A causal study of an emulated network

Hadrien Hours, Ernst Biersack, Patrick Loiseau

EURECOM, Campus SophiaTech
Route des Chappes, 06410 Biot, France
firstname.lastname@eurecom.fr

1. Introduction

With the diversity of applications and technologies present in communication networks, the study of their performance has become a complex issue. The large number of parameters to take into account to model communication network performance increases the risk of spurious associations between the explanatory variables. Therefore studies based on correlation become a challenging approach. On the other hand, causal models, and their representation as Directed Acyclic Graphs (DAGs), offer simple and attractive models. Using simple graphical criteria, one can predict interventions on the system by using data collected through passive measurements. In this paper, we present an example of a causal study of communication network performance. We place ourselves in an emulated environment, where our predictions can be verified, and we show the benefits of the approach.

2. Causal study of an emulated network

For this paper we study the performance of TCP observing FTP traffic. We focus on the network performance, represented by the TCP throughput. To be able to verify our predictions we emulate a network using the Mininet software [1]. The experiment consists in one client that downloads files from a FTP server, both machines are connected to different routers (**R1** and **R2** respectively). In order to create a more realistic scenario we add two other machines, one connected to **R1** and the other to **R2**, creating cross traffic. The FTP traffic is recorded at the server side using *Tcpdump* tool and the dataset that we obtain is presented in Table 1.

Using the PC algorithm [4] with the Hilbert Schmidt Independence Criterion (HSIC) [5], we obtain, from the data summarized in Table 1, the causal model represented Figure 1. While TCP is a feedback based protocol, it is important to notice that we are studying the causal relationships between the parameters defining our system. In this case we obtain a causal graph, represented as a DAG, where the causes of the throughput are disclosed and its direct causes are represented as its parents. In this model we observe the receiver window (*rwin*), delay (*rtt*), and loss (*retrscore*), the empirical causes of the throughput, that are, indeed, found as its direct parents. Due to space constraints we cannot describe all the properties of this model and focus on prediction.

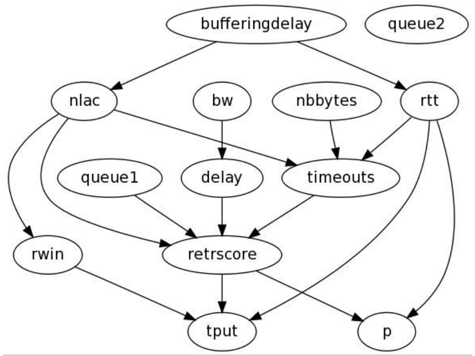

FIGURE 1 – Causal model of emulated FTP traffic

We want to estimate the distribution of the *throughput* after an intervention on the *retrscore*, where its value is set to 1%.

To estimate the total effect, on a parameter Y, of an intervention on a parameter X, we use the *Back-door Criterion* [3] defined as follows.

Definition 1 ***(Back-door criterion)*** *A (set of) variable(s) Z is said to satisfy the back-door criterion relative to an ordered pair of variables (X_i,X_j) in a DAG G if : (i) no node in Z is a descendant of X_i ; and (ii) Z blocks every path between X_i and X_j that contains an arrow into X_i.*

Using this definition, we can predict the effect of an intervention as follows.

Theorem 1 ***(Back-door adjustment)*** *If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and given by the formula :*

$$P(y \mid do(X = x)) = \sum_z P(Y = y \mid X = x, Z = z)P(Z = z).$$

where the $do()$ operator represents the intervention of setting the parameter X to the value x

We estimate the probability density functions (pdfs) by first estimating the marginals using normal kernels and then modeling the multivariate and conditional pdfs using T-Copulae [2].

To predict the effect of an intervention on the *retrscore* parameter, using the causal model presented in Figure 1, we apply the Back-door adjustment formula (Theorem 1) with the blocking set $Z = \{\mathtt{nlac}, \mathtt{rtt}\}$.

Parameter	Definition	Min	Max	Avg	CoV
bw	minimum bandwidth (Mbps)	1	25	7.1	0.91
delay	propagation delay (ms)	30	180	86	0.48
queue1	size of **R1** buffer (pkts)	10	400	98	1.1
queue2	size of **R2** buffer (pkts)	10	400	100	0.99
nlac	Narrow Link Available Capacity (Kbps)	12	3.07e3	630	5
rwin	**C1** advertised receiver window (KB)	74	2.23	290	0.65
bufferingdelay	part of the RTT due to queuing delay (ms)	1	6.76e3	120	2.4
rtt	Round Trip Time (ms)	84	6.91e3	3.1e2	0.99
timeouts	number of timeouts (units)	0	682	79	1.5
retrscore	fraction of retransmitted packets (no unit)	0	0.61	3.7e-3	5.1
p	fraction of loss events (no unit)	0	0.64	3.8e-3	8.4
nbbytes	number of bytes sent by the server (MB)	6	150	110	0.21
tput	throughput (Kbps)	6	1.10e3	280	0.81

TABLE 1 – Summary of Mininet network emulation experiments dataset

To verify our prediction we set up a new experiment where we drop 1% of the packets at the router **R1** and compare the throughput we obtain with the one that was predicted using the equation from Theorem 1. Figure 2 presents the probability density functions corresponding to the throughput prior to intervention (in dash-dot line), the post-intervention throughput estimated with the Back-door criterion (dashed line) and the throughput observed when we manually modify the loss (solid line). While not perfectly matching, the graph-based prediction (dashed line) and the real density after manual intervention (solid line) show similar shapes. In particular, both densities have a single mode around the same value (~ 100 kbps). The fact that the experimental throughput values after intervention (solid line) show less variance can be explained by the small number of samples (20) used to make the estimation. Also our method to estimate the post-intervention throughput (dashed line) uses normal kernels and a T-copula which tend to widen and smooth the estimated post-intervention density.

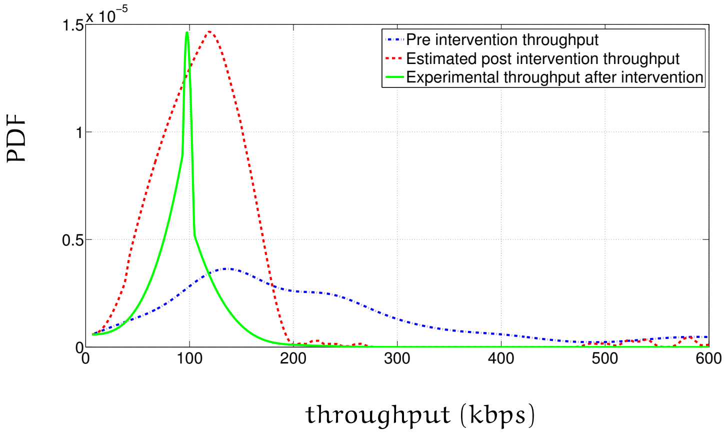

FIGURE 2 – Throughput distribution after an intervention on the retransmission score

3. Concluding remarks

Due to space constraints, many aspects of this study could not be presented. The study of the causal model we obtain discloses many properties of the system that would need further explanations and would lead to future studies. However it should be noticed, from the Back-door adjustment equation (Theorem 1) that the values present in our original dataset dictate the range of predictions that can be made. We are now working on adopting parametric modeling of the probabilities of the parameters that would allow to overcome this limitation.

References

1. N. Handigol, B. Heller, V. Jeyakumar, B. Lantz, and N. McKeown. Reproducible network experiments using container-based emulation. In *CoNEXT '12*, pages 253–264.
2. P. Jaworski, F. Durante, W.K. Härdle, and T. Rychlik. *Copula Theory and Its Applications*. Lecture Notes in Statistics. Springer, 2010.
3. J. Pearl. *Causality : Models, Reasoning and Inference*. Cambridge University Press, 2009.
4. P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, Vol. 9 :62–72, 1991.
5. K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. *CoRR*, abs/1202.3775, 2012.