Computer Networks 000 (2016) 1–11

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

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[d]

[a] Eurecom, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France
[b] Ecole Normale Superieure de Lyon - Site monod, 46 Alle d'Italie - 69007 Lyon, France
[c] CAIPY, 06560 Valbonne, Fance
[d] Torre Telefonica Diagonal 00, Plaza de Ernest Lluch i Martin, 5, 08019 - Barcelona, Spain
[e] Politecnico di Torino - Corso Duca degli Abruzzi, 24 - 10129 Torino, Italy

ARTICLE INFO

Article history:
Received 30 September 2015
Revised 9 June 2016
Accepted 17 June 2016
Available online xxx

Keywords:
Causality
CDN
DNS
Traffic performance
Bayesian networks
Optimization
Monitoring
Knowledge inference
Computer networks

ABSTRACT

Resources such as Web pages or videos that are published in the Internet are referred to by their Uniform Resource Locator (URL). If a user accesses a resource via its URL, the host name part of the URL needs to be translated into a routable IP address. This translation is performed by the Domain Name System service (DNS). DNS also plays an important role when Content Distribution Networks (CDNs) are used to host replicas of popular objects on multiple servers that are located in geographically different areas. A CDN makes use of the DNS service to infer client location and direct the client request to the optimal server. While most Internet Service Providers (ISPs) offer a DNS service to their customers, clients may instead use a public DNS service. The choice of the DNS service can impact the performance of clients when retrieving a resource from a given CDN. In this paper we study the impact on download performance for clients using either the DNS service of their ISP or the public DNS service provided by Google DNS. We adopt a causal approach that exposes the structural dependencies of the different parameters impacted by the DNS service used and we show how to model these dependencies with a Bayesian network. The Bayesian network allows us to explain and quantify the performance benefits seen by clients when using the DNS service of their ISP. We also discuss how the further improve client performance.

1. Introduction

Each time an Internet user wants to access a resource, he uses a human readable name called Uniform Resource Locator (URL), containing the domain name of the administrative entity hosting this resource. However, a domain name is not routable and needs to be translated into the IP address of a server hosting the resource the client wants to access. This is taken care of by the DNS service. At the same time, many popular services such as YouTube, iTunes, Facebook or Twitter, rely on CDNs, where objects are replicated on different servers, and in different geographical locations to optimize the performance experienced by their users. When a client accesses an object hosted by a CDN, its default DNS server contacts the DNS server of the CDN that hosts the resource the client is requesting. Based on the origin of the request, the authoritative CDN DNS redirects the client to the optimal server. Most of the ISPs provide a DNS service, but it is now common to see customers using a public DNS service instead [10]. Clients using the DNS service of their ISP are served by a local DNS server that often provides a more accurate location information to the CDN compared to the information communicated by a public DNS service such as the Google DNS service. Indeed, public DNS servers are usually further away from the clients of a given ISP than the default ISP DNS server. There have been several studies suggesting that public DNS services do not perform as well as local DNS services provided by ISPs, mainly because of the impossibility of public DNS to correctly communicate the location of the clients originating the request [1,7]. This problem is addressed with ECS (edns-client-subnet) [16] but Akamai does not support it currently.

Studying the performance of the users accessing resources in the Internet is a complex task. Many parameters influence the end user experience and the relationships between these parameters is not always observable or intuitive. It is therefore necessary to

* Corresponding author.
E-mail addresses: hadrien.hours@eurecom.fr, hadrien.hours@ens-lyon.fr (H. Hours), erbi@e-biersack.eu (E. Biersack), patrick.loiseau@eurecom.fr (P. Loiseau), alessandro.finamore@telefonica.com, alessandro.finamore@polito.it (A. Finamore), marco.mellia@telefonica.com (M. Mellia).

http://dx.doi.org/10.1016/j.comnet.2016.06.023

use a simple, yet formal model that allows us to understand the role of a given parameter and its dependencies with other parameters. Bayesian networks offer a simple and concise way to represent complex systems [2]. In this paper, we use a Bayesian network to represent the causal model that captures the impact of the DNS service on the throughput performance experienced by clients accessing resources hosted by the Akamai CDN. Bayesian networks capture the dependencies between the different parameters impacting the throughput of the clients. One very interesting property of causal models is their *stability under intervention*. Causal models can be used to predict how the throughput of CDN users would evolve if we would *intervene* on the different parameters influencing the performance of CDN users. Here an intervention consists in isolating a given parameter of the system being studied, removing all its direct and remote causes and fixing its variations to a pre defined value or distribution. Being able to predict the effect of interventions, we can use causal models to understand the observed performance of a given system and to design strategies to improve its performance. In this work, we infer and use the causal model of CDN performance to understand the impact of choosing one DNS service instead of another. From such a model we are able to explain why clients using the DNS service of their ISP experience better download performance than clients using the Google DNS service. We are also able to indicate how to further improve the performance of the clients using the DNS service of their ISP.

Our work differs from previous studies of DNS services in several important points:

- We use a causal approach that formally models the structural dependencies of the different parameters influencing the throughput obtained.
- Observing that the clients using the DNS service of their ISP (referred to as local DNS) experience higher throughput than the clients using the public DNS service (referred to as Google DNS), we can show that this performance difference is due to the fact that clients using the DNS service of their ISP are redirected to closer servers. We are also able to precisely quantify this performance improvement.
- The causal model of our system also reveals that the parameterization of TCP (initial congestion window) of the servers accessed by the users of the Google DNS plays a key role in their throughput performance. Besides fully explaining the observed performance, this result also indicates how to further improve the performance of the clients using the local DNS.

Overall, the main contribution of our work resides in the methodology adopted and in its use of counterfactuals to understand the causal dependencies of a complex system.

In Section 2, we introduce causal models and their use to predict interventions, summarizing some of the main concepts from [11,15]. We then present, in Section 3, the environment of our study and the description of the parameters constituting our system. Section 4 presents our study of the DNS impact on the throughput. In particular we present the causal model of our system where we can observe the impact of the choice of the DNS service on the throughput. Our approach also allows us to predict the improvement that could be achieved by modifying the parameterization of the servers accessed by the users of the local DNS service. Section 5 compares our approach to the related work and Section 6 summarizes our work and proposes directions to further extend our work.

Several methods mentioned in this paper were designed and validated with parallel studies that are presented in an Appendix. The Appendix is available with the online version of this paper.[1] We give references to these studies in the paper.

2. Causal model: Definitions and usage

To model a complex system such as a communication network and to organize the knowledge obtained from its passive observation is very challenging. Existing work typically looks for the presence of correlation between different events observed simultaneously (see [9] and references therein). However, *correlation is not causation and the detection of correlation between two parameters A and B does not inform us on how they are related. A* can impact *B*, or the other way around, or an unobserved parameter can impact both *A* and *B* simultaneously. The difference between correlation and causation plays an important role if we want to find out how to improve our system by partly modifying its behavior. A causal approach uncovers the structural dependencies between the parameters of the system under study. The ability to predict the effects of a manipulation of the parameters of a system is a major strength of causal models as they are *stable under intervention*. Stability under intervention means that a causal model, inferred from the observations of a system in a given situation, is still valid if we manually change the system mechanisms, redefining the systems laws. The manual modification of the system parameters is called an *intervention*. Interventions consist in modifying the behavior of a component of the system, removing the influence of its direct and remote causes and manually setting its variations. The inference of a causal model and of a causal effect [11,15] is made using passive observations only. The causal theory allows us to predict the behavior of the various parameters of the inferred model after an intervention *without the need of additional observations*.

In this section we present the PC algorithm [14] that is used to infer the causal model of our system. We also describe the different properties of a causal model as described in [11,15].

2.1. Causal model: Inference

For our work, we use the PC algorithm [14] to build the Bayesian graph representing the causal model of our system. This algorithm takes as input the observations of the different parameters that characterize our system and infers the corresponding causal model. In our representation of a causal model as a Bayesian network, each node represents one parameter of our system and the presence of an edge from a node X to a node Y ($X \rightarrow Y$) represents the existence of a causal dependence of parameter Y on parameter X.

The PC algorithm starts with a fully connected and unoriented graph, called *skeleton*, where each parameter is represented by a node and connected to every other parameter. The PC algorithm then trims the skeleton by checking for independencies between adjacent nodes:

- First, the unconditional independencies ($X \perp\!\!\!\perp Y$) are tested for all pairs of parameters and the edges between two nodes whose corresponding parameters are found to be independent are removed.
- For the parameters whose nodes are still adjacent, the PC algorithm then checks if there exists a conditioning set of size one that makes two adjacent nodes independent. If this is the case, it removes the edge connecting the corresponding two nodes, otherwise the edge is kept.
- The previous step is repeated, increasing the conditioning set size by one at each step, until the size of the conditioning set reaches the maximum degree of the current skeleton (the

[1] http://dx.doi.org/10.1016/j.comnet.2016.06.023