Critical Abstraction: Generating Simplest Models for Causal Explanation

Brian C. Williams

Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto CA 94304 USA
net address: bwilliams@xerox.com

Abstract

Central to most scientific and engineering tasks — design, diagnosis, analogy and design rationale capture — are descriptions of a device that capture its salient features with respect to how it works. These descriptions are used as the central focus of many engineering reasoning techniques. Yet little has been said about what constitutes a best, or even adequate, description. Likewise, related questions, such as "what is an interesting qualitative landmark?" and "what constitutes a simplest model?" have been left unanswered. Each is an instance of the more general question: what constitutes a goold abstraction? While current reasoning systems are effective users of qualitative abstractions, future systems must also be able to create these abstractions.

We take the perspective that a device works by constructing a topology of interactions between quantities. To construct a parsimonious description of an *interaction topology* we introduce the concept of a *critical abstraction* — a most abstract description of a topology relative to a set of queries, that preserves the link between the individual mechanisms of a device, and the behaviors mentioned in the queries. We demonstrate critical abstraction for equational descriptions of interactions. We present a *generative modelling technique* for computing critical topologies using only the mathematical properties of the interaction topology representation. The use of causal connections during all phases of abstraction highlights the central link between causality and what is interesting.

1 Introduction

Capturing and reasoning about how devices work are central to most scientific and engineering reasoning. For example, when looking for innovative solutions the designer searches among devices that work in qualitatively different ways[19]. Given a faulty device, a diagnostician hypothesizes mechanisms that could account for qualitative differences between the artifacts observed and expected behavior[6]. And to help account for new phenomena a scientist searches for known mechanisms that work analogously[7]. A central focus of all these tasks are abstract accounts of how a device works.

Up to now our reasoning systems have been users of abstractions. Past research has concentrated on the effective use of particular qualitative representations for some task. A key direction for the future are systems that create new abstractions.

An account of how a device works can be represented in many ways; for example, as a system of qualitative equations, as a causal diagram or as a time-varying set of histories. Our long term goal is a unified theory of abstraction that builds on our experience with these different types of representations [16, 17, 18, 19]. The theory must provide a criteria for what constitutes a good abstraction and techniques for generating these abstractions for the various representations. In this paper we introduce one such criteria, called *critical abstraction*, and a generation technique for equational descriptions, *called generative modelling.*

A good abstraction highlights only features of interest, suppressing all superfluous detail. In our experience what is qualitatively and temporally interesting is inextricably tied to the concepts of causality and local interaction. Causality has taken a back seat in most current qualitative reasoning research. In this paper we show that causality plays a central role in all stages of the abstraction process.

The remainder of this introduction provides an overview of what constitutes an account of how a device works, and its overall impact on the abstraction process. Such an account establishes the role played by each of a device's internal mechanisms, with respect to achieving the behavior of interest. This link is crucial, for example, during diagnosis to pinpoint faulty mechanisms, or during design modification to identify the appropriate mechanisms to be changed.

Intuitively, for continuous systems a "device works" by establishing a network of local interactions between quantities, and modulating these interactions over time.[1] Each of these local interactions are produced by basic mechanisms such as processes, components and their interconnections. And the modulation of interactions produces such effects as raising and lowering signals, changing the operating modes of components, and, more generally, pushing the device between different regions of its state space. We refer to such a network as an *interaction topology* [19].

What description of an interaction topology is appropriate? A system of quantitative equations, such as those from physical system dynamics, would be overly detailed. To capture how a device works these quantitative interactions must be abstracted in a way that captures exactly those properties that are essential to achieving the behavior of interest. But in a way that assigns responsibility for each abstract interaction to a single mechanism. When all but the essential properties of the interactions are eliminated we say that the topology is *critically abstracted*, and we refer to the result as a *critical topology.*

A critical topology is by no means an absolute concept. What internal features of interactions are interesting depends on certain external features of interest for particular variables.

[1] The reason we use the term interaction, for example as opposed to equation, is to distinguish between the representation of an object and the object being represented. An interaction is an entity in the world (or abstraction) that we are trying to model or describe. An equation is one of many possible ways of describing an interaction.

and the causal connection between each interaction and these variables. These external features are determine by the external context of the device's use, the goals of the analysis.

For example, designers use a pullup-pulldown transistor pair as an inverter in digital design, and as an amplifier in analog design (figure 1). The features of interest for the device's input-output behavior and the resulting description of how this device works is very different in the two situations. As an inverter we are interested in how the input and output rise and fall between voltage intervals representing "1" and "0". To account for this behavior one considers the transistors' non-linear behavior in several different operating modes — cutoff, unsaturated and saturated — at a qualitative level. When the device is treated as an amplifier, one is interested in the input's and output's incremental variations around a quiescent point. To account for this behavior one considers only the transistors' linear behavior in the unsaturated mode, near the quiescent point. To evaluate the amplifier's gain the device's quantitative behavior must be considered.

The critical topology also depends on the granularity of the internal mechanisms. The account of the above circuit is very different if we decompose the transistor into conductive and capacitive components, as opposed to a single mechanism. To account for these context dependencies, our concept of a critical topology is defined relative to sets of questions being asked, mechanisms that produce local interactions, and independent variables.

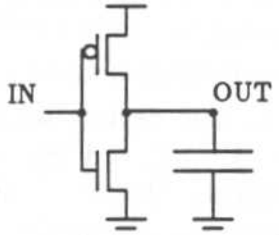

Figure 1: Is it an inverter or an amplifier? How a device is used substantially effects our perception of how it works.

The process of generating a critical topology depends on the interaction topology representation. As we mentioned earlier, there are many ways of representing how a device works, each capturing an interaction in a different way. Interaction may be causal or constraining, dynamic or static, intensional or extensional, and somewhere between qualitative and quantitative. Temporal constraint propagation[17], can be loosely viewed as a technique for generating critical abstractions when the interactions are described by concise histories - that is each interaction is causal, extensional and dynamic. In this case the abstraction process is one of generating a qualitative, causal explanation.

This paper focuses on topologies of static interactions that are described intensionally through a system of (Q1) equations[18]. In this case the process of abstraction is one of model simplification. More specifically this paper reports progress on a technique that automatically generates a critical topology relative to a behavior of interest, given a quantitative, equational description of a device's interactions. The technique is one of the first examples of what we call *generative modeling*. That is, it generates a simpler model of an artifact without prior experience, purely from the mathematical properties of the interaction representation. The simpler, qualitative models preserve the features of interactions necessary to determine the behaviors of interest, and the causal connection between interesting internal mechanisms and these behaviors. Thus the generated models can be used by a qualitative explanation or simulation system.

The work presented is a key element in our program to develop a theory of interaction topologies, as they are used in scientific and engineering problem solving. Elsewhere we report progress on developing appropriate representations for capturing topologies, an axiomatization of this representation, and techniques that use these topologies as the central focus of analysis, diagnosis and design[16, 17, 18, 21, 19, 20].

In the next section we examine several additional uses for critical abstraction and the link to related work. We then elaborate the concept and the abstraction technique in the remaining paper.

2 The Pervasiveness of Critical Abstraction

The endeavor of representing and generating critical topologies addresses issues central to qualitative reasoning, teleological reasoning, and modeling. The use of qualitative abstractions has been an important element of each of these. We argue that the automatic generation of these abstractions is also central, and that critical abstraction is an appropriate starting point to automating this process in each of these cases.

Capturing causal accounts of how devices work has always been a primary concern of qualitative reasoning[3, 5, 9, 16, 17, 15], and the representation we use here for interaction topologies (a variant of Q1 equations [18]) evolved from these ideas. An open problem in qualitative reasoning is that the traditional qualitative representations over abstract [12, 13] — they introduce unintended behaviors that are physically unrealizable. As proposed in [18], this problem is partially solved by hybridizing qualitative and quantitative algebras. This allows interactions to be described at many levels of abstraction, lying between traditional quantitative and qualitative algebras. But the problem remains of automatically selecting the appropriate abstraction in which to describe each interaction — the task addressed here.

A second open issue is: where do qualitative representations come from? Said differently, what are the interesting landmarks of a quantity space[9], and how are they generated? Typically the interesting landmarks for each mechanism type fall directly out of the qualitative process description or component model. Additionally the sign of all quantities and their derivatives are automatically considered interesting[5, 9, 16]. An interesting example of the automatic generation of landmarks is QSIM[12]. QSIM introduces as landmarks any critical points and inflection points that come up during the simulation process. We take a different approach. We propose instead that the concept of "interesting" is relative to the questions being asked. As we see later, the interesting landmarks fall out nicely as a byproduct of constructing a critical topology relative to these questions. While these landmarks may happen to be critical points or inflection points (as in QSIM), these properties are neither necessary or sufficient.

Another important area is teleological reasoning. Key elements of a teleological account are the way individual mechanisms of a device work together to achieve the device's intended behavior, and the role played by each component. This is captured by an explanation of how the device correctly works. Early research identified the basic elements for representing and constructing these explanations — causality, quality and time [3, 5, 9, 16]. Although explanation and teleology has taken a back seat to event-driven simulation, recently some promising research has been devoted to identifying appropriate causal and teleological representations[4, 10, 17]. This paper differs from earlier work in that it introduces a precise definition of a good account — critical abstraction — and proposes a technique for automatically generating these accounts.

Finally, when applied to systems of equations, critical abstraction can be viewed as a model simplification technique. Roughly speaking, recent work in modeling[1, 8, 14, 2] is concerned with selecting or composing models that support efficient simulation. Our motivation is different - constructing parsimonious descriptions versus computational efficiency. But the work shares much in common. As in reference [8], critical abstraction is focused relative to a set of questions, for example, as opposed to global consistency[1, 14]. As in reference [2], critical abstraction generates models from the algebraic properties of interactions, for example, as opposed to constructing abstractions solely by composing and selecting from libraries of models or model fragments [1, 8, 14]. We refer to this as *generative modeling.* Furthermore, the goals of parsimonious description and computational efficiency are strongly linked. A rationale for qualitative analysis is that computation can be faster with simpler, qualitative descriptions than with their corresponding quantitative ones.

3 An Example

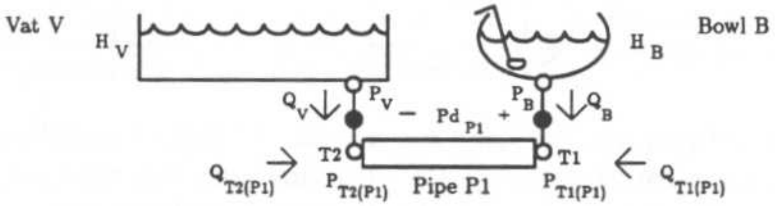

Figure 2: A regulator consisting of a bowl B whose fluid height is regulated, a vat V that supplies fluid, and a pipe P1 that performs the regulation.

Consider the familiar regulation example[18, 20] of figure 2. As punch is removed from bowl B, it is replaced by fluid from vat V, via the pipe $P1$. The quantitative equations (called the *base topology*) describing the interactions produced by these three components and their connections are listed below, and shown diagrammatically in figure 3, left side: [2]

[2] BIn stands for base interaction number n, and CIn stands for critical interaction n.

Bowl Interactions:

$$\begin{array}{lrcl}
\text{BI1)} & V_B & = & H_B \times A_B \\
\text{BI2)} & M_B & = & d \times V_B \\
\text{BI3)} & F_B & = & M_B \times g \\
\text{BI4)} & F_B & = & P_B \times A_B \\
\text{BI5)} & Q_B & = & dV_B/dt \\
\text{BI6)} & dV_B/dt & = & dH_B/dt \times A_B
\end{array}$$

Bowl/Pipe connection:

$$\begin{array}{lrcl}
\text{BI7)} & Q_B & = & -Q_{T1(P1)} \\
\text{BI8)} & P_B & = & P_{T1(P1)}
\end{array}$$

Pipe interactions:

$$\begin{array}{lrcl}
\text{BI17)} & Pd_{P1} & = & P_{T1(P1)} - P_{T2(P1)} \\
\text{BI18)} & Pd_{P1} & = & Q_{T1(P1)} \times R_{P1}
\end{array}$$

Vat Interactions:

$$\begin{array}{lrcl}
\text{BI9)} & V_V & = & H_V \times A_V \\
\text{BI10)} & M_V & = & d \times V_V \\
\text{BI11)} & F_V & = & M_V \times g \\
\text{BI12)} & F_V & = & P_V \times A_V \\
\text{BI13)} & Q_V & = & dV_V/dt \\
\text{BI14)} & dV_V/dt & = & dH_V/dt \times A_V
\end{array}$$

Vat/Pipe connection:

$$\begin{array}{lrcl}
\text{BI15)} & Q_V & = & -Q_{T2(P1)} \\
\text{BI16)} & P_V & = & P_{T2(P1)}
\end{array}$$

$$\begin{array}{lrcl}
\text{BI19)} & Q_{T1(P1)} & = & -Q_{T2(P1)} \\
\text{BI20)} & R_{P1} & > & 0
\end{array}$$

where V denotes fluid volume, H fluid height, A cross-sectional area, M mass, d fluid density, F downward force, g gravitational acceleration, P pressure, Q fluid flow, R fluid resistance.

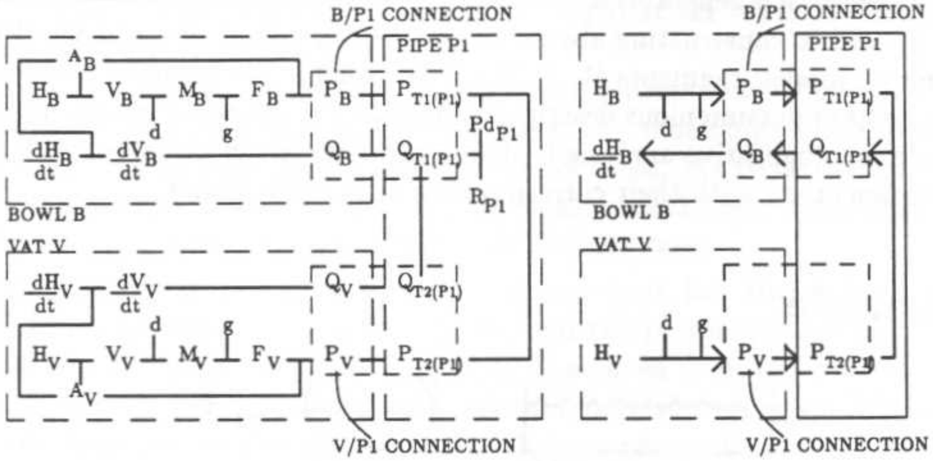

Figure 3: Base topology (left) and its critical abstraction (right) relative to the question "$[dH_B/dt]$?" Each multi-edge denotes an interaction. For example, the edge in the upper left corner, going between H_B, A_B and V_B, corresponds to interaction BI1. A dotted line encircling a set of edges denotes a set of interactions produced by a single mechanism (e.g., component or connection).

Suppose we ask what direction the bowl's height changes (i.e., $[dH_B/dt]$),[3] given the heights H_B and H_V. Equations describing the device's critical interaction topology relative to this question are listed below, and shown diagramatically on the right side of figure 3:[4]

[3] $[e]$ denotes the sign — positive, negative or zero — of real expression e.

[4] $f(x_1 \ldots x_n) \Rightarrow y$ is a *causal interaction* supporting y. $\Rightarrow$ is treated otherwise as an equivalence relation.

Bowl Interactions:
CI1) $H_B \times d \times g \;\Rightarrow\; P_B$
CI2) $[Q_B/dt] \;\Rightarrow\; [dH_B/dt]$
Bowl/Pipe connection:
CI3) $[Q_B] \;\Leftarrow\; \ominus[Q_{T1(P1)}]$
CI4) $P_B \;\Rightarrow\; P_{T1(P1)}$

Vat Interactions:
CI5) $H_V \times d \times g \;\Rightarrow\; P_V$
Vat/Pipe connection:
CI6) $P_V \;\Rightarrow\; P_{T2(P1)}$

Pipe interactions:
CI7) $[Q_{T1(P1)}] \;\Leftarrow\; [P_{T1(P1)} - P_{T2(P1)}]$

This corresponds to the following teleological account:

> *Why does the bowl height change in the direction of the bat/bowl height difference?* The bowl changes its height in the direction of flow into its bottom (CI2), which because of the bottom connection is the same as pipe flow (CI3). The pipe guides its flow in the direction of its pressure drop (CI7), which, because of the connections, is the pressure drop between the vat and bowl (CI4,CI5). Given uniform density and gravity, the vat and bowl guide this in the direction of height difference (CI1,CI5).

Consider the difference between the base and critical topologies. First, a causal order has been imposed, and several superfluous interactions and variables have been eliminated — dH_V/dt, dV_V/dt, $Q_{T2(P1)}$, BI13, BI14 and BI19 — which do not constrain the value of $[dH_B/dt]$. Second, several interactions have been collapsed together. For example, the pipe interactions BI17, BI18 and BI20 are combined into CI7. However, each interaction still corresponds to a single mechanism (e.g., component or connection) making it possible to uniquely assign responsibility for any individual piece of behavior. Finally, many (although not all) of the resulting interactions have been weakened. Some are purely qualitative, relating the signs of quantities; for example, CI2 and CI3. Others are hybrid, using a combination of real values and signs; for example, CI7. And some remain quantitative, such as CI1, CI4, CI5 and CI6.

What falls out naturally from the critical topology is a qualitative representation — the set of interesting landmarks — that is appropriate for answering the query. Moving backward through this topology from dH_B/dt to H_v and H_b, we see that to know the direction of dH_B/dt it is sufficient to know only Q_B and $Q_{T1(P1)}$'s sign. Determining this last sign requires knowing the relative ordering of P_B and P_V. This requires knowing the quantitative relationship between P_V, P_B, H_V, H_B, d and g.[5] From this description we see that the

[5] More precisely, using the quantitative relationshps:

$$\begin{aligned} P_B &= H_B \times d \times g \\ P_V &= H_V \times d \times g \end{aligned}$$

we infer that:

$$\begin{aligned} [P_B - P_V] &= [H_B \times d \times g - H_V \times d \times g] \\ \Rightarrow [P_B - P_V] &= [(H_B - H_V) \times d \times g] \\ \Rightarrow [P_B - P_V] &= [(H_B - H_V)] \otimes [d] \otimes [g] \end{aligned}$$

landmarks of interest for the critical topology are $dH_B/dt = 0$, $Q_B/dt = 0$, $Q_{T1(P1)}/dt = 0$, and $P_B = P_V$ (and the real values of H_B, H_V, d and g).

Above we saw that the critical topology eliminates superfluous interactions, aggregates relevant interactions shared by a common mechanism, and abstracts these aggregate interactions. These three elements reflect the three basic steps of our abstraction process. Each throws away some type of extraneous information. The first throws away unused interactions. The second throws away variables and structure internal to each mechanism. The third abstracts the remaining interactions, and the values of the remaining variables.

The key observation underlying our approach is that *every element of the abstraction process is determined by the causal connection of every internal variable and interaction to the primary behavior of interest.* Thus our approach begins by establishing these causal connections, and then uses them to guide the three stages of the abstraction process.

The remaining sections of this paper are devoted to making more precise the concept of critical topologies and the three stages of the abstraction process. Note that critical abstraction is an ideal, to the degree that we have not proven whether our approach always constructs abstractions that are critical.

4 Problem Statement and Definitions

The task is to construct the critical abstraction of an interaction topology, given a query and a set of mechanisms of interest. The resulting topology makes explicit the role played by each mechanism — exactly those features of a mechanism's interactions, that contribute to the overall device's behavior of interest. Next we consider definitions for interaction, interaction topology, mechanism, query, and critical abstraction.

In general an interaction is a horn clause whose antecedents and consequents are equations in the hybrid algebra Q1 [18], such as:

$$H_{valve} > H_{closed} \rightarrow [Pd_{valve}] = [Q_{T1(valve)}]$$

This is a (dynamic) interaction between a valve's pressure and flow, which is active only when the valve is open. A Q1 equation is composed of =, the standard operators $(+, -, /, \times)$ on the reals $\Re$, the corresponding operators $(\oplus, \ominus, \oslash, \otimes)$[6] on signs $S' \equiv \{[+], [0], [-], [?]\}$, [ ] which returns the sign of a real, constants in $\Re$ and S', and variables ranging over $\Re$. Inequalities are treated as abbreviations of Q1 equations; for example, $a > b \equiv [a - b] = [+]$.

An interaction topology T consists of five parts:

In the second line the quantitative relationships were necessary to put H_B and H_V over a common denominator. for example the following qualitative relations would not have been sufficient:

$$\begin{aligned} [P_B] &= [H_B] \otimes [d] \otimes [g] \\ [P_V] &= [H_V] \otimes [d] \otimes [g] \end{aligned}$$

[6] Roughly speaking, $[x] \otimes [y]$ answers the question "what is the sign of $x \times y$, given only the signs of x and y?" $\oplus, \ominus$ and $\oslash$ have similar interpretations.

- a set of variables V each is mentioned in at least one interaction,
- a set of independent variables IV $\subset$ V,
- a set of interactions I on the variables of V,
- a set of mechanisms M (e.g., components, connections, processes), and
- an onto function IM: $I \rightarrow M$, which associates each interaction with the mechanism that produces it.

For example, for the base topology in the previous section, $V = \{ d, g, V_B, H_B, A_B, M_B, F_B, P_B, Q_B, dV_B/dt, dH_B/dt, V_V, H_V, A_V, M_V, F_V, P_V, Q_V, dV_V/dt, dH_V/dt, Pd_{P1}, Q_{T1(P1}, Q_{T2(P1)}, R_{P1} \}$,[7] IV = $\{H_B, H_V, d, g, R_{P1}, A_B, A_V\}$, and I = $\{BI1 \ldots BI20 \}$. The mechanisms are the vat, bowl, pipe, and their interconnections; that is, M = {B, V, P1, connection(B,P1), connection(V,P1)}. IM maps each of { BI1, ... BI6 } to B, { BI7, BI8 } to connection(B, P1), { BI9, ... BI14 } to P1, { BI5, BI16 } to connection(V, P1), and { BI17, ... BI20 } to P1. The representation of mechanisms is intentionally minimalist – a set of interactions — providing the flexibility of describing mechanisms using lumped elements, processes, kinematic pairs or some other ontology.

Intuitively a question is of the form "is its qualitative value x?" or "what is its qualitative value?", where a qualitative value is defined to be a region of state space[16]. Determining a qualitative value reduces to determining a device's position in state space relative to a set of boundaries (for example, inequalities with respect to some landmarks). This corresponds to determining the values of a set of Q1 expressions. For example, suppose the bowl has the qualitative value "partially full" if $H_B > 0$ and $HT_B > H_B$, where H_B is B's fluid level, and HT_B is its top. This is equivalent to the conjunction $[H_B] = [+]$ and $[HT_B - H_B] = [-]$; thus to answer the question, we are interested in the values of the expressions $[H_B]$ and $[HT_B - H_B]$.

More precisely, given a topology T, a *query* Q is a set of Q1 expressions over the variables V. In the above example Q = $\{[HT_B - H_B],[H_B]\}$. We say that *T answers Q* if the value of each expression in Q is uniquely determined by T and the values of the independent variables.

All of this is a precursor to defining the critical abstraction of a topology (or critical topology for short). Given a question Q about a topology BT, called the *base topology*, the *critical abstraction of BT relative to Q* is the greatest abstraction of BT (called the *critical topology*) that answers Q. That is, no strict abstraction[8] of the critical topology exists which answers Q.

Finally, given topologies T1 and T2, *T2 is an abstraction of T1* if:

1. variables V(T2)[9] $\subset$ V(T1),

[7] Note that the relationship between a quantity and its derivative is not exploited - X and dX/dt are simply treated as two distinct variables.

[8] Strict abstraction has the obvious definition. i is a *strict abstraction* of j if i is an abstraction of j, but j is not an abstraction of i.

[9] V(T2) denotes the variables of topology T2.

2. independent variables IV(T2) $\subset$ IV(T1),
3. mechanisms M(T2) $\subset$ M(T1), and
4. for every mechanism in M(T1), its interactions in T1 entail its interactions in T2. That is, for each i in I(T2) and its corresponding mechanism m = IM(T2)(i), i is satisfied whenever m's base interactions MI(T1)(m) are satisfied.

Most of this is the standard definition for one set of relations being weaker than another set. The key addition is thϵ constraint that the abstraction preserve the connection between interactions and individual mechanisms that produce them. This connection, together with "weakest abstraction" we take to be the essential ingredients for capturing a mechanism's role — that is, they tell us exactly those features that each mechanism contributes to a behavior of interest.

5 Abstracting Static Topologies

Consider the abstraction process. Due to space constraints we make several simplifications. First, we restrict ourselves to topologies of static interactions — interactions with no antecedents. This corresponds to modelling devices as having a single operating region. Second, the interactions and query are restricted to a subalgebra of Q1, called Q2, which is missing sign addition and subtraction ($\oplus$, binary $\ominus$), but preserves sign times and unary minus. The motivation is that Q2 is sufficient to describe qualitative representations consisting of open regions separated by boundaries. The added complexity of introducing binary $\oplus$ and $\ominus$ seems unnecessary. Third, the base interactions are all quantitative equations. They are composed only of = and the operators on $\Re$ ($+, -, \times$, and $/$). Fourth, the equations in the base topology are presumed to be irredundant. The motivation is that standard techniques can be used to remove irredundant quantitative equations. We avoid repeating these techniques here. And finally they do not contain any simultaneities. The purpose of this paper is not to develop new techniques for inferring causal ordering, but to demonstrate the central importance of this causal ordering during all phases of the abstraction process. Eliminating simultaneities makes it easy to determine causality, allowing us to focus the presentation on the traditionally less understood phases of the abstraction process. Additionally, for systems with simultananeities we intend to use a technique under development, called *causal dominance*, to break the simultaneities at their weakest links. This reduces a device description to a simultaneity free system of equations, where the generative modelling technique presented here can be directly applied. Causal dominance is in the spirit of feedback analysis [16], and will be presented in a separate report.

5.1 Causal Ordering and Removing Superfluous Interactions

Given a base topology BI and query Q, we begin by identifying a minimal subset of its interactions I that, together with its independent variables IV, determine each variable in the query expressions. To accomplish this we first determine the "causal" relationships imposed by the interactions on the dependence of the variables (arrows in figure 4). We then remove all but those interactions causally relating the query variables to the independent variables (crossed out interactions).

Computing the causal relations is straightforward for the restricted case considered here, and produces a result roughly analogous to any of causal reasoning techniques presented in [5, 16, 11]. First, by definition the base topology must answer the query; thus, there exists a subset of its interactions that determine each query variable. Second, the topology contains no simultaneities; thus, for each interaction i in this set, one of its variables v is determined by the remaining variables rv, and each of the remaining variables are determined by the independent variables, IV, independent of v. In this case we say that *v is caused by rv through i*, and refer to v as *i's effect*, rv as *i's causes*, and i as v's *causal support.* Finally, the interactions are irredundant; thus, a determined variable is the effect of exactly one interaction. We define the *causal support structure* of a determined variable to be its causal support, together with the support structure of the variable's causes. The support structure of each independent variable is the empty set.

We compute the support of each dependent variable by sweeping out locally through the interactions from the independent variables.[10] The relevant interactions are those in the support structure of each query variable.

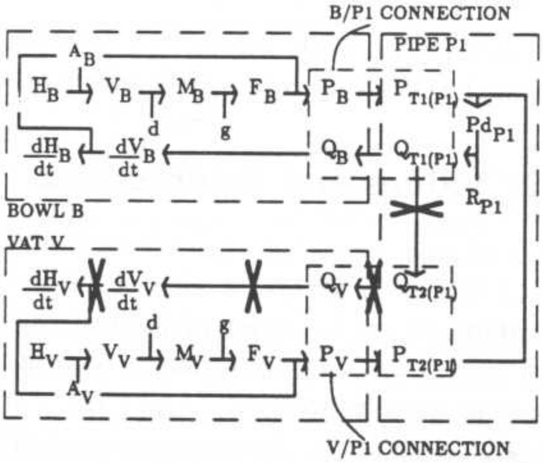

Figure 4: Interactions causally supporting query $[dH_B/dt]$. Arrows indicate causality. X's indicate superfluous interactions.

[10] More precisely, let I initially be all the interactions containing an independent variable. For each i in I, if exactly one of its variables v is unsupported: first, support v with I, specifying the remaining variables as v's cause, and second, add to I all interactions, except i, that mention v. This is analogous to constructing well founded support in a JTMS.

5.2 Eliminating Structure Internal to Each Mechanism

To establish the role of each mechanism in a device's behavior, we've required that each interaction remain attributable to a single mechanism. Nevertheless, the current description often contains structure internal to a mechanism, and is thus superfluous. More specifically, mechanisms can be viewed as communicating internally and externally through shared variables. The *external variables* are the independent variables, variables in query expressions, and any variable shared by interactions of distinct mechanisms. The remaining variables are *internal*; they are used only to communicate between interactions within a single mechanism. For example, consider the bowl interactions, circled in the upper left corner of figure 4. The external variables are $H_B, dH_B/dtP_B, Q_B, A_B, d$ and g, and the internal variables are V_B, M_B and F_B.

The internal communications are abstracted (eliminating the internal variables) by composing the interactions of each mechanism that lie between its external variables. For example, starting with external variable P_B and its support:

$$F_B/A_B \Rightarrow P_B$$

we successively substitute each internal variable (in this case F_B) with its support and simplify, until only external variables remain:

$$H_B \times d \times g \Rightarrow P_B$$

This composition is performed for each *output*.[11] In our example the outputs are dH_B/dt, $Q_{T1(P1)}$, P_V and P_B, and the resulting topology is shown in figure 3, right side.

5.3 Abstracting Values and Interactions

The interactions just produced are sufficient to determine the quantitative value of all remaining variables. The last step abstracts all but those features of these variables and interactions that contribute to the behavior of interest (as specified by the query). These features may be a sign (e.g., $[dH_B/dt]$) an ordering (e.g., expressed as $[P_V - P_B]$), or more generally the sign of a composite expression, (e.g., $[(H_V - H_B)/(d \times g)]$).

The algebra Q1 has a number of interesting properties that can be used to determine the relevant features of an expression's variables, necessary to determine the expression's value. Most important are a set of homomorphisms for the sign operator [ ]. Specifically, [ ] is homomorphic with respect to multiplication, division and negation, although not with respect to addition and subtraction (where the sign operators are strictly weaker):[12]

$$\begin{array}{rcl@{\qquad}rcl} [a \times b] & = & [a] \otimes [b] & [a + b] & \subset & [a] \oplus [b] \\ [a/b] & = & [a] \oslash [b] & [\imath - b] & \subset & [a] \ominus [b] \\ [-a] & = & \ominus [a] & & & \end{array}$$

Thus, given the pipe interaction:

$$Pd_{P1} \Leftarrow R_{P1} \times Q_{T1(P1)}$$

[11] An output of mechanism m is an external variable, supported by an interaction in m. An input is one mentioned in the support of one of m's interactions.

[12] Weaker under subset, where each sign denotes subsets of $\Re$.

If we are interested in the sign of Pd_{P1}, it can be determined knowing only the signs of R_{P1} and $Q_{T1(P1)}$:

$$[Pd_{P1}] \Leftarrow [R_{P1}] \otimes [Q_{T1(P1)}]$$

Identifying the relevant features of values is critical, since often they are known, while the precise value of the variables are unknown. For example, a pipe's resistance is always positive ($[R_{P1}] = [+]$). Thus using the fact that $[+]$ is a multiplicative identity ($[+] \otimes s = s$) the pipe interaction becomes:

$$[Pd_{P1}] \Leftarrow [Q_{T1(P1)}]$$

Stepping back, consider how we weaken the complete set of interactions in a topology. Conceptually, given a query $[q]$, the current set of interactions perform a series of quantitative operations, starting from the independent variables. The quantitative operators in q are then applied, and the abstraction operator, $[\,]$, is applied only at the very end (figure 5,upper).

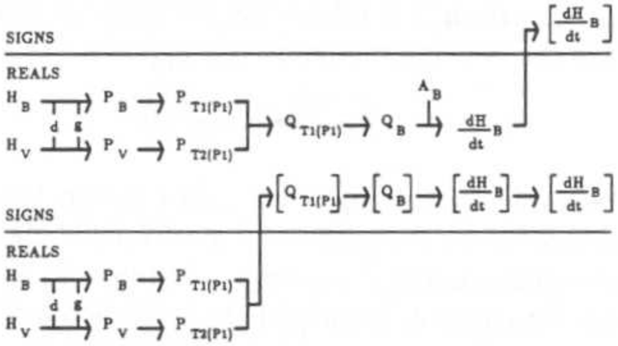

Figure 5: The sequence of operations performed by the topology before (upper) and after (lower) abstraction.

To determine the relevant features of each input and output, we use the homomorphisms to push the abstraction operation as early in the causal structure as possible (figure 5, lower). More precisely, given a query expression $[q]$, we begin by pushing $[\,]$ inward to the smallest "reachable" subexpressions of q. We observed earlier that multiplication, division and negation are homomorphic, but not addition and subtraction; thus, these smallest subexpressions are q's prime factors, f1 ... fn, and [q] becomes $[f1] \otimes \ldots [fn]$. This computation is equivalent to canonicalization in Minima, a symbolic algebra system for Q1, and is described in more detail in [18]. In our example the query expression $[dH_B/dt]$ is already prime.

Once the query is decomposed into prime factors, we begin to abstract each mechanism's interactions, preserving only those properties necessary to determine the sign of the query's factors. The abstraction performed depends on whether the variables contained by each factor f are part of the same or distinct mechanisms.

First, if f's variables are the effects of distinct mechanisms, then, since we cannot aggregate the interactions of distinct mechanisms, we have no choice but to determine the quantitative value of each of the factor's variables. The interactions necessary to accomplish this are those in each of the variable's causal support structure.

Second, if f's variables are the effects of a single mechanism m, then a relevant feature of m is the relation expressed by the factor. For example, the factor $[dH_B/dt]$ tells us that a relevant feature is the sign of H_B's derivative (i.e., the landmark $dH_B/dt = 0$). Later, the factor $[P_V - P_B]$ tells us that a relevant feature is the relative ordering of P_V and P_B (i.e., the landmark $P_V = P_B$). We then construct from m's interactions an interaction that is sufficient to determine f. First we substitute for each *output* variable, v, in f, using the interaction in m which causally supports v. We then canonicalize the result as described above, producing a new expression, e, over m's *input* variables. For example the causal support for the factor $[dH_B/dt]$ is:

$$Q_B/A_B \Rightarrow dH_B/dt$$

substituting into the factor and simplifying produces $[Q_B] \oslash [A_B]$.

Next, just as f told us what is interesting about m's outputs, the factors of e tell us what is interesting about m's inputs. In our example they are the signs of A_B and Q_B. Given that a bowl's area is always positive, then e simplifies to $[Q_B]$. Furthermore, together e and f provide an abstract interaction that relates the relevant features of m's inputs and outputs $e \Rightarrow f$. In our example the new interaction is:

$$[Q_B] \Rightarrow [dH_B/dt]$$

This process is repeated by moving backwards through the mechanisms that causally precede m. The inputs of m are the outputs of mechanisms that immediately precede m in the query's causal support structure. The relevant features of m's inputs, which are described by e's factors, are also the relevant features of the corresponding outputs. Thus, to abstract the rest of the query's causal structure pertaining to f, we repeat the above process on e's factors, terminating on the independent variables.

In our example, the new factor is $[Q_B]$ which is the output of:

$$Q_{T1(P1)} \Rightarrow -Q_B$$

for the mechanism "connection(B,P1)." The new interaction produced is:

$$\ominus[Q_{T1(P1)}] \Rightarrow [Q_B]$$

The next iteration of this process produces:

$$[P_{T1(P1)} - P_{T2(P1)}] \Rightarrow [Q_{T1(P1)}]$$

At this step the two variables of this factor, $P_{T1(P1)}$ and $P_{T2(P1)}$, are outputs of distinct mechanisms, thus, the remainder of the topology remains quantitative. The resulting topology is the critical topology given at the beginning of this paper (figure 3, right side).

The generative modelling process has been implemented in Lisp on top of Minima[18]. It has been tested on a variety of examples, including the one just described.

6 Discussion

We introduced the concept of interaction topology and its critical abstraction, and have argued that this abstraction captures descriptions of how devices work. We argued that this abstraction is determined by the behavior of interest, the granularity of the internal mechanisms considered, and the causal relationships imposed by the independent variables. Finally, we described a generative modelling technique that constructs critical abstractions, based on this context and the mathematical properties of Q1. As argued earlier this technique represents progress with respect to several open representational issues in qualitative reasoning: identifying landmarks that are argued to be interesting based on the query being posed, avoiding the problems of over abstraction in traditional qualitative approaches, modelling from first principles, and automatically constructing teleological descriptions.

A large portion of the research in qualitative reasoning has ignored issues of causality, while others of us believe in the central importance of causality, but have placed it on a back burner. The Importance of the representational concerns mentioned above, the ability of critical abstraction to capture these concerns, and the fundamental role of causality in the generation of critical abstraction convinces us that future progress in qualitative reasoning hinges upon restoring causality to first class status.

7 Acknowledgements

I would like to thank Sanjaya Addanki, Leo Joskowicz, Sesh Murthy and Dan Weld for some very exciting discussions, and my reviewers for several insightful comments.

References

[1] S. Addanki, R. Cremonini, and J. S. Penberthy. Reasoning about Assumptions in Graphs of Models. In *IJCAI-89*, Detroit, MI, August 1989.

[2] B. Falkenhainer and M. Shirley. Explicit Reasoning About Accuracy for Approximating Physical Systems. Automatic Generation of Abstractions and Approximations Workshop, AAAI,, 1990.

[3] C. Rieger and M. Grinberg. Representation and Procedural Simulation of Causality in Physical Mechanisms. In *IJCAI*, Camb., MA, Aug. 1977.

[4] Chandrasekaran, J. Josephson, and A. Keuneke. Functional reprsentation as a basis for explanation generation. (*IEEE conf. on Systems, Man and Cybernetics*), 1986.

[5] J. de Kleer and J. Brown. A Qualitative Physics Based on Confluences. *Artificial Intelligence*, 24, Dec. 1984.

[6] J. de Kleer and B. C. Williams. Diagnosing Multiple Faults. *Artificial Intelligence*, 32, April 1987.

[7] B. Falkenhainer. Learning from Physical Analogies: A Study in Analogy and the Explanation Process. Phd thesis,, U. Ill., December 1988.

[8] B. Falkenhainer and K. Forbus. Setting up Large-Scale Qualitative Models. In *AAAI-88*, pages 301–306, St. Paul, MN, August 1988.

[9] K. Forbus. Qualitative Process Theory. *Artificial Intelligence*, 24, Dec. 1984.

[10] D. Franke. Representing and Qcquiring Teleological Descriptions. (*AAAI Model-based Reasoning Workshop*), 1989.

[11] Y. Iwasaki and H. A. Simon. Causality in Device Behavior. *Artificial Intelligence*, 29, July 1986.

[12] B. Kuipers. Qualitative Simulation. *Artificial Intelligence*, 29, Sep. 1986.

[13] P. Struss. Mathematical Aspects of Qualitative Reasoning. *Artificial Intelligence in Engineering*, 3(3), 1988.

[14] D. Weld. Automated Model Switching: Discrepancy Driven Selection of Approximation Reformulations. Technical Report 89-08-01, Department of Computer Science and Engineering, University of Washington, 1989.

[15] D. S. Weld. Theories of Comparative Analysis. PhD Thesis TR 1035, MIT Artifical Intelligence Lab, May 1988.

[16] B. C. Williams. Qualitative Analysis of MOS Circuits. *Artificial Intelligence*, Dec. 1984.

[17] B. C. Williams. Doing Time: Putting Qualitative Reasoning on Firmer Ground. In *AAAI*, August 1986.

[18] B. C. Williams. MINIMA: A Symbolic Approach to Qualitative Reasoning. In *AAAI*, August 1988.

[19] B. C. Williams. Invention From First Principles via Topologies of Interaction. Phd thesis, MIT Artifical Intelligence Lab, May 1989.

[20] B. C. Williams. Interaction-based Invention: Designing Novel Devices from First Principles. In *AAAI*, July 1990.

[21] B. C. Williams. Minima: Integrating Qualitative and Quantitative Algebraic Reasoning. (*to appear in Artificial Intelligence*), 1991.