Going quantitative in software modeling

TRUST 2nd Workshop

Univ. of Minho, October 23, 2017

J.N. Oliveira

INESC TEC & Univ. of Minho

(POCI-01-0145-FEDER-016826)

Motivation

Software modeling

Simplicity + elegance = **effectiveness** (Dijkstra)

Alloy — *writing less to say more* :-)

However: qualitative features simpler to model than quantitative ones

"Quantitative **abstraction**"?

"Scalable modeling": the *"keep definition, change category"* lemma.

Starting point — what is **modeling language**, after all?

Motivation

Software modeling

Simplicity + elegance = **effectiveness** (Dijkstra)

Alloy — *writing less to say more* :-)

However: qualitative features simpler to model than quantitative ones

"Quantitative **abstraction**"?

"Scalable modeling": the *"keep definition, change category"* lemma.

Starting point — what is **modeling language**, after all?

Category

Abstract language made of arrows which (may) compose with each other, and such that

(a) **associativity**

$$c \cdot (b \cdot a) = (c \cdot b) \cdot a \qquad (1)$$

holds.

$$\begin{tikzcd}
q & \\
k \arrow[u, "c"'] & m \arrow[l, "b"'] \arrow[ul, "c\cdot b"'] \\
& n \arrow[u, "a"'] \arrow[ul, "b\cdot a"]
\end{tikzcd}$$

(b) every **object** a has an **identity** such that:

$$1 \cdot a = a \cdot 1 = a \qquad (2)$$

$$\begin{tikzcd}
m & m \arrow[l, "1"'] \\
n \arrow[u, "a"] & n \arrow[l, "1"] \arrow[u, "a"'] \arrow[ul, "a"']
\end{tikzcd}$$

Thus, arrows form a **monoid**.

Enriched categories

Arrows can be added

$$n \xrightarrow{a} m \qquad n \xrightarrow{b} m \qquad n \xrightarrow{a+b} m$$

and can be multiplied

$$n \xrightarrow{a} m \qquad n \xrightarrow{b} m \qquad n \xrightarrow{a\times b} m$$

such that, under $(\cdot)$, $\times$ and $+$, arrows form two **semirings**:

$$
\begin{array}{ll}
a + (b + c) = (a + b) + c & a + 0 = a = 0 + a \\
a \times (b \times c) = (a \times b) \times c & a \times \top = a = \top \times a \\
a + b = b + a & \\
a \times (b + c) = a \times b + a \times c & a \times 0 = 0 = 0 \times a \\
a \cdot (b + c) = a \cdot b + a \cdot c & a \cdot 0 = 0 = 0 \cdot a
\end{array}
$$

"Dagger" categories

Further structure — for every arrow $k \xrightarrow{a} q$ there exists an arrow $k \xleftarrow{a^\circ} q$, the **converse** of a, such that:

$$
\begin{aligned}
(a^\circ)^\circ &= a \\
(a \cdot b)^\circ &= b^\circ \cdot a^\circ \\
(a + b)^\circ &= a^\circ + b^\circ \\
(a \times b)^\circ &= a^\circ \times b^\circ
\end{aligned}
$$

NB: "dagger" because a° often written as $a^\dagger$.

Famous **counter**-example: category of sets and functions.

Idempotency

Additive operator $+$ makes a difference.

$+$-**idempotency**: wherever $a + a = a$ holds for all a, then

$$a \leqslant b \quad \stackrel{\mathrm{def}}{=} \quad a + b = b \tag{3}$$

is a partial order.

Clearly, $0 \leqslant a$ for all a and $(+)$ is the *lub* with respect to $\leqslant$:

$$a + b \leqslant c \quad \equiv \quad a \leqslant c \wedge b \leqslant c \tag{4}$$

NB: $c := a + b$ in (4) means $a + b$ is upper bound; $\Leftarrow$ means it is the **least** upper bound (*lub*).

Relational algebra is an example of such idempotency (next slide).

Binary relations

The algebra of **binary relations** is a well known example of such enriched categories:

Categorial	**Binary relations**	**Description**
$x \cdot y$	$R \cdot S$	composition
$x + y$	$R \cup S$	union
$x \times y$	$R \cap S$	intersection
0	$\perp$	empty relation
1	id	identity relation
$\top$	$\top$	top relation
x°	R°	converse relation
$x \leqslant y$	$R \subseteq S$	inclusion

$\cup$-**idempotency** brings about the $R \subseteq S$ partial order, thus enabling recursion, iteration etc. — but it hinders implicit expression of **quantities** (cf. numbers in Alloy).

Matrices

In case addition is **not** idempotent — eg. $x + x = 2\,x$ — we get a typed **linear algebra** of matrices ("as arrows"):

Categorial	Matrices	Description
$x \cdot y$	$M \cdot N$	MMM
$x + y$	$M + N$	pointwise addition
$x \times y$	$M \times N$	Hadamard product
0	$\bot$	everywhere-0 matrix
1	id	identity matrix
$\top$	$\top$	everywhere-1 matrix
x°	M°	transpose matrix

$\{0, 1\}$-valued (Boolean) matrices represent binary relations, where

$$M \cap N = M \times N$$
$$M \cup N = M + N - M \times N.$$

(So the $+$-semiring must be a **ring**.) By default, in this talk we assume $\mathbb{Z}$-valued matrices.

Functions

Functions are Boolean matrices (relations) such that $! \cdot f = !$, where $k \xrightarrow{\ !\ } 1 = \top$. ($!$ is itself a function; $1 \xrightarrow{\ !\ } 1 = id$.)

Functions enjoy quite a number of properties, in particular, for f and g functions,

$$
\begin{aligned}
y\,(g^\circ \cdot M \cdot f)\,x &= (g\,y)\,M\,(f\,x) && (5)\\
y\,(f \cdot M)\,x &= \langle \textstyle\sum z : y = f\,z : z\,M\,x \rangle && (6)\\
y\,(M \cdot f^\circ)\,x &= \langle \textstyle\sum z : x = f\,z : y\,M\,z \rangle && (7)
\end{aligned}
$$

For relations, similar laws hold just by replacing $\sum z$ by $\exists\, z$.

In the sequel, we shall denote by $\mathbb{R}$ — resp. $\mathbb{M}$ — the category of binary **relations** — resp. $\mathbb{Z}$-valued **matrices**.

Abstract model

In $\mathbb{R}$, to begin with:

$$
\begin{tikzcd}
 & Keyword & \\
Paper \arrow[ur, "K"] \arrow[r, "C"] & Paper \arrow[r, "A"] \arrow[d, "m"] & Author \\
 & Medium &
\end{tikzcd}
\qquad (8)
$$

where

- $c' \; C \; c$ means c' is **cited** by c or c cites c'
- $k \; K \; p$ means that paper p has **keyword** k
- $m \; p$ is *the* **publication** medium of paper p (a function)
- $a \; A \; p$ means a is among the **authors** of paper p.

Alloy

$$
\begin{array}{l}
\textbf{sig}\ \mathit{Paper}\ \{\\
\quad C : \textbf{set}\ \mathit{Paper},\\
\quad K : \textbf{set}\ \mathit{Keyword},\\
\quad A : \textbf{set}\ \mathit{Author},\\
\quad m : \textbf{one}\ \mathit{Medium}\\
\}
\end{array}
$$

Papers cannot cite themselves: $C \subseteq \neg id$, that is

$$\textbf{fact}\ \{\textbf{no}\ \ C\ \&\ \mathit{iden}\}$$

since $\neg R = R \Rightarrow \bot$ and implication is defined by GC

$$X \cap Y \subseteq Z \Leftrightarrow X \subseteq Y \Rightarrow Z. \tag{9}$$

Triangular patterns

In category $\mathbb{R}$:

$$
\begin{tikzcd}
 & Keyword \\
Paper \arrow[ur, "K"] \arrow[ur, bend left, "R"] \arrow[r, "C"] & Paper \arrow[u, "K"] \arrow[u, bend right, "S"']
\end{tikzcd}
\qquad
\left\{
\begin{array}{l}
R = K \cap K \cdot C \\
S = K \cap K \cdot C^{\circ}
\end{array}
\right.
$$

R is not particularly interesting.

But S is so,

$$k \; S \; p \Leftrightarrow k \; K \; p \wedge \langle \exists \; q \; : \; p \; C \; q : \; k \; K \; q \rangle$$

meaning: *paper* p *is cited by at least another (btw different) paper* q *"in the same area" (keyword* k*).*

Triangular patterns (composition)

$$\begin{tikzcd} & Keyword & \\ Paper \arrow[ur, "K"] \arrow[r, "C"'] & Paper \arrow[u, "S"] \arrow[r, "A"'] & Author \arrow[ul, "Q"'] \end{tikzcd}$$

$$\begin{cases} S = K \cap K \cdot C^\circ \\ Q = S \cdot A^\circ \end{cases}$$

Then

$$Keyword \xleftarrow{Q} Author \ = S \cdot A^\circ$$

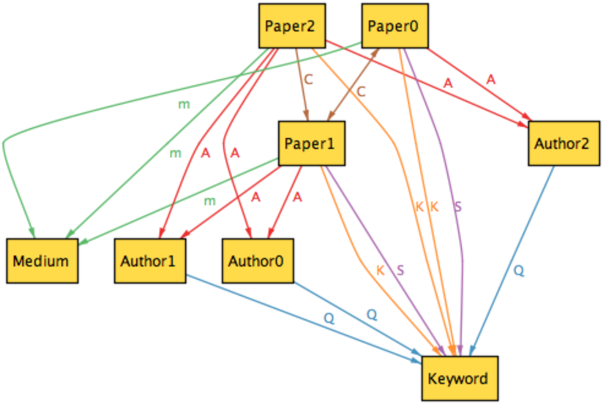
is such that

$$k \ Q \ a = \langle \exists \ p \ : \ a \ A \ p : \ k \ S \ p \rangle$$

telling which authors have cited papers in particular areas (keywords).

Alloy

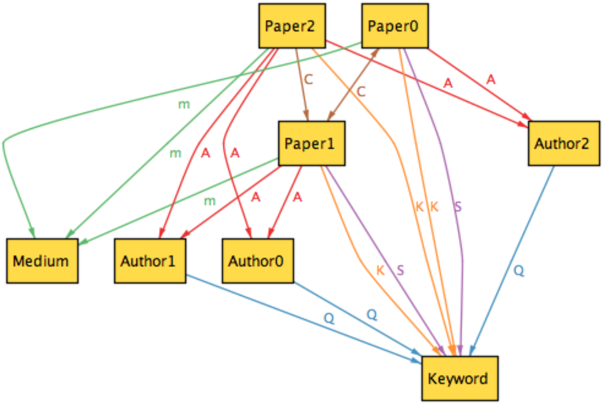
We can do model analysis...



... but no bibliometrics! Why? Idempotency!

Alloy

We can do model analysis...



... but no biblio<u>metrics</u>! Why? Idempotency!

Keep definition!

Shall we add **quantitative** information to the model?

No! *Recall scalable modeling: "keep definition, change category".*

It suffices to interpret the **same** (abstract) model in category $\mathbb{M}$ — e.g. pattern

$$Keyword \xleftarrow{S} Paper \;= K \times (K \cdot C^\circ)$$

will now **count** how many papers cite a given one, all within the same area:

$$\begin{array}{l} k\ S\ p = \\ \quad \textbf{if}\ (k\ K\ p)\ \textbf{then}\ \langle \sum\ q\ :\ p\ C\ q \wedge k\ K\ q :\ 1 \rangle\ \textbf{else}\ 0 \end{array}$$

Keep definition!

In $\mathbb{M}$, the arrows in

$$\begin{tikzcd}
& Keyword & \\
Paper \arrow[ur, "K"] \arrow[r, "C"] & Paper \arrow[u, "S"] \arrow[r, "A"] & Author \arrow[ul, "Q"']
\end{tikzcd}
\qquad
\left\{ \begin{array}{l} S = K \times (K \cdot C) \\ Q = S \cdot A^{\circ} \end{array} \right.$$

are still relations but, as the category changed,

$$Keyword \xleftarrow{Q} Author \;\; = S \cdot A^{\circ}$$

is such that

$$k \; Q \; a = \langle \textstyle\sum \; p \; : \; a \; A \; p : \; k \; S \; p \rangle \tag{10}$$

— it gives, for each author, her/his histogram of citations per keyword, within the same area.

Percentiles

Pushing further, $\mathbb{M}$ can be enriched so that $\times$ forms a **group**, bringing division in:

$$Keyword \xleftarrow{Z} Author \;=\; \frac{S \cdot A^\circ}{S \cdot \top} \tag{11}$$

This makes such histograms relative to the **grand total** of citations in each area (keyword) k:

$$k\ Z\ a = \frac{\langle \sum p\ :\ a\ A\ p :\ k\ S\ p \rangle}{\langle \sum q\ ::\ k\ S\ q \rangle}$$

That is, $k\ Z\ a$ gives the percentile of author a when evaluated (with)in keyword (area) k.

Example: $k\ Z\ a = n\ 10^{-5}$ means that n-many citations among 10^5 citations in area k are of papers by a.

h-index?

Z metric better than h-index because it takes into accout the cardinality of each community (cf. keywords).

h-index harder to encode (is there a "ranking" semiring?)

(Thinking about this — too many frustrating committees!)

More triangular patterns (metaphors)

In $\mathbb{R}$, another triangular pattern is

$$
\begin{tikzcd}
T \arrow[dr, "g"'] & & V \arrow[ll, "\frac{f}{g} = g^{\circ} \cdot f"'] \arrow[dl, "f"] \\
& A &
\end{tikzcd}
\qquad t \, \frac{f}{g} \, v \Leftrightarrow (g \; t) = (f \; v) \tag{12}
$$

— where f, g are functions — called a **metaphor**.

By (5), this has the **same** meaning in category $\mathbb{M}$.

Nice properties, recalling **rational numbers**, e.g.

$$
\frac{f}{id} = f \tag{13}
$$

$$
\left(\frac{f}{g}\right)^{\circ} = \frac{g}{f} \tag{14}
$$

and so on.

Rational matrices / relations

$\frac{f}{g}$	$a1$	$a2$	$a3$	$a4$	$a5$
b_1	0	0	1	0	1
b_2	1	0	0	0	0
$b3$	0	1	0	1	0
$b4$	0	1	0	1	0
$b5$	0	0	0	0	0

$$
f = \begin{matrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{matrix}
$$

$$
g = \begin{matrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{matrix}
$$

Most specifications are rational relations / matrices, eg.

$$
Sort = \frac{bag}{bag} \times \frac{true}{ordered} \qquad \left(= \frac{bag \triangledown true}{bag \triangledown ordered}\right)
$$

where $(f \triangledown g)\ a = (f\ \ a, g\ \ a)$.

Quantitative invariants

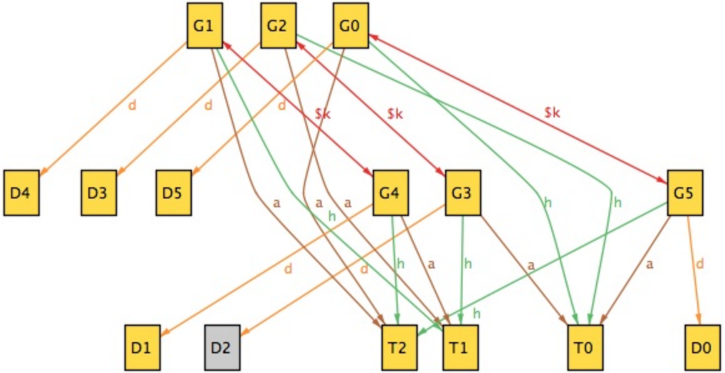
The teams (T) of a football league play games (G) at home (h) or away (a), and every game takes place in some date (d):

$$\begin{tikzcd} T & G \arrow[l, "h"'] \arrow[r, "a"] \arrow[d, "d"'] & T \\ & D & \end{tikzcd}$$

Invariantly,

- All teams play against each other exactly once but never against themselves.
- No team can play two games on the same date.

Three teams playing



Clearly, $k = \frac{a^\triangledown h}{h^\triangledown a}$ should be a bijection (cf. team swapping).

Quantitative invariants

All teams play again each other exactlty once but never against themselves, in $\mathbb{R}$:

$$h \cdot a^{\circ} = \top - id \tag{15}$$

meaning, for all teams t, t'

$$\langle \exists\, x \ :: \ t = h\, x \wedge t' = a\, x \rangle \Leftrightarrow t \neq t'$$

Exactly once? In $\mathbb{M}$ we write exactly the same as above,

$$h \cdot a^{\circ} = \top - id$$

capturing everything:

For all teams t, t',

$$\langle \sum x \ : \ t = h\, x \wedge t' = a\, x : \ 1 \rangle = \textbf{if } t = t' \textbf{ then } 0 \textbf{ else } 1$$

Quantitative invariants

No team can play two games on the same date, in $\mathbb{R}$:

$$d^{\circ} \cdot d \subseteq id \cup \neg(I^{\circ} \cdot I)$$

where $I = a \cup h$, $t \; I \; x$ meaning "team t is involved in game x".

That is, for all $x \not\equiv x'$,

$$\langle \exists \; t \; :: \; t \; I \; x \wedge t \; I \; x' \rangle \Rightarrow (d \; x) \not\equiv (d \; x')$$

Interestingly, in $\mathbb{M}$ this invariant is rendered much simpler,

$$d \cdot (a + h)^{\circ} \leqslant \top$$

cf.

$$\langle \forall \; y, t \; :: \; \langle \sum x \; : \; y = d \; x : \; t \; a \; x + t \; h \; x \rangle \rangle \leqslant 1$$

Quantitative invariants

Recall that $k = \frac{a \triangledown h}{h \triangledown a}$ should be a bijection.

Bijection = function (= deterministic + total) + injective + surjective

---

In $\mathbb{M}$:

$$! \cdot \frac{a \triangledown h}{h \triangledown a} = !$$

$$! \cdot \frac{h \triangledown a}{a \triangledown h} = !$$

It all has to do with **totals** — **counting** how many 1s the (Boolean) matrices have per column /row !

Wrapping up

Main idea:

"Scalable modeling": the "keep definition, change category" lemma.

In the previous TRUST workshop I played the same game with another category, that of **Markov chains**.

Questions:

- What is the best path towards **quantitative abstraction**?
- Some pointfree statements simpler if idempotency is removed
- What would it mean for Alloy to drop $+$-idempotency? (cf. SMT backend)

Annex

Annex

Pointwise details of (10):

$$
\begin{aligned}
k\ Q\ a &= \langle \sum\ p\ :\ a\ A\ p :\ k\ S\ p \rangle \\
&= \langle \sum\ p\ :\ a\ A\ p \wedge k\ K\ p :\ \langle \sum\ q\ :\ p\ C\ q \wedge k\ K\ q :\ 1 \rangle \rangle \\
&= \langle \sum\ p, q\ :\ a\ A\ p \wedge k\ K\ p \wedge p\ C\ q \wedge k\ K\ q :\ 1 \rangle
\end{aligned}
$$