

Indicator frameworks

Joshua Tan¹, Christine Kendrick², Abhishek Dubey³, and Sokwoo Rhee⁴

¹Department of Computer Science, University of Oxford

²Bureau of Planning and Sustainability, City of Portland

³Department of Electrical Engineering and Computer Science, Vanderbilt University

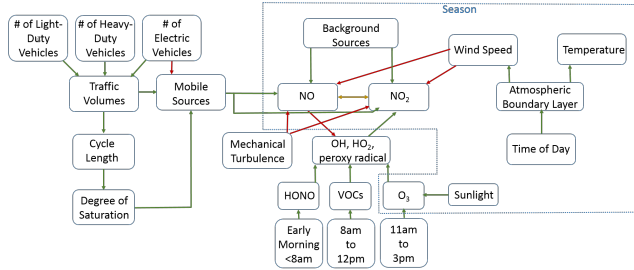
⁴National Institute for Standards and Technology

ABSTRACT

We develop a diagrammatic tool for constructing correlations between random variables, called an abstract indicator framework. Abstract indicator frameworks are modeled off operational (key performance) indicator frameworks as they are used in city planning and project governance, and give a rigorous, statistically-motivated process for constructing operational indicator frameworks.

1. INTRODUCTION

We take as our starting point a diagrams of simple correlations, as below:



This diagram correlates the measurement variables of an air pollution monitoring system with other, partially-observable variables like traffic composition, mechanical turbulence, and the presence of sunlight. It presents an intuitive and apparently useful description of a system at large. We would like to clarify the meaning of this diagram, and of others like it, by giving its interconnections a precise mathematical meaning. Clarifying the meaning of the diagram will not only make it more useful; it will allow us to connect this local, correlation-based picture of a system with other local pictures, as well as with more sophisticated scientific models of the world.

In this paper, we develop *abstract indicator frameworks*,

a diagrammatic tool for constructing causally-linked sets of random variables and their correlations. Abstract indicator frameworks are modeled off operational (key performance) indicator frameworks, especially as they are used in city planning and project governance. Such operational indicator frameworks have three main uses: (1) to communicate quantitative information and strategic priorities to a wide audience, (2) to enable policy reactions to data, especially in the optimization of processes, and (3) to restrict attention to a set of ‘relevant’ indicators—thus discarding the information from many other, ‘non-relevant’ indicators.

In city planning, there are several strategy-setting frameworks for constructing operational indicator frameworks, from balanced scorecards [3] to SMART [2] to more specialized urban planning frameworks; in such frameworks, the indicators are often designed by mayors, chief strategy officers, and sizable expert committees in tandem with new projects, new policies, and new processes. Even assuming that the participants adhere to a framework, the process of choosing indicators is often ad hoc, the results do not account for statistical relationships between the indicators, and the generated data is hard to translate across localities.

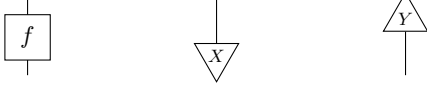
We propose an alternative. Instead of constructing operational indicator frameworks expensively and internally, meaning indicator-by-indicator, we can specify them abstractly and externally, by means of their causal and statistical relationships to other, already extant sets of indicators. Our approach is especially suited to the new ‘smart’ projects and smart cities where a primary challenge is to relate and integrate many large, heterogeneous data sets.

For example, cities are often interested in understanding the second-order impacts of specific projects, e.g. the impacts on health, crime, and jobs of a smart shuttle system. Assuming the existence of a local indicator framework for the shuttle system, and the existence of a top-level indicator framework representing broad priorities such as health, crime, jobs, and so on, then we can construct a mediating indicator set whose indicators satisfy certain statistical and causal relationships generated by the given indicator frameworks; these mediating indicators represent the second-order impacts of the local project to the city’s other priorities.

This motivates the definition of abstract indicator frameworks, which we define as the objects of a certain “category of diagrams of random variables”, *Ind*. We will apply *category theory*, originally developed to relate and analyze topological spaces, as an efficient language for relating and analyzing the causal and statistical aspects of indicator frame-

works.

Let \mathcal{X} stand for an indicator set. Abstract indicator frameworks have a notion of *process* that transforms one indicator set into another, a notion of *state* that represents the process of picking a specific indicator in \mathcal{X} , and a notion of *effect* that represents the process of “measuring” or computing the correlation with respect to a specific indicator in \mathcal{X} . Processes, states, and effects are represented, respectively:



These are needed to capture the operations of composing processes in sequence, called *composition*, and combining them in parallel, called *tensoring*. The composition $g \circ f$ (first f , then g) and tensor $f \otimes g$ are represented as:



We call any formalism with a notion of composition and tensoring a *process theory*. The semantics of process theories and their diagrams are governed by the theory of monoidal categories, which is surveyed in [11]. The goal of the paper is to specify an appropriate symmetric monoidal category, **Rand**, representing the appropriate operations on random variables, after which we can define a causal model as a strong monoidal functor from a causal theory into **Rand**. These causal models—essentially, diagrams in **Rand**—will be the promised abstract indicator frameworks. In Section 2, we will consider a preliminary version of **Rand** along with some of the possible alternatives. In Section 3, we will give the statistical justification for our choice of **Rand**, review the notion of a causal model from [4], and then give the full definition of abstract indicator frameworks.

2. BACKGROUND

As mentioned above, there are a variety of approaches to choosing indicator frameworks as part of the process of strategic priorities. Of the many specialized approaches to choosing indicator frameworks in various fields, Niemeijer and de Groot [9] have suggested a similar methodology for choosing environmental indicator sets based on explicit causal networks of environmental forces and societal response; while their methodology is still largely qualitative rather than formal or statistical, their paper handily illustrates how (diagrams of) causal models can facilitate the selection of relevant indicator sets. In statistics, Horvath [6] also takes a compositional approach to correlation by focusing on weighted correlation networks, which represent random variables by nodes in a graph and edges between variables by a soft threshold on their correlation. These correlation networks have proved useful for analyzing high-dimensional data sets, especially gene expression data.

Even within the constraints of a process theory, there are still a number of diagrammatic approaches to probability. In this section, we will go over three examples: the traditional Hilbert space interpretation of random variables, the original category of probabilistic mappings suggested by Lawvere [8], and the diagrammatic approach of Coecke and Spekkens [1]

to Lawvere’s work. We also briefly discuss graphical models such as those surveyed in [7], which are the most obvious applications of [8] and [1], e.g. see [4].

The traditional approach, which we call **Rand**, uses the fact that real-valued random variables over some fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ form a Hilbert space H where the inner product $\langle X, Y \rangle$ is just the covariance $\mathbb{E}(XY)$. Assuming that we restrict ourselves to standard variables with zero mean and unit variance, the covariance equals the correlation, and we can represent both by the inner product in **Rand**.

This inner product can be written as a process diagram, namely as the composition of a state and an effect in **Rand** into the following diagram:

$$\text{Cor}(X, Y) = \text{Cov}(X, Y) = \langle X, Y \rangle = \begin{array}{c} \triangleup Y \\ \downarrow \\ \triangle X \end{array}$$

As we will see in the next section, **Rand** is actually already very close to what we want; the problem is that the obvious categorification of **Rand** does not give a natural way of analyzing the data of “intermediate” correlations.

Rand is closely related to older work on the categorical foundations of probability initiated by Lawvere in [8] and developed in Giry [5]:

Definition 2.1 *The category **Stoch** of stochastic processes is defined by the following data:*

1. *objects are measurable spaces (A, Σ_A) of sets A with a σ -algebra Σ_A*
2. *morphisms $P : (A, \Sigma_A) \rightarrow (B, \Sigma_B)$ are stochastic kernels, i.e. functions $P : A \times \Sigma_B \rightarrow [0, 1]$ that assign to (a, σ_B) the probability of σ_B given a , denoted $P(\sigma_B|a)$*
3. *composition $Q \circ P : A \times \Sigma_C \rightarrow [0, 1]$ of $P : (A, \Sigma_A) \rightarrow (B, \Sigma_B)$ and $Q : (B, \Sigma_B) \rightarrow (C, \Sigma_C)$ is defined by*

$$(Q \circ P)(\sigma_C|a) = \int_{b \in B} Q(\sigma_C|b) dP_a,$$

i.e. marginalization over B

As suggested by the notation, morphisms in **Stoch** represent probability measures on an event/outcome space (A, Σ_A) . If we restrict to the subcategory **FinStoch** whose objects are *finite* measurable spaces—since the outcomes are finite, one can imagine these spaces as sets of natural numbers $\{1, \dots, n\}$ —then we can think of stochastic kernels as stochastic matrices, i.e. matrices whose column entries sum to 1. Taking $1 = (\{*\}, \Sigma_*)$ as the monoidal unit, we can see that a probability distribution in **FinStoch** is just a vector $P : 1 \rightarrow (A, \Sigma_A)$ whose entries are the probabilities of all the possible atomic outcomes in A . The usual tensor product described in **Stoch** is the functor $\otimes : \mathbf{Stoch} \times \mathbf{Stoch} \rightarrow \mathbf{Stoch}$ that assigns to two probability distributions $P : 1 \rightarrow (A, \Sigma_A)$, $Q : 1 \rightarrow (B, \Sigma_B)$ their product measure, i.e. the map $PQ : 1 \rightarrow (A \times B, \Sigma_A \otimes \Sigma_B)$ s.t. $PQ(*, (a, b)) = P(*, a)Q(*, b)$.

Using the language of symmetric monoidal categories, Coecke and Spekkens [1] give a graphical calculus for **FinStoch** and use it to elaborate Bayesian reasoning (in particular, a diagrammatic representation of Bayes’ rule). As above, objects of **FinStoch** are natural numbers, morphisms from m to

n are $n \times m$ stochastic matrices, composition is matrix product, and the monoidal product is the matrix tensor product. States are probability distributions over the set $1, \dots, n$:

$$\begin{array}{c} \downarrow \\ \triangleleft P \triangleright \end{array} : 1 \rightarrow A = (p_1, \dots, p_n) \text{ such that } \sum_{j=1}^n p_j = 1$$

A joint state is a state over the composite object $1, \dots, mn$. A joint state is “uncorrelated”—one should be careful, since this is correlation between probability distributions, not between random variables per se—when it can be decomposed into a tensor product, and perfectly correlated when it can be represented as a delta function. Uncorrelated and perfectly correlated joint states are depicted, respectively:

$$\begin{array}{c} \downarrow \quad \downarrow \\ \triangleleft P \triangleright \quad \triangleleft Q \triangleright \end{array} : 1 \rightarrow A \otimes B = (p_1 q_1, \dots, p_n q_m)$$

$$\begin{array}{c} \cup \\ \bullet \end{array} : I \rightarrow A \otimes A = \mathbf{e}^T = (\delta_{i,i'} | i, i' \in \{1, \dots, n\}).$$

A perfectly anti-correlated joint state in **FinStoch** is just a cup with a NOT-gate attached to one end. More generally, any correlation can be obtained by attaching a suitable box to one of the ends of the cup.

There are two major problems with this graphical formalism, and with **Stoch** and **FinStoch** in general. First and foremost, there is not a very convenient way of talking about *random variables*. Technically, a real-valued random variable is given by the diagram below, of a measurable function $X : (\Omega, \Sigma_\Omega) \rightarrow (\mathbb{R}, \Sigma_\mathbb{R})$ which takes possible outcomes in Ω to their numerical representations in \mathbb{R} (technically, X is not a function but a stochastic matrix whose columns represent point probabilities), a probability measure $P : 1 \rightarrow (\Omega, \Sigma_\Omega)$ on the outcomes in Ω , and finally the pushforward $X(P)$ of P along X , which represents the probability distribution of the random variable X .

$$\begin{array}{ccc} 1 & \xrightarrow{P} & (\Omega, \Sigma_\Omega) \\ & \searrow X(P) & \downarrow X \\ & & (\mathbb{R}, \Sigma_\mathbb{R}) \end{array}$$

Besides being difficult to work with, the point of view of this paper is that probability measures are not random variables nor are they sufficient replacements; traditionally, one views probability measures as something probability theorists invented in order to talk about random variables. While useful in the context of Bayesian networks, which can be articulated primarily in terms of stochastic processes, probability measures are less visible in cases driven by data and by correlational arguments.

The second problem is that there is not a good interpretation of *effect*, i.e. of “measuring” or computing something with respect to a specific state. An effect in **FinStoch**

$$\begin{array}{c} \triangleleft X \triangleright \\ \uparrow \end{array}$$

is defined to be a morphism $X^\dagger : (A, \Sigma_A) \rightarrow 1$, i.e. a function $X^\dagger : A \times \Sigma_* \rightarrow [0, 1]$. The problem is that 1 is terminal in **Stoch**: there is only one possible morphism from any object to 1 due to the constraint on morphisms of being a probability measure. In particular, for any (A, Σ_A) and for all $a \in A$,

the unique map $X^\dagger : (A, \Sigma_A) \rightarrow 1$ is given by $X^\dagger(*|a) = 1$ and $X^\dagger(\emptyset|a) = 0$. In other words, any ‘measurement’ of a state (i.e. a probability distribution) in **FinStoch** and **Stoch** simply kills the state.

Finally, in Bayesian statistics and machine learning, a variety of more generic graphical approaches, called graphical models, have been developed to model the conditional (in)dependence of multivariate random variables; the joint distribution over all the random variables in a graphical model is the product of their conditional distributions. Among the most familiar examples of graphical models are (directed) Bayesian networks and (undirected) Markov networks. The upshot is that complex questions about joint distributions of many interrelated variables can be answered in terms of the topology of the graph. We mention these graphical models, and especially Bayesian networks (and by extension stochastic matrices), since they ground many existing approaches to integrating causality with probability, including that of **Stoch** and **FinStoch**.

3. INDICATOR FRAMEWORKS

In this section, we define the category **Ind** of abstract indicator frameworks and give an example, with diagrams, of a concrete indicator framework.

Before giving the definition of the category **Ind** of abstract indicator frameworks, we will go through some of the statistical justification. Suppose that we have a correlation between random variables X and Y and another one between Y and Z . What can we say about the correlation between X and Z ? One obvious guess would be

$$\text{Cor}(X, Z) = \text{Cor}(X, Y)\text{Cor}(Y, Z). \quad (1)$$

Of course we know that Equation 1 is, in general, false.¹ But it is, under certain conditions, still the best guess. The following result is a standard exercise in statistics.

Lemma 3.1 *If $a = \text{Cor}(X, Y)$ and $b = \text{Cor}(Y, Z)$, then*

$$\text{Cor}(X, Z) \geq ab - \sqrt{1 - a^2}\sqrt{1 - b^2} \quad (2)$$

$$\text{Cor}(X, Z) \leq ab + \sqrt{1 - a^2}\sqrt{1 - b^2} \quad (3)$$

PROOF. WLOG, assume that A, B, C are standard variables with zero mean and unit variance, since the correlation is invariant under changes to mean and variance. We can write $X = aY + E_{Y,X}$ and $Z = bY + E_{Y,Z}$ where, by construction, $E_{Y,X}, E_{Y,Z}$ are random variables uncorrelated with Y .

Then $\langle X, Z \rangle = \text{Cor}(X, Z) = \langle aY + E_{Y,X}, bY + E_{Y,Z} \rangle = ab + \langle E_{Y,X}, E_{Y,Z} \rangle$, and we can use the Cauchy-Schwarz inequality to bound $\langle E_{Y,X}, E_{Y,Z} \rangle$ from above and from below, giving the lemma. \square

The lemma tells us that there is a range of possible values, centered around $\text{Cor}(X, Y)\text{Cor}(Y, Z)$, for the composite correlation; unfortunately, in practice that range can so large

¹Correlations are rarely composed in practice because (1) the computation is usually false and (2) because we can usually compute the composite correlation directly from the data. It is only when we lack the data (which is quite often in studies of complex systems) that we use a causal model to infer the correlation. Unfortunately, causal models are often invoked in the process of imposing a learned model such as a Kalman filter or a dynamical Bayesian network, which will often conflate the statistical and causal contributions.

as to be useless. In such a situation, we may ask what is the obstruction, given $\text{Cor}(X, Y)$ and $\text{Cor}(Y, Z)$, to knowing the canonical or ‘true’ correlation of their composite, and whether we can reduce or get around that obstruction. Reading the proof of the lemma, we know the obstruction is just the correlation $\langle E_{Y,X}, E_{Y,Z} \rangle$; that is, if $\langle E_{Y,X}, E_{Y,Z} \rangle$ were 0, our guess would be valid.

One may also derive the above result from the definition of the partial correlation of X and Z , fixing Y . Recall that the partial correlation $\rho_{XZ \cdot Y}$ is defined as the correlation between the residuals of X and of Z , fixing Y . In terms of their component correlations,

$$\rho_{XZ \cdot Y} = \frac{\text{Cor}(X, Z) - \text{Cor}(X, Y)\text{Cor}(Y, Z)}{\sqrt{1 - \text{Cor}(X, Y)^2} \sqrt{1 - \text{Cor}(Y, Z)^2}}.$$

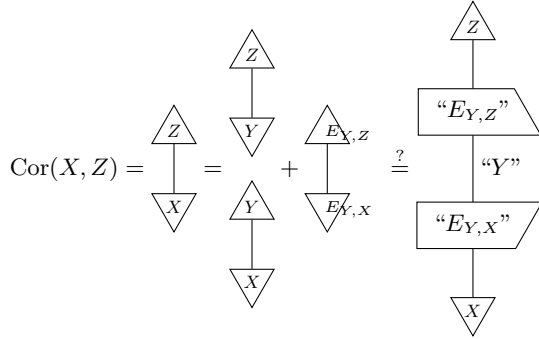
Thus

$$\langle E_{Y,X}, E_{Y,Z} \rangle = \rho_{XZ \cdot Y} \sqrt{1 - \text{Cor}(X, Y)^2} \sqrt{1 - \text{Cor}(Y, Z)^2}.$$

In other words, Equation 1 is correct just when the partial correlation $\rho_{XZ \cdot Y} = 0$, when $\text{Cor}(X, Y) = 1$ or -1 (i.e. X and Y are linear functions of each other), or when $\text{Cor}(Y, Z) = 1$ or -1 . This allows us to produce another guess:

$$\text{“Cor}(X, Z)\text{”} = Y \text{ s.t. } \rho_{XZ \cdot Y} = 0 \quad (4)$$

Explicitly, $E_{Y,X}$ measures the nonlinear component of the relation between X and Y . But one may also think of it as a measure of the ‘noise’ or ‘error’ between X and Y , at least as it concerns the correlation. The idea of Equation 4 is that, if we are lucky in choosing Y , then the noise factors $E_{Y,X}$ and $E_{Y,Z}$ will “cancel out” to produce the true correlation $\text{Cor}(X, Z) = \text{Cor}(X, Y)\text{Cor}(Y, Z)$. Heuristically, we can represent this process as below:



That is, the correlation between X and Z can be computed by applying a transformation “ $E_{Y,X}$ ”, representing some sort of structured noise factor, then applying a transformation “ $E_{Y,Z}$ ” that cancels out the noise introduced by “ $E_{Y,X}$ ”.

We can formalize this intuition. Recall that real-valued, square-integrable random variables over a given probability space form a Hilbert space $L^2(\Omega, \Sigma, \mathbb{P})$ whose inner product is just the covariance.

Definition 3.2 *The category of random variables, Rand , is defined by the following data:*

1. *objects are finite-dimensional Hilbert spaces*

$$\mathcal{X} = L^2(\Omega_{\mathcal{X}}, \Sigma_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$$

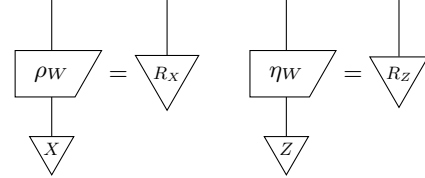
of square-integrable random variables (under the equivalence relation $X_1 \sim X_2$ if $\mathbb{P}_{\mathcal{X}}(X_1 = X_2) = 1$) with

inner product $\langle X, Y \rangle = E(XY)$, defined over probability spaces $(\Omega_{\mathcal{X}}, \Sigma_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$, with an associated basis $B_{\mathcal{X}} = \{X_1, X_2, \dots, X_n\}$

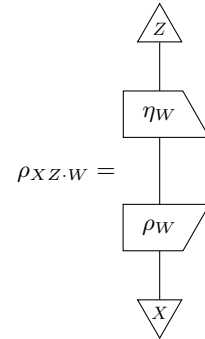
2. *morphisms $F : \mathcal{X} \rightarrow \mathcal{Y}$ are bounded linear operators*
3. *the composition is the usual composition of bounded linear operators*
4. *the tensor product of \mathcal{X} and \mathcal{Y} is the pushout over their joint support in $\Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$*

Example 3.3 *Suppose that the transportation department buys a new bus and designates an indicator, X , that counts the number of riders on the bus per day. Elsewhere, the education department tracks an indicator, Z , that counts the number of students per day who are absent from class across the whole city.*

In this case, integrating data is not a problem, so we can compute the correlation between X and Z . But even if we did, the correlation may be very small, or conversely it may be absurdly high, especially if there is some antecedent variable correlated with both X and Z , e.g. an economic boom. Suppose the federal government tracks a separate variable, W , on aggregate economic performance per quarter. The first step is to get rid of the influence of W , i.e. compute the residuals R_X, R_Z of X, Z resulting from their linear regression with W . Assuming that X and Z live in indicator sets \mathcal{X} and \mathcal{Z} respectively, we can represent the residual as transformations ρ_W, η_W on $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$, respectively:



Then by definition, we have



As shorthand, we will sometimes refer to the space of random variables \mathcal{X} as a set of variables or indicators; in such cases, we always mean the basis set of random variables, $B_{\mathcal{X}}$.

It will help to think of random variables as representing column vectors or “dimensions” of data in a table of such data, where row vectors in that table represent particular data points. The correlation between two column vectors is just their sample correlation. This has several benefits: it makes the inner product (correlation) and tensor product (entity resolution) very concrete, it is what a data analyst

actually looks at, and it highlights the restrictions and challenges imposed by the presence and absence of data. In fact, we can define a category **Data** explicitly in such terms:

Definition 3.4 *The category of \mathbb{R} -valued data tables, **Data**, is defined by the following data:*

1. objects $\mathcal{X} = (\mathcal{X}, \Omega_{\mathcal{X}}, \mathbb{I}_{\mathcal{X}})$ of **Data** are $m \times n$ tables of \mathbb{R} -valued data vectors whose rows are assigned an index key given by $\mathbb{I}_{\mathcal{X}} : \Omega_{\mathcal{X}} \rightarrow \mathbb{R}$ and whose columns, $B_{\mathcal{X}} = \{X_1, \dots, X_n\}$, represent indicators
2. morphisms $f : \mathcal{X} \rightarrow \mathcal{Y}$ are linear transformations of the column values by addition and scalar multiplication
3. the composition is just the matrix product
4. the tensor product of $\mathcal{X} \otimes \mathcal{Y}$ is the integrated table of their data values over a table of linkages, $S \subset \Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$

Suppose we are working in a 3-dimensional Hilbert space \mathcal{X} with a basis of random variables $B_{\mathcal{X}} = \{X, Y, Z\}$. In this basis, the random variable $E_{Y,X}$ is just the vector $X - \langle X, Y \rangle Y$ (and similarly with $E_{Z,X}$), but the problem with this space... is that there is no problem! In **Rand**, having a basis in X, Y, Z corresponds to the observation, in **Data**, that we already have the tabular data we need to compute $\langle X, Z \rangle$ directly. But in many situations of interest in a complex, open system, e.g. in computing the second-order impacts of local and/or technical projects, such broad-based data is difficult to obtain.

So we lack data. But to take just one example, in a database setting there are ways to reason about “missing data”, e.g. database nulls, especially when that data is the subject of a data migration or integration, as described in [12]. In particular, one can impose a set of algebraic equations that each null value must satisfy, where the equations are given by a diagram of database schema mappings. More generally, almost every diagram in a category articulates a set of constraints on the objects of that category.

Example 3.5 *Recall our earlier example, where X stands for the number of riders on a particular bus in a city, and Z stands for the number of absent students across a city. Suppose that we have already controlled X and Z for economic performance (i.e. W) along with any number of other confounding variables, that we have found (or suspect) a small but significant correlation between X and Z , and that we are now interested in understanding how X correlates with Z .*

There may be a variety of possible explanations for why this correlation exists: maybe dropping the price of a ticket (thus promoting more bus ridership) allows more students to go to school, or perhaps additional bus ridership decreases traffic, which gives harried parents more time to track their truant children. Without choosing any one explanation, we can represent the statistical properties of a set of mediating, “explanatory” variables \mathcal{Y} by a ‘sum’ of the possible explana-

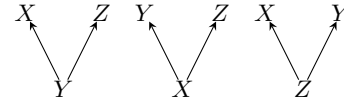
tions:

$$\text{Cor}(X, Z) = \sum_{Y \in \mathcal{Y}} \left(\begin{array}{c} \triangle Z \\ \downarrow \\ \eta_Y \\ \downarrow \\ \mathcal{Y} \\ \downarrow \\ \rho_Y \\ \downarrow \\ \triangle X \end{array} \right) \quad (5)$$

The equation above succinctly represents a set of constraints that we can impose on the intermediate framework \mathcal{Y} , and motivates the following definition.

Definition 3.6 *A mediating framework between two sets of random variables \mathcal{X}, \mathcal{Z} is a set of random variables \mathcal{Y} such that Equation 5 is satisfied for all variables $X \in \mathcal{X}, Z \in \mathcal{Z}$.*

Rand and **Data** supply the basic statistical foundation for a theory of indicator frameworks. We would now like to incorporate a causal foundation. There are several reasons for doing so. First: many statistical arguments, e.g. partial correlation, actually rely on an implicit choice of causal model—see discussions related to confounding and mediating variables by Pearl [10], among others. For example, given three random variables X, Y, Z ; the partial correlations $\rho_{XZ \cdot Y}$, $\rho_{XY \cdot Z}$, or $\rho_{YZ \cdot X}$ are all equally valid; which one we take as ‘true’ depends on which of the following causal structures we believe is true:



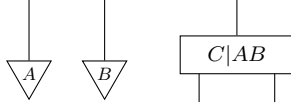
Second, many operational indicator frameworks are constructed based on experts’ causal models of the indicators, e.g. as in [9] or as in the air pollution diagram shown at the very beginning of this paper. Any story of indicator frameworks would be incomplete without mentioning causation. And third, the pattern developed here for causation will be useful later, when we want to incorporate not only causal models but arbitrary scientific models (such as those in Bayesian networks) into our indicator frameworks.

We recall the definition of causal theory from Fong [4], as a certain symmetric monoidal category induced from a directed acyclic graph (i.e. the causal structure), such as any of the three graphs above. Without going into the details, the idea is that given such a causal structure, we can specify a symmetric monoidal category whose objects are collections of the letters $\{X, Y, Z\}$, and whose morphisms are generated by the counit (representing ‘deletion’) and comultiplication (representing ‘copying’), depicted respectively by



and by a set of causal mechanisms generated from the causal structure, $[A] : \emptyset \rightarrow A$, $[B] : \emptyset \rightarrow B$, and $[C|AB] : AB \rightarrow$

\mathcal{C} , depicted as



Given a causal theory, i.e. a symmetric monoidal category, we can define a model of that causal theory \mathcal{C} in **Rand** as a strong monoidal functor $F : \mathcal{C} \rightarrow \mathbf{Rand}$. To specify such a functor, it suffices to define its behavior on every atomic variable and every generating map in \mathcal{C} , i.e. the counit, comultiplication, and causal mechanisms of \mathcal{C} , since the values of the functor on the rest of \mathcal{C} is specified up to isomorphism by the definition of a strong monoidal functor.

What is the functor doing? If A is an atomic causal variable of the causal theory \mathcal{C} , then F sends A to a one-dimensional Hilbert space, e.g. one with basis set $\{X\}$. Then on tensor products of atomic causal variables, $F(A \otimes B)$ gives the tensor product $F(A) \otimes F(B)$, i.e. the integrated random variable with basis set $\{F(A), F(B)\}$ and probabilities inherited from the product measure. On morphisms, $F([A]) : \mathbb{R} \rightarrow F(A)$ is just a random variable, and a causal mechanism $[C|AB]$ becomes a transformation $F([C|AB]) : F(A \otimes B) \rightarrow F(C)$.

We can now state the definition of **Ind**.

Definition 3.7 *The category **Ind** of abstract indicator frameworks is defined by the following data:*

1. *an object I of **Ind** is a strong symmetric monoidal functor $\mathcal{C} \rightarrow \mathbf{Rand}$ from a causal theory \mathcal{C} to the category of random variables.*
2. *a morphism η between abstract indicator frameworks is a natural transformation of strong symmetric monoidal functors*

One may directly compare **Ind** with the category of stochastic causal models in [4], which are generalizations of Bayesian networks.

4. CONCLUSION

In this paper, we sought to give a rigorous mathematical alternative to the traditional, indicator-by-indicator process of constructing indicator frameworks, especially in city planning and project governance. We proposed that indicator frameworks could be defined (and optimized) by means of their relationships to other indicator frameworks. These relationships were probabilistic as well as causal. Therefore, we sought to develop a semantics for the problem of constructing indicator frameworks that could accommodate both probabilistic and causal modes of reasoning.

We examined several options for the semantics of probability, including **Bayes** [1] and **Stoch** [8]. After reflecting on the practical necessities of data analysis, we decided to base our construction on a category more directly in terms of random variables and correlations, and defined the symmetric monoidal category **Rand** random variables and correlations. We then introduced the idea of a causal model from [4], and used this to motivate the definition of the category **Ind** of abstract indicator frameworks as models of a causal theory in **Rand**.

We then used **Ind** as the setting for an optimization problem: how to construct a mediating indicator framework that

best explains the relationship between a given set of indicators, such as those of a specialized project in a city, and another set of indicators, such as headline indicators of broad interest to the public. Such a mediating framework can be used to answer the question, “what are the secondary impacts of my project?”

This paper is the subject of ongoing research; future versions will address applications to more complicated, real-world examples in city administration, as well as the integration of other mathematical models beyond causal ones into the framework. Other additional future work include the possibility of studying the constraints on data-supported applications introduced by constraints on their underlying indicators, as well as closer examinations of the phenomenon of tensoring or ‘gluing’ data sets, and the possible obstacles such gluings might introduce to producing a consistent global picture of the complex system.

5. ACKNOWLEDGEMENTS

We would like to thank Bob Coecke, Bilin Guvenc, Levent Guvenc, Derek Loftis, and Ed Griffior for helpful conversations in the writing of this paper.

Disclaimer: certain commercial products may be identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

6. REFERENCES

- [1] B. Coecke and R. Spekkens. Picturing classical and quantum bayesian inference. *Synthese*, 186(3):651–696, June 2012.
- [2] G. T. Doran. *Management review*, 70(11):35–36, 1981.
- [3] M. J. Epstein and J.-F. Manzoni. The balanced scorecard and tableau de bord: translating strategy into action. *Strategic Finance*, 79(2):28, 1997.
- [4] B. Fong. Causal theories: A categorical perspective on bayesian networks. Preprint, April 2013.
- [5] M. Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, volume 915 of *Lecture Notes in Mathematics*, pages 68–85. Springer-Verlag, 1982.
- [6] S. Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer, 2011.
- [7] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
- [8] F. W. Lawvere. The category of probabilistic mappings. Unpublished seminar notes.
- [9] D. Niemeijer and R. S. de Groot. A conceptual framework for selecting environmental indicator sets. *Ecological Indicators*, 8:14–25, 2008.
- [10] J. Pearl. *Causality*. Cambridge University Press, 2009.
- [11] P. Selinger. A survey of graphical languages for monoidal categories. Preprint, August 2009.
- [12] D. I. Spivak. Functorial Data Migration. *ArXiv e-prints*, Sept. 2010.