Appendix

A Study of the Impact of DNS Resolvers on CDN Performance Using a Causal Approach

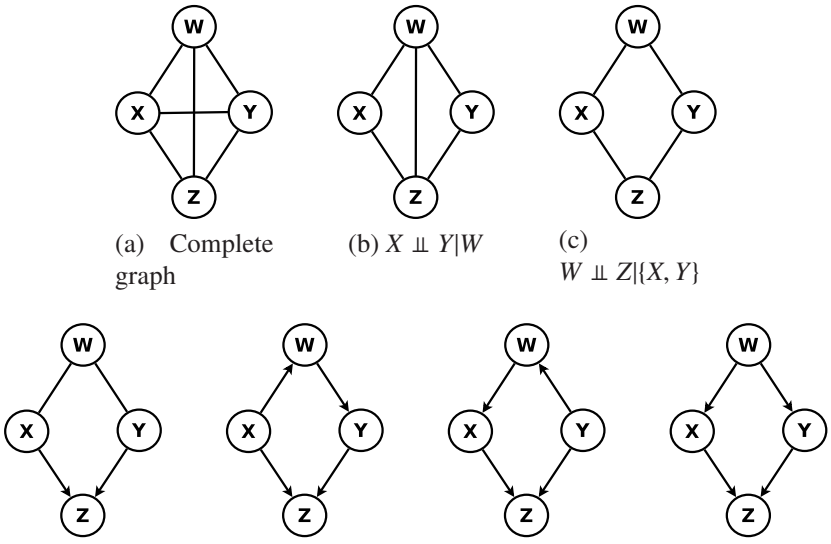
Hadrien Hours[a,b], Ernst Biersack[c], Patrick Loiseau[a], Alessandro Finamore[d,e], Marco Mellia[e]

[a] EURECOM, email: firstname.lastname@eurecom.fr
[b] ENS Lyon, email: firstname.lastname@ens-lyon.fr
[c] Caipy, email: erbi@e-biersack.eu
[d] Telefonica, email: firstname.lastname@telefonica.com
[e] Politecnico di Torino, email: firstname.lastname@polito.it

Appendix A. Causal model inference

Appendix A.1. PC algorithm

We have described the PC algorithm in Section 2.1. In Figure A.7 we now illustrate the different steps. In this example we try to infer the causal model of the system corresponding to four parameters, W, X, Y, Z. We could detect two independences: $\mathbf{I}_1 = (X \perp\!\!\!\perp Y|W)$ and $\mathbf{I}_2 = (W \perp\!\!\!\perp Z|\{X, Y\})$. In Figure A.7b and Figure A.7c the edge between X and Y and the edge between W and Z are removed for the independences detected with a conditioning set size of 1 and 2 respectively. Because of detected $X \perp\!\!\!\perp Y$ but $X \not\perp\!\!\!\perp Y|Z$ but we can orient $X - Z - Y$. However, the orientation of the second V-structure, $X - W - Y$, cannot be deduced from the set of detected independences. The three orientations presented in Figure A.7e, Figure A.7f and Figure A.7g verify the independences $\mathbf{I}_1$ and $\mathbf{I}_2$.



(a) Complete graph (b) $X \perp\!\!\!\perp Y|W$ (c) $W \perp\!\!\!\perp Z|\{X, Y\}$

(d) Orientation V-structure (e) Orientation V-structure choice 1 (f) Orientation V-structure choice 2 (g) Orientation V-structure choice 3

Figure A.7: Different steps of the inference of causal model corresponding to a system of four parameters, $\{W, X, Y, Z\}$ with the detected independences $\mathbf{I}_1 = (X \perp\!\!\!\perp Y|W)$ and $\mathbf{I}_2 = (W \perp\!\!\!\perp Z|\{X, Y\})$

Appendix A.2. Independence test

The accuracy of the PC algorithm comes from the accuracy of the test used to test parameter independences. Compared to our previous works [Hours et al., 2015], one difference comes from the presence of a categorical variables, like the DNS service used by the clients observed in our study or the destination IP address. The test we use in our study is the KCI test [Zhang et al., 2012] combined with a bootstrap approach to solve numerical issues in its use of Cholesky factorization and to parallelize computations and decrease the algorithm completion time [Hours et al., 2015].

To validate the use of the KCI test in the presence of categorical variable, we generate two artificial datasets as follows:

- Dataset 1:
  - X_1 is a categorical variable with 4 levels: $X_1 \sim \mathbf{U}\{c_1, c_2, c_3, c_4\}$.
  - X_2 is a deterministic mapping of X_1, adding 20% of Gaussian noise: $X_2 = f_2(X_1) + \varepsilon$.
  - X_3 is a deterministic mapping of X_1, adding 20% of Gaussian noise, $X_3 = f_3(X_1) + \varepsilon$.
  - X_4 is a function of X_2 and X_3, adding 20% of Gaussian noise: $X_4 = f_4(X_2, X_3) + \varepsilon$.
  - with
    - f_2 and f_3 defined as $f_i(c_j) = c_{i,j}$ for $j \in \{1, 2, 3, 4\}$, with $c_{ij} \neq c_{i'j'}$ if $i \neq i'$ or $j \neq j'$
    - $f_4(x, y) = \sqrt{x + y}$.
    - ε an error terms following normal distribution with a mean equals to 0 and a variance equals to $0.2 \times \sigma_{f_i(X_i)}$.
- Dataset 2:
  - X_1 is a categorical variable with 4 levels: $X_1 \sim \mathbf{U}\{c'_1, c'_2, c'_3, c'_4\}$.
  - X_2 is a deterministic mapping of X_1, adding 20% of Gaussian noise: $X_2 = f'_2(X_1) + \varepsilon$.
  - X_3 is a categorical variable with 4 levels, the probability of each level depends on X_1: $X_3 = f'_3(X_1) + \varepsilon$.
  - X_4 is a function of X_2 and X_3, adding 20% of Gaussian noise: $X_4 = f'_4(X_2, X_3) + \varepsilon$.
  - with
    - f'_2 defined by, $f'_2(c'_j) = c_{2,j}$ for $j \in \{1, 2, 3, 4\}$,
    - $f'_3(c'_j) \in \{c''_1, c''_2, c''_3, c''_4\}$, each value c''_k drawn from 4 different distributions, chosen base on the value of c'_j
    - $f'_4(x, y) = c(x) + \sqrt{y}$, with $c(x)$ defined as a deterministic mapping of x, similarly to f'_2.
    - ε an error terms following normal distribution with a mean equals to 0 and a variance equals to $0.2 \times \sigma_{f_i(X_i)}$.

The graphical causal model corresponding to these dependencies is presented in Figure A.8. The definition of the two datasets leads to two independences $\mathbf{I}_1 = (X_2 \perp\!\!\!\perp X_3 | X_1)$ and $\mathbf{I}_2 = (X_1 \perp\!\!\!\perp X_4 | \{X_2, X_3\})$.

We test different independences for 20 artificial datasets of size 1000, generated according the definitions given above. The average p-values of the different tests for the two classes of artificial datasets are presented in Table A.3. We can see that the KCI performs correctly even in the presence of complex dependencies including the presence of categorical variables. The conclusion of this study is that the KCI performs as expected in the presence of categorical variable and can be used in our study.

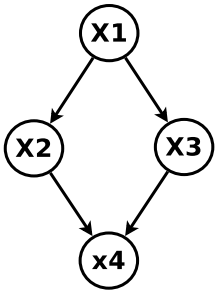


Figure A.8: Graphical causal model illustrating the dependencies of the artificial dataset parameters

Table A.3: Results of the KCI test when testing independences with the presence of categorical parameters

Independence	p-value dataset 1	p-value dataset 2
$X_1 \perp\!\!\!\perp X_2$	0	0
$X_1 \perp\!\!\!\perp X_3$	0	0
$X_1 \perp\!\!\!\perp X_4$	0	0
$X_2 \perp\!\!\!\perp X_3$	0	0
$X_2 \perp\!\!\!\perp X_4$	0	0
$X_3 \perp\!\!\!\perp X_4$	0	0
$X_1 \perp\!\!\!\perp X_2 \mid X_3$	0	0
$X_1 \perp\!\!\!\perp X_2 \mid X_4$	0	0
$\boldsymbol{X_2 \perp\!\!\!\perp X_3 \mid X_1}$	**0.4**	**0.6**
$X_2 \perp\!\!\!\perp X_3 \mid X_4$	0	0
$X_3 \perp\!\!\!\perp X_4 \mid X_1$	0	0
$X_3 \perp\!\!\!\perp X_4 \mid X_2$	0	0
$X_1 \perp\!\!\!\perp X_4 \mid X_2$	2e-14	2e-2
$X_1 \perp\!\!\!\perp X_4 \mid X_3$	0	3e-4
$\boldsymbol{X_1 \perp\!\!\!\perp X_4 \mid \{X_2, X_3\}}$	**0.7**	**0.5**

Appendix A.3. Markov Equivalence class

In our study of the impact of the DNS service choice on the Akamai CDN performance, we obtain the Bayesian network represented in Figure A.9. Using our understanding of CDN and the parameters present in our model, we orient the undirected edges and obtain the Bayesian network representing the causal model of our system represented in Figure A.10. Using the Tetrad software [Spirtes et al., 2001], it is easy to represent all the members of what is called the *Markov Equivalence Class*. There can be several graphs that represent the set of independences that were detected from the tests performed on a given series of observations. The set of all these graphs can be represented by a partially oriented graph, Figure A.9. In Figure A.11 we represent the eight members of the Markov Equivalence Class corresponding to the set of independences that were

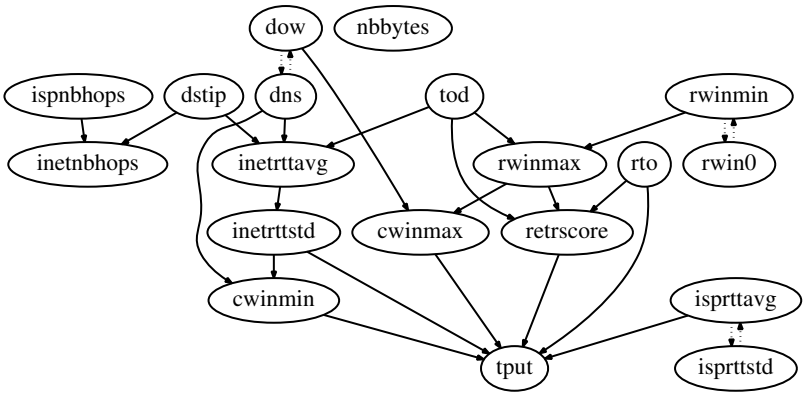


Figure A.9: Output of the PC algorithm when no domain knowledge is used

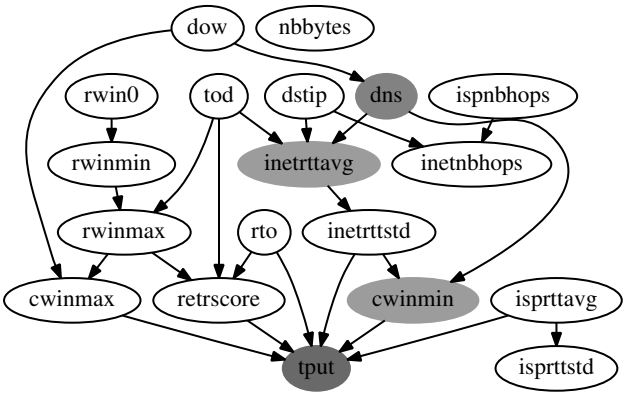


Figure A.10: Bayesian network representing the causal model of Web performance using two different DNS: the public Google DNS and the DNS of the local ISP

detected from the series of tests performed on the observations of the parameters of our system.

Appendix B. Predicting interventions

Appendix B.1. Theory

Appendix B.1.1. Atomic interventions

The description of atomic intervention was presented in Section 2.2.1. In this section, we only repeat the rules of Do-calculus that will be used in the following sections to present the details of our method.

If we use G to denote the Bayesian graph that represents the causal relationships between the parameters of our system, we use $G_{\overline{X}}$ to denote the sub-graph of G where all the edges entering X are removed and $G_{\underline{X}}$ the sub-graph of G where all the edges exiting X are removed. We can use the rules of *do-calculus* from [Pearl, 2009] to estimate the distributions of the parameters of

our system after an intervention based on their distributions prior to this intervention. Note that these rules do not rely on any assumption regarding the distributions or functional dependencies of the parameters. In particular, P represents the (possibly multivariate) probability distribution specified by the probability mass function or probability density function depending on the nature of the parameters.

Theorem 2 (3.4.1 from [Pearl, 2009]). *(Rules of do calculus) Let* G *be the directed acyclic graph associated with a causal model [...] and let $P(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y and Z we have the following rules.*

Rule 1*(Insertion/deletion of observation):*

$$P(y|do(x), z, w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}} \tag{B.1}$$

Rule 2*(Action/observation exchange):*

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}} \tag{B.2}$$

Rule 3*(Insertion/deletion of intervention):*

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{XZ(W)}}}, \tag{B.3}$$

where Z(W) is the set of Z-nodes that are not ancestor of any W-nodes in $G_{\overline{X}}$.

Appendix B.1.2. Enforcing intervention with a given probability

In our study of the impact of the DNS service on CDN performance (throughput), we are interested in estimating the effect of interventions on parameters influenced by the DNS service and influencing the throughput. In such case, we do not limit ourselves to atomic interventions but we are interested in intervening on a given parameter to change its distribution.

From [Pearl, 2009, Section 4.2], if we want to predict how an intervention on X affects Y, where the intervention on X is enforced with the conditional probability distribution $f^*(X|Z)$, we obtain:

$$f(y)_{|f^*(x|z)} = \int_{D_X} \int_{D_Z} f_{Y|do(X),Z}(y, x, z) f^*(x|z) f(z) \mathrm{d}x \mathrm{d}z. \tag{B.4}$$

From Equation (B.4), one should notice that we need to integrate on the intervention parameter, X. For performance reasons, the estimations of $f_{Y|do(X),Z)}$ are made in parallel on different machines. Therefore, the estimation of $f_{Y|do(X),Z}(y, x, z)$ is done on a different machine for each x. As the data used in our study is not publicly available, we do not present the different parameterizations of the density estimation used for predicting $f(y)_{|f^*(x|z)}$.

Appendix B.2. Adapting the theory to our problem

Appendix B.2.1. Intervening on the DNS service

To understand how choosing one DNS service instead of another impacts CDN performance, we want to predict the throughput of a client who used a DNS service s_1 if the same client would have used a different DNS service, s_2, instead. To do so, we are interested in the impact of the DNS service on a given parameter, X, that in turn influences the throughput. We use the Theorem 2 and

Equation (B.4) to estimate the distribution of the throughput of a client using a given DNS service, s_1, if we intervene on the distribution of a parameter X, forcing its distribution to follow the one of X if the DNS service s_2 would have been used instead. This study is equivalent to study the impact of the parameter X for clients using the DNS service s_1 when intervening on the parameter X enforcing this intervention with the distribution $f_{X|DNS=s_2}$. If we denote Y the parameter capturing the performance of CDN users[5], we want to estimate:

$$f(Y|DNS = s_1, do(X \sim f(X|DNS = s_2))). \tag{B.5}$$

If we denote W the set of parameters blocking the spurious associations between X and Y, according to Theorem 2, we have

$$f(Y \mid DNS = s_1|do(X \sim f(X|DNS = s_2))) = \int_X \overbrace{\int_W f(Y \mid X = x, DNS = s_1, W) f(W)}^{f(Y|DNS=s_1,do(X=x))} f(X = x \mid DNS = s_2) P(DNS = s_2) \tag{B.6}$$

We can see from Equation (B.6) that the distribution of Y for users of the DNS s_1, if we intervene on X and fix its distribution to follow the distribution of X seen by the users of DNS s_2, is a weighted sum of the distribution of Y for DNS $= s_1$ after an atomic intervention on X ($do(X{=}x)$) with weights being the probability of observing $X = x$ under DNS $= d_2$.

Such approach allows to *(i)* Capture the effect of the DNS on a given mechanism influencing the performance of CDN users; *(ii)* Divide our prediction in a set of predictions of atomic interventions that can be estimated from the results of Theorem 2. Finally, we can use Equation (B.4) to estimate the final distribution of Y for the intervention on X that modifies its distribution.

Appendix B.2.2. Interventions and conditional multivariate distributions

It should be noticed that, as X is a continuous variable, the probability of observing a given value is 0. Therefore, instead of selecting samples for which $X = x$ is observed, we define an interval I_x corresponding to $[x-\delta_X; x+\delta_X]$ and assume that the samples for which the X parameter falls into this interval can be approximated to take the value x.

From the samples where $X \in I_x$ and DNS $= s_1$, we estimate the cumulative distribution function (CDF) of Y and W conditionally to DNS $= s_1$. We then estimate $f(Y|X = x, W = w, DNS = s_1)$ using the Sklar theorem.

The Sklar theorem stipulates that, if F is a multivariate cumulative distribution function with marginals $(F_1, \ldots, F_i, \ldots, F_n)$, there exists a copula C such that

$$F(x_1, \ldots, x_i, \ldots, x_n) = C(F_1(x_1), \ldots, F_i(x_i), \ldots, F_n(x_n)). \tag{B.7}$$

If we take the example of the bivariate distribution of two parameters X_1 and X_2, which marginals are denote F_1 and F_2 and f_1 and f_2 for the CDFs and PDFs respectively, we obtain, taking the derivative of Equation (B.7) with respect to X_1 and X_2:

$$f(x_1, x_2) = c(F_1(x_1), F_2(x_2)) f_1(x_1) f(x_2), \tag{B.8}$$

[5]In our study, we use the throughput to measure user performance

with f the bivariate PDF of X_1 and X_2.

As we have:

$$f_{X_1|X_2}(x_1, x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \tag{B.9}$$

for values of X_2 for which $f_2(x_2) \neq 0$, we can deduce that:

$$f_{X_1|X_2}(x_1, x_2) = c(F_1(x_1), F_2(x_2))f_1(x_1). \tag{B.10}$$

We can then estimate the conditional CDFs, $F(Y|X = x, DNS = s_1)$ and $F(W|X = x, DNS = s_1)$, using kernels and the samples where $X \in I_x$ and DNS = s_1, and use the previous formula to estimate $f(Y|X = x, W = w, DNS = s_1)$:

$$\begin{aligned} f(Y|X = x, W = w, DNS = s_1) =& c(F(Y|X = x, DNS = s_1), F(W|X = x, DNS = s_1)) \\ & f(Y|X = x, DNS = s_1), \end{aligned} \tag{B.11}$$

Finally, by integrating Equation (B.6) on W we obtain the distribution $f(Y|do(X = x), DNS = s_1)$.

We then select the samples for which $DNS = s_2$ and use normal kernels to estimate $f(X|DNS = s_2)$ and frequencies to estimate $P(DNS = s_2)$.

After these steps, we have all the factors present in Equation (B.6) and we can integrate over X to obtain the distribution of Y post intervention.

Appendix B.2.3. Estimation of marginals

Some practical issues are silenced in the sequence of steps described in Appendix B.2.2:

1. How to define the intervals I_X ?
2. As we are working with continuous variables, the two distributions conditionally to different DNS values might not have the same support. How do we define the conditional probability so that we have common values to integrate on ?

The last point is solved by always defining the PDFs domains as equally spaced points between the minimum and maximum observed value of the corresponding parameter in the whole dataset (DNS = s_1 or DNS = s_2). We use normal kernels to estimate the different distributions.

The first point, however, is more complicate as many possibilities exist and there is the constraint of finding enough samples in each interval to estimate the CDFs from which the copula parameters will be estimated and the conditional distributions derived.

Several solutions have been tested

- Variable bin width histogram,
- Fixed bin width and interpolation,
- Fixed bin width, filtering, rescaling.

Variable bin width histograms. The first method consists in fixing an objective number of samples and, starting from a fixed width bin histogram, merging adjacent bins until obtaining bins of different size but with a minimum number of samples:

Pros This method ensures the maximum number of atomic predictions being successful.

Cons As many of the parameters we observe have a long tail distribution, and as it is often in the tail of the parameter distributions that we find the values for interesting predictions, we obtain very large bins for the extreme values of the intervention parameter. The approximation stating that this bin represents a single value is then too strong. Additionally, when multiplying the atomic intervention fixing X parameter to the value x by the value of the PDF conditionally to the other DNS for $X = x$ (Equation (B.6)) we need to take into account the actual range of X that the atomic intervention represents to have a consistent approximation.

Fixed bin width. The second method is going in the opposite direction. We use histograms with fixed bins and then try to make predictions of $f(Y|do(x))$ with the number of samples we found in a given bin corresponding to a X value. If the estimation of the prediction fails, then we use interpolation of the $f(Y|do(x)$ for the X values where the post intervention PDF of Y could be estimated.

Pros This method solves the issues of approximation inconsistency of the variable bin width method. We fix the bin width and decide on the approximation of assimilating an interval to a given value.

Cons It can often happen that we manage to predict $f(Y|do(x)$ even when there are few sample in the interval I_x corresponding to the x value. However, with very few samples, it is very likely that this estimation is not accurate. This lack of accuracy impacts in a very negative way the overall post intervention PDF accuracy, as the PDF of Y post intervention that could be computed with few values will be used in the interpolation to recover the PDF Y for x values where the estimation of the post intervention PDF failed.

Fixed bin width and high pass filter. The last method, the one eventually adopted, is based on the fact that, if there are very few samples in a given area of the distribution of $X|DNS = s_1$ then this value is very unlikely to be observed and, as the previous method will impact negatively the overall results by including these samples, we consider that the PDF of $X|DNS = s_1$ in the domains with few samples is null.

This method uses a fixed bin histogram and selects only the bins where a minimum number of samples is observed to predict atomic interventions. After predicting the distribution of Y after intervention we rescale it based on the following observation:

$$\int_Y f(Y|do(X = x), DNS = s_1)dY = 1 \tag{B.12}$$

In practice, instead of varying the bin width and threshold we do the following: We select an objective number of samples, Δ_S, under which the atomic intervention $f(Y|X_2 = 0, do(X = x))$ is considered as null, and search for the optimal quantization leading to the maximum bins with a number of samples $\geq \Delta_S$. To do so we use this very simple algorithm:

Data: Vector X
Result: Set of bins (X_{final}) with fixed width (δ) defining the intervals around the values on which atomic interventions will be predicted:

$nB = 10$;
$nV_{old} = 0$;
$nV_{cur} = 1$;
X_{old} =[];
X_{cur} = [];
H_{old} = [];
H_{prev} = [];
while $nV_{cur} > nV_{old}$ **do**
    $H_t, E_t = hist(X, nB)$;
    $nV_{old} = nV_{cur}$;
    $nV_{cur} = nvalues(H_t > \Delta_S)$;
    $X_{old} = X_{cur}$;
    $X_{cur} = E_t$;
    $H_{old} = H_{cur}$;
    $H_{cur} = H_t$;
    $nB = 2 * nB$;
end
$nV_{final} = nV_{old}$;
$X_{final} = X_{old}$;
$H_{final} = H_{old}$;
$\delta = X_{final}(2) - X_{final}(1)$;

Algorithm 1: Dynamic quantization

This method solves the previous issues. There is a risk of not having any prediction for values in the tail of the distribution of $X|DNS = s_1$, that can represent the zone of overlap of the two distributions $X|DNS = s_1$ and $X|DNS = s_2$. However, limitations due to the lack of samples cannot be overcome by simple approximations as seen in the second method. The only way to solve this issue is to use parametric distributions (for example a mix of normal, gamma, beta, log-normal distributions). The studies made so far have shown that the approximation of the distributions of the parameters of our systems for values that are actually observed offers an acceptable accuracy. However, these models become inaccurate as soon as we try to use them to estimate the distribution of a parameter for values that have not been observed.

Appendix C. Parameterization of the method

In this section, we present a study that aims to answer the following questions:

1. How to choose the copula family that will model the dependencies between the different marginals ?
2. In which proportion the absence of values observed for both conditional distributions impacts the prediction accuracy and how to improve the accuracy in this case ?
3. As discussed previously, we estimate a complex intervention as the weighted average of atomic (simpler) interventions. Each atomic intervention is estimated on a sub domain

corresponding to an (small) interval around the value corresponding to our atomic intervention. On one hand, the bigger the interval is the more data we have to calibrate our model and the better prediction for the corresponding atomic intervention will be. On the other hand, if we manage to estimate many atomic interventions, we will have more inputs for estimating the global intervention we are interested in. The question boils down to finding the trade-off between the number of atomic interventions we estimate and the quality of each of these estimates.

To study these aspects we need a ground truth to estimate the accuracy of our prediction for different strategies and parameterization. Therefore, we generate a set of artificial datasets where the intersection between the two conditional distribution (DNS = s_1 and DNS = s_2) domains is modified and its impact on the accuracy of our method is studied.

Appendix C.1. Simulated dependencies

To simulate the same situation as the one met in the study of DNS service impact on the CDN performance, we randomly generate 4 parameters, X_1, X_2, X, Y, with dependencies illustrated by the graph presented in Figure C.13. To be closer to the situation observed in our study of the impact of DNS on CDN performance, we generate X_1 by randomly re-sampling the throughput observed in this study and we generate X_2 by randomly re-sampling the observed *DNS* from the same study. We eventually convert X_2 to binary value (*0* or *1*). The presence of a categorical data (the DNS in the corresponding study) is an important aspect that we want to keep in this study.

Appendix C.2. Intervention prediction

First, from the causal model, G, represented by the Bayesian network of Figure C.13, we can use the d-separation criterion to deduce the following independences:

- $(X \perp\!\!\!\perp Y | X_1, X_2)_{G_{\underline{X}}}$
- $(X_2 \perp\!\!\!\perp X)_{G_{\underline{X_2}}}$

From the do-calculus rules from [Pearl, 2009] we can deduce that the distribution of Y under the condition $X_2 = 0$ intervening on X to fix its distribution to $X \sim X|do(X_2 = 1)$ is given by

$$f(Y|X2 = 0, do(X \sim X|do(X2 = 1))) = \int_X \int_{X_1} f_{Y|X,X_1,X_2}(x, x_1, 0) f_{X_1}(x_1) f_{X|X_2}(x, 1) Pr(X_2 = 1) dx_1 dx \tag{C.1}$$

Appendix C.3. Ideal situation

In this case we generate our first artificial dataset as follows:

$\mathbf{X_1}$ = random_re-sampling(Throughput)

$\mathbf{X_2}$ = random_re-sampling(DNS)

$$\mathbf{X} \sim \begin{cases} \Gamma(k_1, \theta_1) + \sqrt{X_1} * \frac{\mu_1}{2} + \varepsilon : X_2 = 0 \\ \Gamma(k_2, \theta_2) + \sqrt{X_1} * \frac{\mu_2}{2} + \varepsilon : X_2 \neq 0 \end{cases}$$

$$\mathbf{Y} \sim \begin{cases} 10.\sqrt{5.X + 10.X_1} + \varepsilon : X_2 = 0 \\ 25.\sqrt{3.X + 6.X_1} + \varepsilon : X_2 \neq 0 \end{cases}$$

with ε representing an error term.

In this first case we chose the following values $\{k_1 = 5, \theta_1 = 1.0, k_2 = 2.0, \theta_2 = 2.0, \mu_1 = 5, \mu_2 = 8\}$. The resulting distributions are presented in the Figure C.14. To make sure that the parameters are correctly generated, we infer the corresponding causal model using the PC algorithm [Spirtes and Glymour, 1991] with the independence test from [Zhang et al., 2012].

Notice that, for testing our method under the same main constraint we limit our sample size to 10000 (against 7500 in the real case scenario) and we keep the ratio between the number of samples where $X_2 = 0$ is observed and the number of samples where $X_2 = 1$ is observed equal to the ratio between the number of connections where the ISP DNS service was observed (80%) and the number of connections where the Google DNS service was observed (20%) in the real case scenario.

Appendix C.3.1. Prediction of Y given X2 = 0 after intervention on X giving it the distribution of X given X2 = 1

Using the Equation (C.1) we are able to compute the PDF $f(Y|X_2 = 0, do(X \sim X|X_2 = 1))$ and obtain the expected value $\mathbb{E}[Y|X_2 = 0, do(X \sim X|X_2 = 1)]$.

As mentioned in Appendix B.2.3, the choice of the number of bins and the threshold to decide or not to estimate the post atomic intervention PDF, should have an impact on the prediction accuracy. Consequently, in this ideal scenario where the two distributions of $X|X_2 = 0$ and $X|X_2 = 1$ have very similar domains, we vary these two parameters and study their impact on the prediction accuracy.

To obtain the distribution of $f_{Y|X_2=0,do(X\sim X|X_2=1)}$ we generate X and Y as following:

$$\begin{aligned} \mathbf{X} &\sim \Gamma(k_2, \theta_2) + \sqrt{X_1} * \frac{\mu_2}{2} + \varepsilon : X_2 = 0 \\ \mathbf{Y} &\sim 10.\sqrt{5.X + 10.X_1} + \varepsilon \end{aligned} \tag{C.2}$$

We obtain an expected value of $\mathbf{E[Y|X_2 = 0, do(X \sim X|X_2 = 1)]}$ of 101.5.

We summarize the different results we obtained in Table C.4.

We can see that the use of a T-copula (*T-cop*) in the modeling of the multi dimensional PDF gives slightly better results, in terms of accuracy, than the modeling of multidimensional PDF with a Gaussian copula (*G-cop*). An interesting advantage of the T-copula comes from its ability to capture tail dependencies between the different components of the multivariate distribution.

If X is a d-dimensional random vector following a multivariate t -distribution with ν degrees of freedom, mean vector μ and a positive-definite dispersion Σ, denoted $X \sim t_d(\nu, \mu, \Sigma)$, [Demarta and McNeil, 2005] showed that the tail dependency coefficient is given by:

$$\lambda = 2t_{\nu+1}(\sqrt{\nu + 1}\sqrt{1\rho}/\sqrt{1 + \rho}) \tag{C.3}$$

where ρ is the off-diagonal element of the correlation matrix implied by the normalization of the scatter matrix Σ.

This result shows that, by tuning the different parameters of the T-copula we can better capture the tail dependencies between the different components of the multi-variate distribution we want to

Table C.4: Effect of varying Δ_S on the prediction accuracy for the first artificial dataset for two different multi dimensional PDF modeling, using a T-copula (*T-cop*) and Gaussian copula (*G-cop*). *#AI* stands for the number of Atomic Interventions

Δ_S	$\hat{\mathbf{E}}[Y\|X_2 = 0, do(X \sim X\|X_2 = 1)]$		Error		#A.I.	#Failures	
	G-cop	T-cop	G-cop	T-cop		G-cop	T-cop
20	93.80	96.22	7.8%	5.5%	195	0 (0%)	5 (2.6%)
30	93.28	95.90	8.6%	6.0%	118	0 (0%)	3 (2.5%)
50	95.40	96.79	6.5%	5.2%	76	0 (0%)	2 (2.6%)
70	94.41	96.94	7.5%	5.0%	54	0 (0%)	1 (1.9%))
100	94.24	96.74	7.6%	5.2%	39	0 (0%)	0 (0 %)

model. This property is really interesting when modeling communication networks performance, namely the throughput, as it is often the case that we find dependencies in the tail of parameter distributions, such as the one of delay or loss, with the throughput. As stated previously, the counter part of the T-copulae in practice comes from the higher sensitivity to data shortage.

Very likely due to the functions used for generating the artificial dataset and the use of a Gamma distribution, the T-copula is not always able to model the PDFs $f_{Y|X,X_1,X_2}(y, x, x_1, 0)$ that are necessary to compute the post intervention PDF of Y in Equation (C.1).

It is also important to notice that the choice of using a Gaussian copula was motivated by the prediction of an intervention in the opposite case ($f_{Y|do(X\ X|X_2=0),X_2=1}(y)$), see next section. In addition, we generated the artificial dataset in order to have the same domains for $X|X_2 = 0$ and $X|X_2 = 1$ and do not observe the same tail dependence as we would have for the throughput in the real case. From this perspective, the usage of T-copula is not fully justified and the usage of a Gaussian copula, for this particular dataset, should still be preferred.

Appendix C.3.2. Prediction of Y given X2 = 1 after an intervention on X giving it the distribution of X given X2 = 0

Without repeating the explanations given in the previous section, we use Equation (C.1) to predict the expected value of $\mathbf{E}[Y|X_2 = 1, do(X \sim X|X_2 = 0)]$.

The results are presented in Table C.5 in which we compared the expected value obtained using Equation (C.1) and G-copulae or T-copulae with the value obtained from the definitions of the parameters, Equation (C.2). We also study the impact of the number of samples used for estimating the atomic post-intervention distribution, Δ_S. Note that by increasing the minimum number of samples required to estimate a given atomic intervention we decrease the number of atomic interventions used for predicting the distribution of $f(Y|X_2 = B, do(X \sim X|X_2 = \neg B)$ (with $B \in 0, 1$), represented by the parameter **#AI**.

Several important remarks can be made from the results we obtain:

- The usage of a T-copula for modeling the PDF of $f_{Y|X,X_1,X_2}(y, x, x_1, 0)$ for different values of X fails more than 50% of the times.
- The utilization of wider bins, gathering more data to approximate $X = x$, does not improve the the success rate of the predictions of the PDFs post-intervention.

Table C.5: Effect of varying Δ_S on the prediction accuracy for the first artificial dataset for two different multi dimensional PDF models, using a T-copula (*T-cop*) and Gaussian copula (*G-cop*). *#AI* stands for the number of Atomic Interventions

Δ_S	$\hat{\mathbf{E}}[Y\|X_2 = 1, do(X \sim X\|X_2 = 0)]$		Error		#A.I.	#Failures	
	G-cop	T-cop	G-cop	T-cop		G-cop	T-cop
20	173.9	N.A.	2.5%	N.A.	43	0 (0%)	23 (53%)
30	173.9	N.A.	2.5%	N.A.	32	0 (0%)	20 (63%)
50	175.7	N.A.	1.4%	N.A.	17	0 (0%)	13 (76%)
70	175.7	N.A.	1.6%	N.A.	13	0 (0%)	10 (77%)
100	174.0	N.A.	2.3%	N.A.	9	0 (0%)	8 (89%)

- The usage of a G-copula gives better results (in terms of prediction accuracy) than the previous predictions of interventions.

Remarks. It should be noticed that, despite the apparent symmetry between the prediction of Y conditionally to $X_2 = 1$ if we perform an intervention on X where we fix its distribution to the one of $X|X_2 = 0$ and the one of the prediction of the value of Y conditionally to $X_2 = 0$ if we perform an intervention on X where we fix its distribution to the one of $X|X_2 = 1$, the problematic is different. From an external point of view, looking at the completion time and success rate, Gaussian copulae are less data demanding. We generated X_2 by randomly re sampling the DNS parameters from our real dataset. Doing so, we have 80% of the samples where $X_2 = 0$ is observed against 20% where $X_2 = 1$ is observed.

This second scenario shows the sensitivity of our approach to resource limitation. Even in this "optimal scenario" where both conditional PDF of $X|X_2 = 0$ and $X|X_2 = 1$ have the same domains, the shortage of data for the second conditional PDF prevents us from using a model which could be more accurate (T-copula). This can be seen when comparing the predictions made in this section with the ones made when we had more data (Appendix C.3.1) where the T-copula model gave more accurate predictions.

Appendix C.3.3. Concluding remarks

In this section we presented the optimal case where we have both conditional PDFs having almost perfectly overlapping domains. We tried to predict interventions where we condition on one value of the categorical parameter, X_2, and intervene on another parameter, X, and fix its distribution to follow the conditional distribution corresponding to the complementary value of the categorical parameter. The results of this study represent two important findings:

- Our method works and makes an accurate prediction (error $< 3\%$) when enough data is present
- The use of T-copulae better captures the multi dimensional PDFs dependences but fails as soon as the data becomes scarcer, while G-copulae still give an accurate prediction.

Again, these conclusion are made using an artificial dataset (see definition in Appendix C.3) where we could not faithfully mimic the tail dependencies of the parents of the throughput that

Table C.6: Effect of varying Δ_S on the prediction accuracy for the second artificial dataset for two different multi dimensional PDF modeling, using a T-copula (*T-cop*) and Gaussian copula (*G-cop*). *#AI* stands for the number of Atomic Interventions

Δ_S	$\hat{\mathbf{E}}[Y\|X_2 = 0, do(X \sim X\|X_2 = 1)]$		Error		#A.I.	#Failures	
	G-cop	T-cop	G-cop	T-cop		G-cop	T-cop
20	96.73	97.12	4.7%	4.3%	195	0 (0%)	5 (2.6%)
40	95.97	96.13	5.5%	5.3%	97	0 (0%)	2 (2.1%)
60	96.39	96.69	5.1%	4.8%	60	0 (0%)	1 (1.7%)
80	95.16	95.87	6.3%	5.6%	48	0 (0%)	1 (2.1%)
100	94.24	95.05	7.2%	6.4%	39	0 (0%)	0 (0.0%)

we observed in our real case scenario nor the variability of the observed values. These choices are inherent to the design of an artificial dataset and should not be seen as a limitation of the presented method.

Appendix C.4. Removing samples for $X \mid X_2 = 1$ outside of the zone of the concentration of the distribution $f_{X|X_2=0}$

To estimate the impact of the absence of some values in both domains of the conditional PDFs of $X|X_2$, we remove some samples from the dataset where $X_2 = 1$. For this second artificial dataset, we do not remove samples in the domain where the distribution $f_{X|X_2=0}$ takes its biggest values. Figure C.16 represents the original distributions of $X|X_2$ (our first artificial dataset) and Figure C.17 presents the resulting distributions of $X|X_2$ after removing samples from the distribution of $X|X_2 = 1$ (we remove the samples for $X_2 = 1$ where $X \in [0, 5] \cup [15, 20]$).

As in Appendix C.3.1 and Appendix C.3.2, we estimate the expected value of Y conditionally to $X_2 = 0$ when we intervene on X and we fix its distribution to the one of $X \sim X|X_2 = 1$. We also estimate the expected value of Y conditionally to $X_2 = 1$ when we intervene on X and we fix its distribution to the one $X \sim X|X_2 = 0$

We first estimate the effect of intervening on X to fix its distribution to the one of $X \sim X|X_2 = 1$. We can see from Table C.6 that the conclusions drawn previously are still valid when we remove samples from the distribution of $X|X_2 = 1$. We can make the following remarks:

- The precision decreases with the increase of Δ_S and the decrease of the number of atomic interventions,
- The number of failures of the multidimensional PDF modeling using a T-copula is sensibly similar to the number of failures that were observed for the first dataset.

When looking at the opposite case (conditioning on $X_2 = 1$ and giving to X the distribution of $X \sim X|X_2 = 0$) we can observe that the modeling of the conditional PDFs using a T-copula fails often and prevents the estimation of the post-intervention PDF, see Table C.7. The estimation of conditional PDF when we use G-copulae gives very good results.

Table C.7: Effect of varying Δ_S on the prediction accuracy for the second artificial dataset for two different multi dimensional PDF modeling, using a T-copula (*T-cop*) and Gaussian copula (*G-cop*). *#AI* stands for the number of Atomic Interventions

Δ_S	$\hat{\mathbf{E}}[Y\|X_2 = 1, do(X \sim X\|X_2 = 0)]$		Error		#A.I.	#Failures	
	G-cop	T-cop	G-cop	T-cop		G-cop	T-cop
20	176.63	N.A.	0.6%	N.A.	40	0 (0%)	23 (57.5%)
40	176.95	N.A.	0.4%	N.A.	22	0 (0%)	14 (63.6%)
60	178.21	N.A.	0.3%	N.A.	12	0 (0%)	10 (83.3%)
80	179.73	N.A.	1.2%	N.A.	10	0 (0%)	9 (90.0%)
100	177.43	N.A.	0.1%	N.A.	9	0 (0%)	8 (88.9%)

Appendix C.4.1. Concluding remarks

In this scenario we removed samples from our initial dataset to study the impact of data shortage on our method precision. In this case, we removed samples from the conditional distribution $X_2 = 1$ corresponding to x values where the conditional probability distribution $f_{X|X_2=0}$ takes small values. We predicted the effect on $Y|X_2 = b$ when intervening on $X|X_2 = b$ and enforcing this intervention with probability $f_{X|X_2=\bar{b}}$, where b can take the value 0 or 1 and $\bar{b}$ is b complementary.

The prediction of this intervention consists in *i)* Predicting the effect, on Y, of the atomic intervention $f_{Y|do(X=x),X_2=b}$ *ii)* Assign to each atomic intervention a probability being $f_{X|X_2=\bar{b}}$.

$$f(Y|do(X \sim X|do(X_2 = \bar{b}), X_2 = b) = \int_X f(Y|do(X = x), X_2 = b) f(X = x|X_2 = \bar{b}) dx \qquad \text{(C.4)}$$

Therefore, the prediction is more complex in the case where we give to $f_{Y|do(X=x),X_2=b}(y, x)$ a probability $f_{X|X_2=\bar{b}}(x) > 0$ but no value is actually observed for $X = x|X_2 = b$.

This appears in the second case where we want to predict the distribution $f_{Y|X_2=1,do(X\sim X|X_2=0)}$. As there are x values for which $X|X_2 = 1$ is not observed, we cannot enforce the conditional distribution $f_{Y|do(X=x)|X_2=1}(y, x)$ with the distribution $f_{X|X_2=0}(x)$.

Nevertheless, because we removed observation corresponding to x values where $f_{X|X_2=0}(x)$ takes small values, the impossibility to compute the result of Equation (C.4) for some x values has a relatively small impact on the estimation of the overall post-intervention distribution and our method keeps performing well.

The next section presents a more complex scenario where we remove samples in intervals where the PDF of $X|X_2 = 0$ takes important values (values of x where $X|X_2 = 0$ has a high probability).

Appendix C.5. Removing samples $\boldsymbol{X \mid X_2 = 1}$ at the zone of the concentration of the distribution $\boldsymbol{f_{X|X_2=0}}$

We generate a third artificial dataset with a distribution $X|X_2$ as represented in Figure C.18. For this new dataset, we remove samples from the domain where the distribution $f_{X|X_2=0}$ takes its biggest values. The results for the two interventions are presented in Tables C.8- C.9 with:

- The prediction of expected value of Y after intervention corresponding to PDF of Y: $f_{Y|X_2=0,do(X\sim X|X_2=1)}(y)$ in Table C.8

Table C.8: Effect of varying Δ_S on the prediction accuracy for the third artificial dataset for two different multi dimensional PDF modeling, using a T-copula (*T-cop*) and Gaussian copula (*G-cop*). *#AI* stands for the number of Atomic Interventions

Δ_S	$\hat{\mathbf{E}}[Y\|X_2 = 0, do(X \sim X\|X_2 = 1)]$		Error		#A.I.	#Failures	
	G-cop	T-cop	G-cop	T-cop		G-cop	T-cop
20	66.79	80.81	34.2%	20.4%	195	0 (0%)	5 (2.6%)
40	76.15	85.80	25.0%	15.5%	97	0 (0%)	2 (2.1%)
60	67.07	82.14	33.9%	19.1%	60	0 (0%)	1 (1.7%)
80	69.55	83.08	31.5%	18.2%	48	0 (0%)	1 (2.1%)
100	N.A.	N.A.	N.A.	N.A.	39	0 (0%)	0 (0.0%)

Table C.9: Effect of varying Δ_S on the prediction accuracy for the third artificial dataset for two different multi dimensional PDF modeling, using a T-copula (*T-cop*) and Gaussian copula (*G-cop*). *#AI* stands for the number of Atomic Interventions

Δ_S	$\hat{\mathbf{E}}[Y\|X_2 = 1, do(X \sim X\|X_2 = 0)]$		Error		#A.I.	#Failures	
	G-cop	T-cop	G-cop	T-cop		G-cop	T-cop
20	153.68	N.A.	13.5%	N.A.	12	0 (0%)	7 (58.3%)
40	154.50	140.18.	13.0%	21.1%	6	0 (0%)	3 (50.0%)
60	157.90	N.A.	11.1%	N.A.	5	0 (0%)	3 (60.0%)
80	163.52	N.A.	8.0%	N.A.	3	0 (0%)	2 (66.7%)
100	157.25	N.A.	11.5%	N.A.	2	0 (0%)	1 (50%)

- Prediction of expected value of Y after intervention corresponding to PDF of Y: $f_{Y|X_2=1,do(X\sim X|X_2=0)}(y)$ in Table C.9

Appendix C.5.1. Concluding remarks

We can observe the accuracy of the prediction of the expected value of Y conditionally to X_2 when setting X distribution to the one of $X|X_2 = 1$ is highly impacted by the absence of samples in the zone where the PDF $f_{X|X_2=0}$ takes important values. The usage of a T-copula for conditional PDF gives slightly better results than the predictions based on the usage of a Gaussian copula. However with 20% error rate, we cannot use our method any more.

For the prediction of the expected value of Y conditionally to $X_2 = 1$ when intervening on X and fixing its distribution to $f_{X|X_2=0}(x)$, we penalize the usage of T-copula for modeling conditional PDFs more than the G-copula. Gaussian copulae seem to require less data to estimate a parameterization offering an acceptable modeling of the dependencies between the marginals of the multivariate distribution we need to estimate our model.

Appendix C.6. Conclusion

The questions we wanted to answer were:

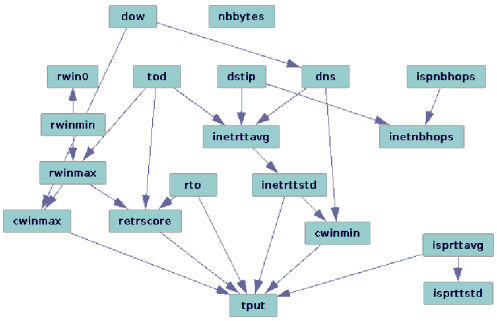
- How to choose the copula family that will model the dependencies between the different marginals ?
- In which proportion the absence of values observed for both conditional distributions impacts the prediction accuracy and how to improve the accuracy in this case ?
- Should we favor the number of atomic predictions from which the final distribution is estimated or the quality of the atomic intervention predictions by increasing the number of samples from which these atomic predictions are computed ?

The first question can be answered by "*data dictates the choice*". In our case, as we are working with a limited amount of data, Gaussian copulae are used to capture the dependencies between the marginals of the multivariate distributions we need to estimate to predict the effect of interventions on parameters impacted by the DNS parameter.

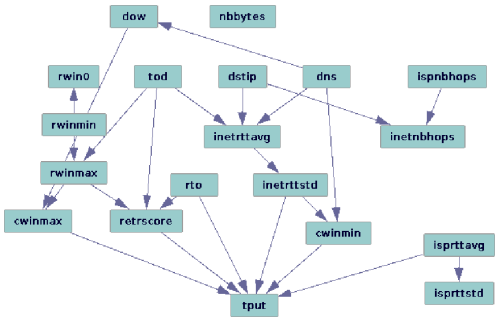
The absence of values observed for both conditional distributions can be overcome using Gaussian copulae if we observe enough values in the zones corresponding to high probabilities for the conditional distributions. T-copulae suffer more from data shortage than G-copula but if the observations of the conditional distributions are too sparse then both models become inaccurate and cannot be used.

The number of atomic interventions should be preferred to the number of samples used for estimating a given intervention but a minimum number of samples should be present (≥ 30).

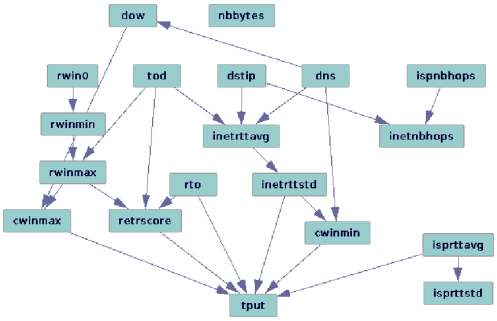
Given these conclusions, we have defined and parameterized the methods that can be used to study the impact of DNS on CDN performance.



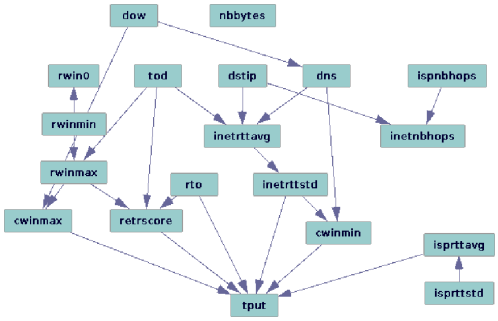
(a) First member



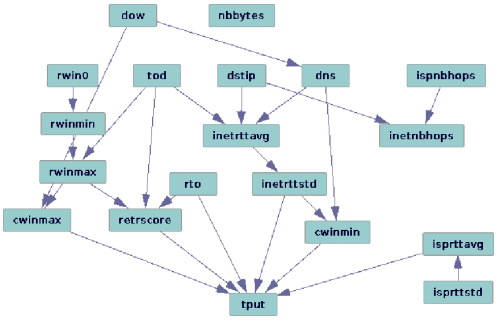
(b) Second member



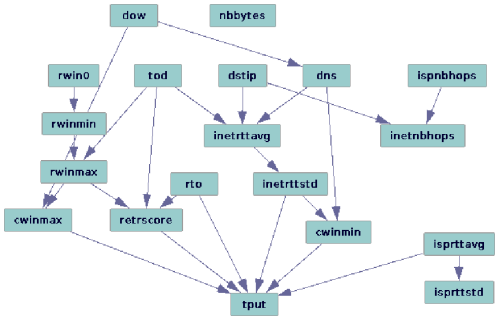
(c) Third member



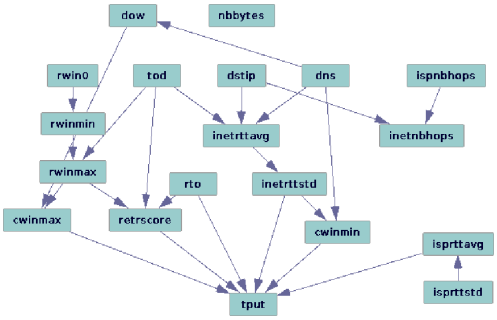
(d) Fourth member



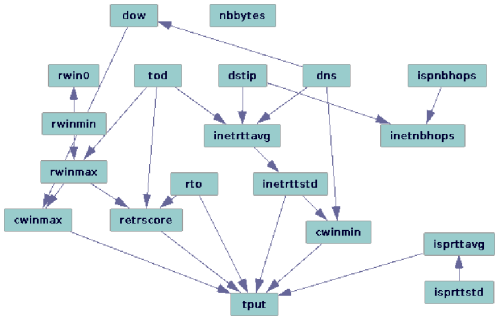
(e) Fifth member



(f) Sixth member

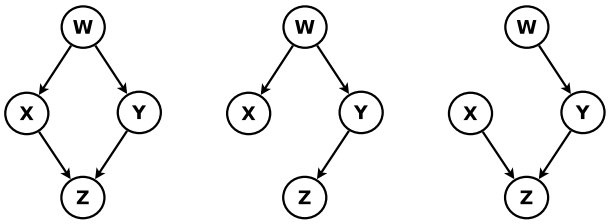


(g) Seventh member



(h) Eighth member

Figure A.11: The eight members of the Markov Equivalence Class corresponding to the set of independences that were detected from our observations. These graphs were obtained using the Tetrad software [Spirtes et al., 2001]



(a) G (b) $G_{\underline{X}}$ (c) $G_{\overline{X}}$

Figure B.12: Illustration of the different subgraphs $G_{\underline{X}}$ and $G_{\overline{X}}$ for a Bayesian network representing the causal model of a four parameter system $\{W,X,Y,Z\}$

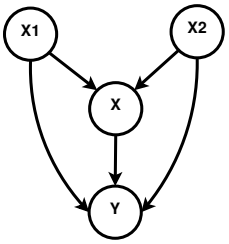


Figure C.13: Artificial dataset dependencies

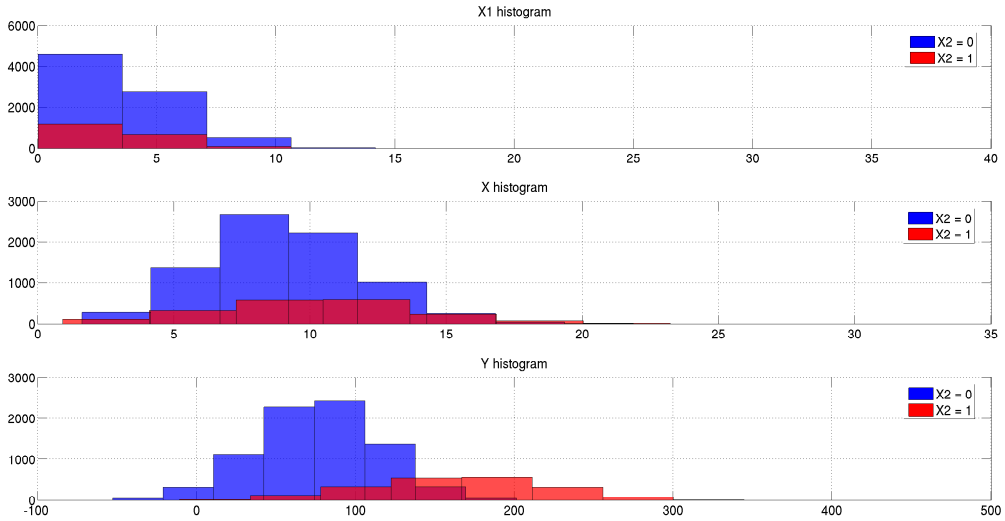


Figure C.14: Distribution of the different parameters for both values of X_2

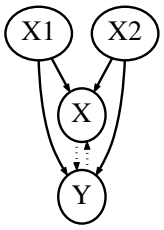


Figure C.15: Causal Model of the first dataset

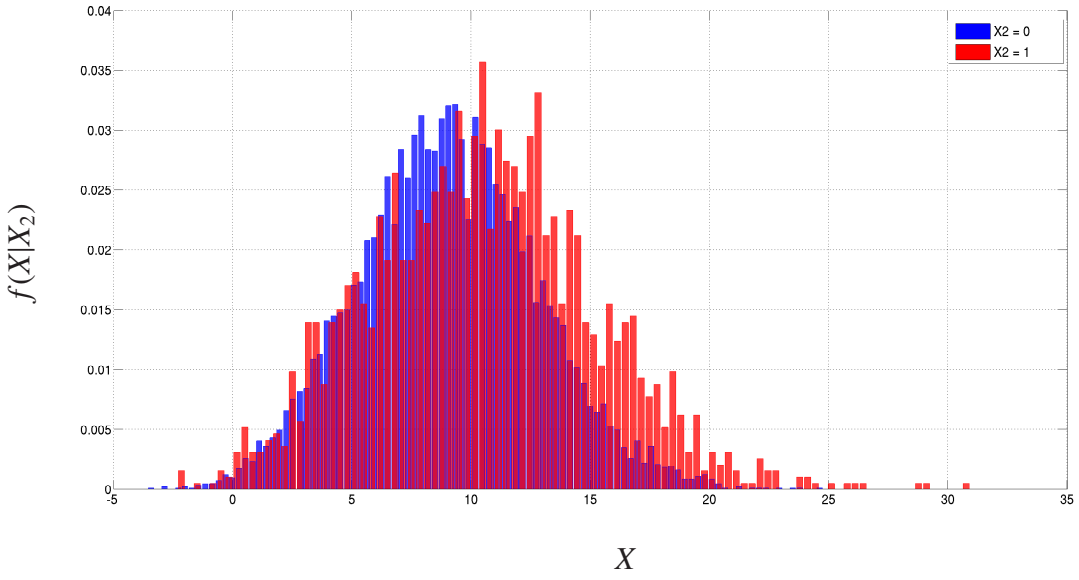


Figure C.16: Conditional probability density function of X conditional on X_2 in the original dataset

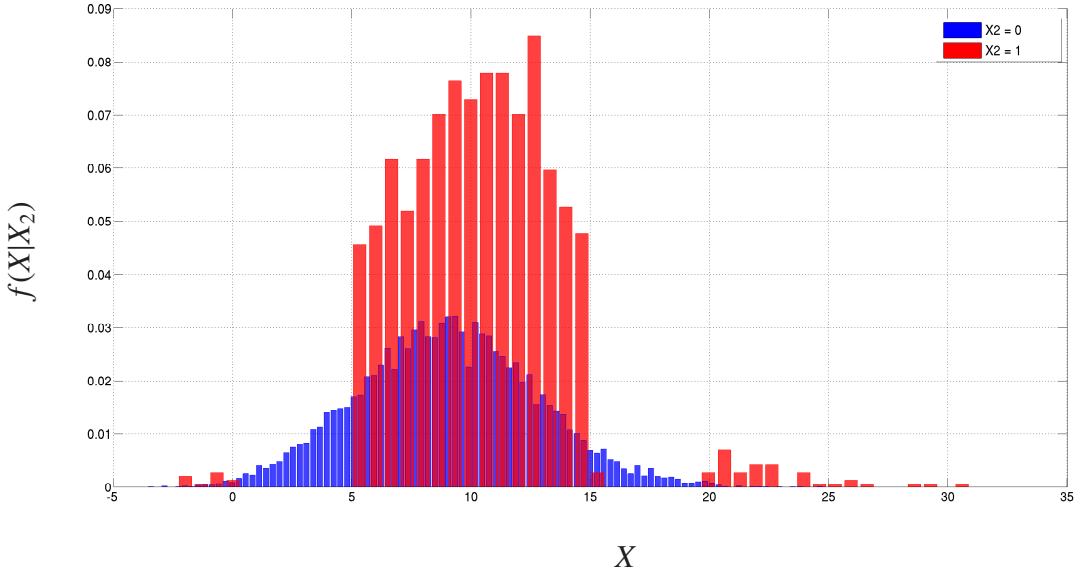


Figure C.17: Conditional probability density function of X conditional on X_2 after removing samples from $X_2 = 1$ in the domain of $X|X_2 = 0$ where the distribution $f_{X|X_2=0}$ is not taking high values, second artificial dataset

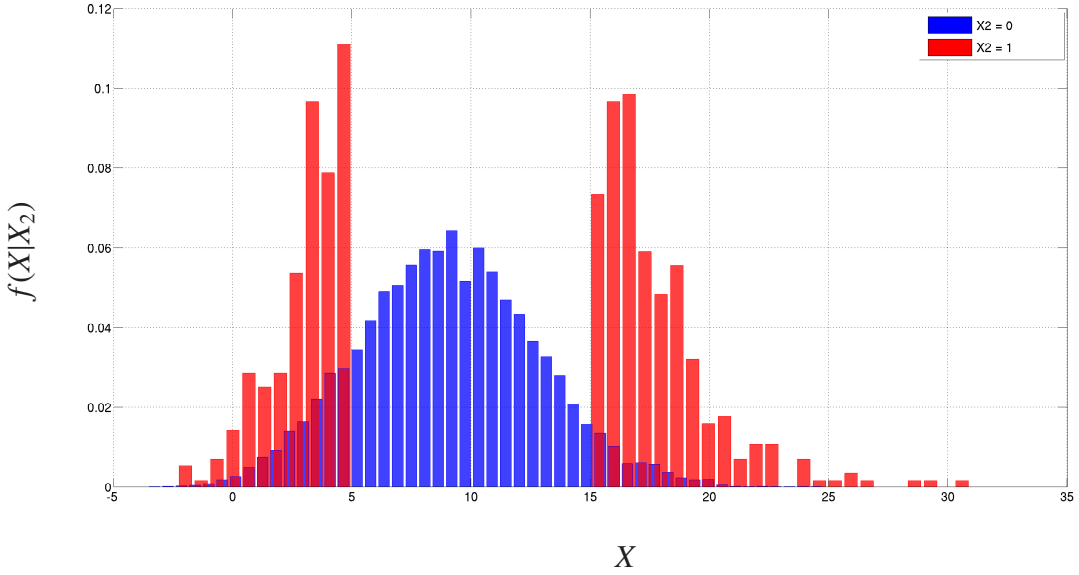


Figure C.18: Conditional probability density function of X conditional on X_2 after removing samples from $X_2 = 1$ in the domain of $X|X_2 = 0$ where the distribution $f_{X|X_2=0}$ is concentrated, third artificial dataset