Internship proposal 2017/2018

Topic: Why is my Internet slow? Big data approach for diagnosing large delays on the Web

Duration: 4 to 6 months

Hosting team: MiMove, Inria Paris (https://mimove.inria.fr/)

Apply at: https://goo.gl/forms/CUYdaBmCA4iYaYpL2

Mentor:
Renata Teixeira, Directrice de Recherche, Inria (https://who.rocq.inria.fr/Renata.Teixeira/)
Jérémie Jakubowicz, Maître de conférences, Telecom Sud Paris (http://www-public.tem-tsp.eu/~jakubowi/)
Benoit Boireau, Chief Technical Officer, Ip-label

Keywords: Internet measurements, Web performance, network diagnosis, machine learning

Description:
Most people who use Internet services and applications in a daily basis feel the frustration when services are slow (e.g., when a website takes forever to load). A slow access affects the business of many online services. For example, the Bing search engine experiences reduced revenue of 4.3% with just a 2-second delay, and a 400-ms delay resulted in a 0.59% decrease in searches per user on the Google search engine. As a result, Internet services and applications put a lot of effort to reduce or mask delays (with the use of CDNs and a number of other optimizations). Still, problems such as flash crowds, poor WiFi in the users' home, or oversubscribed access links may temporarily increase delays. Users are often helpless when online services and applications are slow because problems can lie anywhere from the user's device to any of the many pieces that compose today's complex Internet services and applications. Our long-term goal in the BottleNet project[1] funded by the ANR is to provide tools to help unskilled users, content providers, and ISPs in diagnosing poor Internet performance.

The goal of this internship is to develop lightweight methods to first detect when Web services are slow and then pinpoint the root cause. We envision a data-driven approach, where we leverage datasets already available within the BottleNet project, in particular: (1) Fathom dataset, Fathom [1] is a Firefox extension that runs on the background on the user's browsers and collect periodically baseline measurements of the network performance such as the amount of cross traffic on the local host, the perceived WiFi signal quality, network delays to the home gateway and the access network, CPU and memory utilization, the system load, and page load times of few popular domains; and (2) Ip-label dataset, Ip-label[2] is an SME that provides solutions to measure QoE and QoS for all IP services, with a large customer base, ip-label collects periodic Web performance measurements for a number of popular websites.

Machine learning has already been successfully applied, e.g. to detect anomalies in data centers, to predict clicks on advertisement banners, to recommend content to users, among many other recent and less recent achievements. It is also increasingly popular in the domain of network analysis, be these networks physical or social. The goal of this internship is to evaluate the potential benefits of using machine learning techniques to predict when high delays are going to occur, and much more importantly to explain why. The "why" part can be addressed by considering that machine learning models are not black boxes, but should instead be investigated as first order objects. Statisticians are used to interpret generalized linear model parameters and their statistical relevance; however, interpreting other models such as support vector machines, forests and neural networks is known to be trickier. This internship will focus on: (i) assessing the predictive power of machine learning models on network delays and (ii) extract relevant insight from the models themselves to explain these delays and find their root cause when they appear.

The student should develop scientific skills on Internet measurements and troubleshooting as well as scientific writing and presentation. If the student is interested, there is a possibility of staying for the doctoral studies after the internship.

Desirable skills:
- Comfortable communicating in English
- Knowledge of network measurements
- Knowledge of data analysis techniques
- Knowledge of matlab or gnu R

References:

[1] https://project.inria.fr/bottlenet/
[2] http://www.ip-label.fr/

[1] M. Dhawan, J. Samuel, R. Teixeira, C. Kreibich, M. Allman, N. Weaver, and V. Paxson. Fathom: A Browser-based Network Measurement Platform. in Proc. of ACM Internet Measurement Conference, 2012.
[2] S. Clémençon and J. Jakubowicz: Scoring anomalies: a M-estimation formulation. AISTATS 2013