Different Models to Represent Different Aspects of Software
- Model represents certain specific features
- Others are suppressed, not represented
- Aim is to clarify some issues by not having to worry about the others
- But this leads to different representations of reality
- How to see/infer the whole picture?

Plato’s Allegory of the Cave
The question of how to synthesize multiple views of “reality” is at least 2000 years old

Plato’s Allegory of the Cave
What is most tangible is least “real”
What is least tangible is most “real”
Plato could have been (was?) a great software engineer
**Graphs as Visualization Aids**

- A graph is a picture of a relation
- Graphs are mathematical structures with obvious visualizations that seem often to help many stakeholders visualize these relations.
- A graph’s edges visually represent the ordered pairs that compose the relation
- If the pairs in E are ordered, then G is a directed graph, and its edges are depicted with arrowheads
  - If not, the graph is called an undirected graph

**Example: Elevator Controller**

- What is it supposed to do?
  - Stop on every floor it is called to
  - Maybe not so easy: multiple elevators
  - Service users “first come, first served”
  - May conflict with optimal strategies
- These are hard to be precise about, reason about in “plain English”
- Code helps with some of these, but not all
  - What should it never do?
  - Allow elevator to move with doors open
  - Graphs are frequently used to specify this kind of thing
  - Many kinds of graphs for many purposes
  - A graph is a picture of a relation
Pictures are not enough

- Want to be able to reason about them
- Answer (stakeholder) questions
  - Can this data item ever reach that statement?
  - Can these two events ever happen in sequence?
  - What is the maximum time to execute that sequence?
- Pictures can leave ambiguous impressions
  - How to be sure what they say?
- Graphs, representing mathematical relations, can support deriving definitive answers
  - About the relation

What is This Graph Specifying?

And what can we safely infer/conclude from it?

Annotations Provide Intuitions

But are they suggesting too much?

What’s wrong with this diagram

There is ambiguity and misuse of notation here:
- one circle is a test, others are functions
- some edge annotations are data, some predicates
- are multiple arrows in and out “and” or “or”?

Back to Basics

- Review fundamental Finite Mathematics
  - Set theory
  - Graph theory
  - Predicate Calculus
  - Etc.
- But with an emphasis on why this is of interest in software engineering
Some Properties of Relations

Some familiar properties of ordered binary relations, $R$, over the set $S = \{s\}$:

- Symmetry: $s \in S, \text{ if } s \text{ R } t \Rightarrow t \text{ R } s$
- Reflexivity: $s \in S, \text{ for all } s \text{ in S}$
- Transitivity: $s \in S, \text{ for all } s \text{ in S}$

A relation that is symmetric, reflexive and transitive is called an equivalence relation.

If $R = \{(s, s)\}$ is transitive, then $R = \{(s, s) \mid \text{if } s \text{ calls } s, \text{ then } R \text{ is an irreflexive relation}\}$

If $R = \{(s, s)\}$ is irreflexive, then $R = \{(s, s) \mid \text{if } s \text{ calls } s, \text{ then } R \text{ is an irreflexive relation}\}$

Examples

If $s = \{\text{all subroutines written in Fortran}\}$ $s \in S, \text{ if and only if } s = \{\text{calls } s, \text{ then } R \text{ is an irreflexive relation}\}$

Let $PS = \{c, \text{ all the statements in a program that consists of a set of modules, } M = \{m\}\}$,

$\text{INMOD = } \{(c, c) \mid c \text{ and } c \text{ appear in the same module } m\}$

INMOD is an equivalence relation.

The relation ImmFol (earlier slide) is not transitive.

Change ImmFol to Fol, by defining Fol = \{(L1, L2) \mid \text{the execution of L2 may follow the execution of L1 for some execution of P}\}

Fol is still not transitive.

Graphs

A Graph, $G = (N, E)$ is an ordered pair, consisting of a node set, $N$, and an edge set, $E = \{(n, n)\}$

If $OG = (N, E)$ is an ordered graph with $E = \{(n, n)\}$, then its unordered version, $UG = (N, U)$, where $U = \{<n, n>\}$

Some Examples

Let $I = \{\text{all integers}\}$

Define $Q = \{(x, y, z) \mid x, y, z \text{ are integers and } y = x^2, z = x^3\}$

Let $S = \{\text{all states of the U.S.}\}$

Define $B = \{(S, S) \mid S \text{ and } S \text{ are states that share a border}\}$

Let $L = \{\text{all statements L in a program, P}\}$

Define ImmFol = \{(L1, L2) \mid the execution of L1 may immediately follow the execution of L2 for some execution of P\}

Examples

If $s = \{\text{all subroutines written in Fortran}\}$ $s \in S, \text{ if and only if } s = \{\text{calls } s, \text{ then } R \text{ is an irreflexive relation}\}$

Let $PS = \{c, \text{ all the statements in a program that consists of a set of modules, } M = \{m\}\}$,

$\text{INMOD = } \{(c, c) \mid c \text{ and } c \text{ appear in the same module } m\}$

INMOD is an equivalence relation.

The relation ImmFol (earlier slide) is not transitive.

Change ImmFol to Fol, by defining Fol = \{(L1, L2) \mid \text{the execution of L2 may follow the execution of L1 for some execution of P}\}
Graphs as Visualization Aids

- Graphs are mathematical structures with obvious visualizations that seem often to help many stakeholder communities to visualize key relations.
- A graph’s edges visually represent the ordered pairs that compose the relation.
- If the pairs in E are ordered, then G is a directed graph, and its edges are depicted with arrowheads.
- If not, the graph is called an undirected graph.

• Basic components of a data flow diagram:
  - Boxes (sometimes circles), represent I/O data
  - Both augmented by separate annotation relations
  - Edges, represented by arrows, are data flows between units
  - Nodes, represented by circles (boxes), are functional units

Paths

A path, \( P \), through an ordered graph \( G=(N,E) \) is a sequence of edges, \( (n_i, n_{i+1}) \) such that \( n_i \rightarrow n_{i+1} \) for all \( 2 \leq k \leq n \)

A path, \( UP \), thru an unordered graph \( UG=(N,U) \) is a sequence of edges, \( \{n_i, n_{i+1}\}, \{n_{i+1}, n_{i+2}\}, ...,\{n_{i+k}, n_{i+k+1}\} \) such that all of the \( \{n_i, n_{i+1}\} \) can be ordered to assure that \( n_i \rightarrow n_{i+1} \) for all \( 2 \leq k \leq n \).

In either case, \( n_1 \) is called the start node and \( n_n \) is called the end node.

The length of a path is the number of edges in the path.

A graph \( G \) is connected if and only if, for every pair of nodes, \( n_i, n_j \), there is a path from one of them to the other with \( G \) considered to be an unordered graph.

These graph constructs appeal visually to many stakeholders and often effectively support answering their questions.

Trees

A cycle in a graph \( G \) is a path whose start node and end node are the same.

A simple cycle in a graph \( G \) is a cycle such that all of its nodes are different (except for the start and end nodes).

If a graph \( G \) is connected and has no path through it that is a cycle, then the graph is called acyclic.

An acyclic, connected, unordered graph is called a tree.

A collection of trees is called a forest.

If the unordered version of an ordered connected graph is acyclic, the graph is called a directed acyclic graph (DAG).

If the unordered version of an ordered graph has cycles, but the oriented graph has no cycles, then the graph is called a Directed Acyclic Graph (DAG).

Other Types of Graphs

A Multigraph \( MG \) is an ordered pair \( MG = (N, C) \) where \( N \) is a set of nodes \( (n_i) \) and \( C \) is a collection of pairs of nodes (edges) with repetitions allowed (ie. \( C \) can be a multiset).

A Hypergraph \( HG \) is an ordered pair \( HG = (N, T) \) where \( N \) is a set of nodes \( (n_i) \) and \( T \) is a set of \( t \)-tuples of nodes, where \( t > 2 \).

A Hypermultigraph is a hypergraph where the set of \( t \)-tuples can be a multiset.

A bipartite graph \( BG \) is an ordered pair, \( BG = (BN, E) \) where \( BN \) is a node set that is the union of two disjoint subsets, \( \{n_i\} \) and \( \{n_j\} \), and no edge in \( E \) has both nodes in either \( \{n_i\} \) or \( \{n_j\} \).

A bipartite graph is often called a 2-colorable graph.

A \( k \)-colorable graph is defined analogously, with \( BN \) being the disjoint union of \( k \) subsets.

DATA FLOW Graph (Again)

- Capture system functionality.
- What does system do? How?
- Basic components of a data flow diagram:
  - Nodes, represented by circles (boxes), are functional units.
  - Edges, represented by arrows, are data flows between units.
  - Both augmented by separate annotation relations
  - Boxes (sometimes circles), represent I/O data

  - This is actually yet another relation

Differences in Graphs Result from Different Relations

- Data Flow:
  - Nodes represent set of sites where data is generated/used
  - Each edge is a (data generated, data used) node pair
- Control Flow:
  - Nodes represent units of functionality
  - \((n_i, n_j)\) is an edge if and only if unit \( n_i \) can execute immediately after \( n_j \) executes (ImmFol relation)
- Hierarchy:
  - Models “consists of” or “is a part of”
  - Key to divide-and-conquer approaches to understanding
- Finite State Machines
  - Nodes represent all possible different “execution states”
  - \((s_i, s_j)\) is an edge if and only if it is possible for state \( s_i \) to immediately succeed \( s_j \).
  - Called a transition from \( s_i \) to \( s_j \).
  - Edges annotated with transition condition
- Petri Nets
  - Multiple node and edge types in the same diagram
Formalizing DFGs as Relations

- \((i, j) \in \text{DataFlow}_G\) if node \(i\) creates data that node \(j\) uses in \(G\)
- \(\text{INPUT}_G\) is the set of nodes (input) that is a provider of input to \(G\) from an external source
- \(\text{OUTPUT}_G\) is the set of nodes (output) that is a conveyor of artifacts computed by \(G\) to an external source
- \((e, \text{operand}) \in \text{EdgeAnnotation}_G\), if the operand is the name of the artifact that flows along an edge \(e \in E\)
- \((n, \text{text}) \in \text{NodeAnnotation}_G\), if the string \(\text{text}\) describes the functioning of node \(n\)

Questions this helps answer:
- Why create this data? Who uses this data? What results does the end user see? What does the end user have to input?
- How does a node do its job?

Flowgraphs

Let \(S = \{\text{all statements } s \text{ in a program, } P\}\)
Let \(\text{ImmFol} = \{(s_i, s_j) \mid \text{The execution of } s_j \text{ immediately follows the execution of } s_i \text{ for some execution of } P\}\)
Then: If \(FG = (S, \text{ImmFol})\), \(FG\) is called the flowgraph of \(P\)
\(FG\) is an ordered graph
Every execution sequence (ie. the sequence in which the statements of \(P\) are executed for a given execution of \(P\)) corresponds to a path in \(FG\).
However---the converse is not true. A path through \(FG\) may not correspond to an execution sequence for \(P\)
A loop in \(P\) appears as a cycle in \(FG\)

Callgraphs

Let \(\text{PROC} = \{\text{procedures } S \text{ that the program } P \text{ comprises}\}\)
Let \(\text{CALLS} = \{(S_i, S_j) \mid S_j \text{ is directly invoked from } S_i \text{ during some execution of } P\}\)
Then \(CG = (\text{PROC}, \text{CALLS})\) is called the Call Graph of \(P\)
\(CG\) is a directed graph
If \(P\) is written in a language that does not allow recursion, then \(CG\) will be acyclic
A cycle in \(CG\) indicates that the nodes along the cycle participate in a recursive calling chain

NOTE: DEPICTIONS OF THESE GRAPHS MAY BE SUPERIMPOSED OVER EACH OTHER TO CLARIFY (?)

Consistency is a principal concern

- Are the diagrams consistent with each other?
- Top view consistent with elaborations?
  - Arrows consistent
  - Data flows consistent
  - Other semantics?
- Are we seeing different shadows of the same object?
- Invitation to subtle errors

Hierarchy

- Enables incrementally adding detail
- Increased precision too
- Draws upon innate human mental capability
  - Abstraction
  - Encapsulation
- A typical solution to the problem of needing detail, but needing to avoid overload
- But creates potential problems

Plato’s Allegory of the Cave

The question of how to synthesize multiple views of “reality” is at least 2000 years old
Plato’s Allegory of the Cave

What is most tangible is least “real”
What is least tangible is most “real”

Plato could have been (was?) a great software engineer

Simple Hierarchical Elaboration

Something a Little More Complex

Hierarchical Elaboration
(of “Select New Floor”)

Is this consistent with its parent?

Inconsistencies with parent
Hierarchical Elaboration (of “Select New Floor”)

Current Floor
Floors to
Target
Floor
Direction
Floors in
the right
direction
Floor to
Move to
Floor Other
Cars
Do Updates

Request List
View of Request List
Do Updates

Some Example of “real world” DFGs

- IDEF
- Kepler
- Scientific Workflows

A Kepler Example

Scientific Workflow Graph

The need for this mathematical precision

- Appeals to some stakeholders
  - Verifiers, regulators, innocent bystanders (e.g. people living near nuclear reactors)
- Supports tools that can verify these consistency relations automatically
- Reduces the need to trust fallible humans
More specifically

Let \((n, \text{somenode}) \in \text{NodeAnnotation}_G\) for some \(n, \text{somenode} \in N\), where \(G = (N, E)\) is the “parent graph”

Let \(\text{somenode} = (N', E')\)

This is the DFG elaborating on \(\text{somenode} \in N\)

Two easy consistency properties

- If \(\exists m \text{ such that } (m, \text{somenode}) \notin E\),
  then \(\text{Cardinality} (\text{INPUT}_{\text{somenode}}) \neq 0\)
- If \(\exists m \text{ such that } (\text{somenode}, m) \notin E\),
  then \(\text{Cardinality} (\text{OUTPUT}_{\text{somenode}}) \neq 0\)

(Some) Consistency definitions

- Let \((n, \text{somenode}) \in \text{NodeAnnotation}_G\) for some \(n, \text{somenode} \in N\), where \(G = (N, E)\)
- Let \(\text{somenode} = (N', E')\)
  This is the DFG elaborating on \(\text{somenode} \in N\)
- More specifically
  - Let \(\text{INS}_{\text{somenode}, a}\) be the subset of \(E\), \((*, \text{somenode})\)
    for which \(\text{somenode}\) is the second component.
  - Then \(\text{INS}_{\text{somenode}, a} \models \text{INPUT}_{\text{somenode}}\)

And similarly:

- Let \(\text{OUTS}_{\text{somenode}, a}\) be the subset of \(E\), \((\text{somenode}, *)\)
  for which \(\text{somenode}\) is the first component.
- Then \(\text{OUTS}_{\text{somenode}, a} \models \text{OUTPUT}_{\text{somenode}}\)

More Semantics

- The annotation on each input to \(\text{somenode} \in N\), the node set of \(G\), must match the annotation on some edge, \(e' \in E\)
  \(\text{INPUT}_{\text{somenode}} \subseteq E'\), the edge set of \(\text{somenode} = (N', E')\)

  MORE PRECISELY

- Suppose \(e = (n, \text{somenode}) \in E\), and
  \(e, \text{name} \in \text{EdgeAnnotation}_{\text{somenode}}\)

  Then \(\exists e' = (n', \text{somenode}) \in E'\) \(\text{somenode}\), where \(e' \in \text{INPUT}_{\text{somenode}}\)
  for which \((e', \text{name}) \in \text{EdgeAnnotation}_{\text{somenode}}\)

- And conversely

  Similarly for OUTPUT
Specifying AND vs OR

- Node cannot begin until data arrives along all in-edges
  - Let $G = (N, E)$, then $\forall n \in N$, let $\text{INS}_n$
    - $e \in E \mid e = (n', n)$, for some $n' \in N$

Then $n$ represents a function,

$$f(i_1, i_2, ..., i_t) = \text{INANNS}_n$$

Need something similar to specify DFG for which ANY input is sufficient to trigger $n$

And something similar for AND and OR outputs

These sorts of constraints can support additional types of reasoning:

Eg. about parallelism

Why the need for this mathematical precision?

- Appeal to some stakeholders
  - Verifiers, regulators, innocent bystanders (e.g. people living near nuclear reactors)
- Supports tools that can verify these consistency relations automatically
  - Reduce the need to trust fallible humans

Back to Previous Example

Data needs precise specification too

- DFG's focus on functionality, using data as a vehicle
- Data shown as unstructured atomic units—usually unrealistic
- Complex functions cannot be adequately defined without delving into the details of how they handle structured data
- Sub-DFG's can show how the high level data that high level DFG's deal with is decomposed
  - But this is implicit data definition
  - Can be hard to read/inconsistent
- Data specification is worth doing explicitly, carefully
- Usually using Disciplined Natural Language—eg. Templates
- Formal specification of data is an important future topic

Elaboration of “Select New Floor” Suggests Artifact Issues

- DFG’s focus on functionality, using data as a vehicle
- Data shown as unstructured atomic units—usually unrealistic
- Complex functions cannot be adequately defined without delving into the details of how they handle structured data
- Sub-DFG’s can show how the high level data that high level DFG’s deal with is decomposed
  - But this is implicit data definition
  - Can be hard to read/inconsistent
- Data specification is worth doing explicitly, carefully
- Usually using Disciplined Natural Language—eg. Templates
- Formal specification of data is an important future topic
More Broadening

- Use of "open boxes" to indicate data store
  - A different set, with different semantics
  - Not a computation function
  - Methods are: put, get, search(?)

Still More Broadening

- Different shapes of boxes
- Different pictures instead of boxes
- Different colors
- Different lines

Central questions: What are the semantics? Does this really help? Or confuse?

IDEF0

- Commercial DFG formalism
- Some formality and rigor behind it
- Primarily pictorial
- In wide use
- Additional semantics in IDEF1, IDEF2, etc.

Note: The edges here do not comprise a set. They comprise a collection.
Scientific Workflow Graph

But iteration complicates matters

Kepler—Another DFG Technology

- Data Flow Graph notation
- Has hierarchical decomposition
- Capability for specifying DFG semantics
  - For each diagram
  - Can be different at each level of hierarchy (!)
- Based on Ptolemy II system
What kinds of questions are well addressed by DFGs?

- Overall structure of functional capabilities
  - What does this piece do?
- System outputs and inputs
- How might changes be made?
- What functions create what data entities

Given that Precision is essential

- What about the other three dimensions?
- Detail
  - Gained from hierarchical elaboration
- Breadth
  - Comes from different (sub)types of DFG
- Clarity
  - Seems to be reduced by increased Detail and Breadth
  - With the need for Precision

Considerable Appeal, but Limited Value, to most stakeholders

- Users think they have sufficient understanding
  - But have trouble being able to see easy things (iteration)
- Developers have same problem
- Managers may only care to see easy things (!)
  - Although they should be interested in more
- Bystanders may be shown only easy things
  - Which could be a real problem

Final observations

- Very primitive representation
  - Very limited semantics
- But actually more a family of model types
  - Different sets of semantics
- The actual relation(s) are rarely made clear and precise
- Powerful aid to intuition and efficiency of communication
  - Clear advantages over natural language
- But is intuition misled by ambiguity, misinterpretation?
- Does not help explain HOW things get done