Computer Science 520/620
Spring 2013
Prof. L. Osterweil

Software Models and Representations
Part 2
Data Flow Graphs

Being
Clear
Precise
Comprehensive
Detailed
About Software

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

- 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.

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

Data Flow Graph

Better Edge Annotations

Control Flow Graph

Finite State Machine
Petri Net

- Up button Pressed
- Stop at Floor
- Car going up
- No Press
- Go to Next Floor
- Up button Pressed

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?

- func. 1
- func. 2
- func. 3
- input arg 1
- input arg 2
- t.2.in
- t.3.in
- out 1
- print

Annotations Provide Intuitions

- input height
- check arg. values
- compute area
- print area
- input width
- width
- area
- output error msg
- print error msg

What’s wrong with this diagram

- input height
- width
- check arg. values
- compute area
- args
- bad
- output error msg
- print error msg

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.
Relations:

A relation, R, over a set, S = \{s\} is a set of tuples

R = (r), where r = (s_i, s_j, ..., s_n)

An n-ary relation is a relation where all of the tuples are n-tuples

A Binary relation is a relation where all the tuples are 2-tuples

If (s_i, s_j) is an element of R, then we often write s_i R s_j

Another view of relations:

The relation, R, over the set S can be defined as: R = \{(s_i, s_j) \mid PRED(s_i, s_j) = \text{True, for some predicate, PRED}\}

If the tuples are ordered, the relation is called an ordered relation

If the tuples, <s_i, t_j, ..., u_k> are unordered, the relation is an unordered relation

Some Properties of Relations

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

Symmetry: \( s \, R \, s \) \iff \( s \, R \, s \) for all pairs, s and s in S

Reflexivity: \( s \, R \, s \) for all s in S

Transitivity: \( s \, R \, s_i \) and \( s_i \, R \, s_j \) \implies \( s \, R \, s_j \) for all s, s_i, s_j in S

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

If \( R = \{(s_i, s_j)\} \) is transitive, then \( C_R = \{ (s_i, s_j) \} \) there exists a sequence, \( i_1, i_2, ..., i_n \), in such that \( s_{i_1} = s_i, R \, s_i, s_j = s_n \), .... \( s_{i_n} \, R \, s_n = s_j \) is called the transitive closure of R

Antisymmetry: \( s \, R \, s_j \iff \neg(s \, R \, s_i) \) for all pairs, s and s in S

Irreflexivity: \( \neg s \, R \, s \) for all s in S

Examples

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

Let S = \{all states of the U.S., S\}
Define \( B = \{(S, S) \mid S \text{ and } S \text{ are states that share a border}\}\)

Let L = \{all statements L in a program, P\},
Define \( \text{ImmFol} = \{(L, L) \mid \text{the execution of } L \text{ may immediately follow the execution of } L \text{ for some execution of } P\} \)

Some Examples

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

Let S = \{all states of the U.S., S\},
Define \( B = \{(S, S) \mid S \text{ and } S \text{ are states that share a border}\}\)

Let L = \{all statements L in a program, P\},
Define \( \text{ImmFol} = \{(L, L) \mid \text{the execution of } L \text{ may immediately follow the execution of } L \text{ for some execution of } P\} \)

Examples

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

Let PS = \{c, \text{ all the statements in a program that consists of a set of modules, } M=\{m\}\},
INMOD = \{(c, c) \mid c \text{ and } c \text{ appear in the same module } m\}

INMOD is an equivalence relation

The relation \( \text{ImmFol} \) (earlier slide) is not transitive

Change \( \text{ImmFol} \) to Fol, by defining \( \text{Fol} = \{(L_1, L_2) \mid \text{the execution of } L_2 \text{ may follow the execution of } L_1 \text{ for some execution of } P\} \)

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, \text{ U}), where U=\{(n, n)\}
Graphs as Visualization Aids

- Graphs are mathematical structures with obvious visualizations that seem often to help many stakeholders 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.

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 ordered graph itself has no cycles, then the graph is called a Directed Acyclic Graph (DAG).

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.

Paths

A path, P, through an ordered graph G=(N, E) is a sequence of edges, ((n_1, n_2), (n_2, n_3), ..., (n_k, n_{k+1})) such that n_{i+1} = n_i for all 2 <= k <= n.
A path, UP, thru an unordered graph UG=(N, U) is a sequence of edges, ((n_1, n_2), (n_2, n_3), ..., (n_k, n_{k+1})) such that all of the (n_i, n_j) can be ordered to assure that n_{j+1} = n_j for all 2 <= k <= n.
In either case, n_1 is called the start node and n_{k+1} 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_1, n_2, 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.

Other Types of Graphs

A Multigraph MG is an ordered pair MG = (N, C) where N is a set of nodes (n) and C is a collection of pairs of nodes (edges) with repetitions allowed (i.e. C can be a multiset).
A Hypergraph HG is an ordered pair HG = (N, T) where N is a set of nodes (n) and T is a set of t-tuples of nodes, where t > 1.
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_1 U N_2, and no edge in E has both nodes in either N_1 or N_2.
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.

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 \(G\)
• \(\text{INPUT}_G\) – set of nodes (input) such that input is a provider of input to \(G\) from an external source
• \(\text{OUTPUT}_G\) – set of nodes (output) such that output is a conveyor of artifacts computed by \(G\) to an external source
• \((e, \text{operand}) \in \text{EdgeAnnotation}_G\), if \(e\) is the name of the artifact that flows along an edge \(e \in E\), where \(G = (N,E)\)
  – Preferably the data artifact is defined rigorously
• \((n, \text{text}) \in \text{NodeAnnotation}_G\), if the string \text{text} describes the functioning of node \(n \in N\), where \(G = (N,E)\)
  – This may imply different semantics (e.g. abstraction)
• \(\text{InputAnnotation}\), \(\text{OutputAnnotation}\) are similar

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?
• Questions this can’t answer: What is the exact sequence of events?

Questions this can’t answer: What is the exact sequence of events?

How does a node do its job?

Flowgraphs

Let \(S = \{\text{all statements } s \text{ in a program, } P\}\)

Let \(\text{ImmFol} = \{(s_i, s_j) | \text{the execution of } s_j \text{ immediately follows the execution of } s_i \text{ for some execution of } P\}\)

Then: if \(FG = (S, \text{ImmFol})\), \(FG\) is called the flowgraph of \(P\)

\(FG\) is an ordered graph

Every execution sequence (i.e., 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 \(\text{PROC} = \{\text{procedures } S_i \text{ that the program } P \text{ comprises}\}\)

Let \(\text{CALLS} = \{(S_i, S_j) \mid S_j \text{ is directly invoked from } S_i \text{ during some execution of } P\}\)

Then \(\text{CG} = (\text{PROC}, \text{CALLS})\) is called the Call Graph of \(P\)

\(\text{CG}\) is a directed graph

If \(P\) is written in a language that does not allow recursion, then \(\text{CG}\) will be acyclic

A cycle in \(\text{CG}\) 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 (?!)

Things

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

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

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

Something a Little More Complex

Hierarchical Elaboration (of “Select New Floor”)

Hierarchical Elaboration
(of “Select New Floor”)

Is this consistent with its parent?

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

Back to Previous Example

Elaboration of “Select New Floor”
Suggests Artifact Issues

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

(Some) Consistency definitions
• Let (n, somenode) ∈ NodeAnnotation_{ParentGraph}
  for some n, somenode ∈ N, where ParentGraph = (N, E)
• Let somenode = (N’, E’)
  --This is the DFG elaborating on “somenode”
• Some consistency properties
  --If ((m, somenode)) ≠ ø, m, somenode ∈ N,
    then Cardinality (INPUT(somenode)) ≠ 0
  --If ((somenode, m)) ≠ ø, m, somenode ∈ N,
    then Cardinality (OUTPUT(somenode)) ≠ 0
• Maybe some others(?)
  --If Cardinality ((m, somenode)) = k, m, somenode ∈ N,
    then Cardinality (INPUT(somenode)) = k
  --If Cardinality ((somenode, m))) = k, m, somenode ∈ N,
    then Cardinality (OUTPUT(somenode)) = k

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.

Broadening DFG Semantics

- Node cannot begin until data arrives along all in-edges
  - Let $\text{DFG} = (N, E)$, where $N$ is set of nodes, $(n_i)$. If $n_i$ is a function, then it must be defined on the set of all artifacts $\mathcal{A}_i = \{a_{j,i}\}$, such that $((n_j, n_i), a_{j,i}) \in \text{EdgeAnnotation}_{\text{DFG}}$

- How to adapt this for
  - any semantics?
  - for output semantics?
  - exactly one Output node
  - Etc.

These sorts of constraints can support additional types of reasoning:
Eg. about parallelism
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?

Scientific Workflow Graph

Attempt to make this more precise

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
    "Modeling of Sensor Nets in Ptolemy II", in Proc. of
A Kepler Example

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

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

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

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