Activity Report 2016

Team MUSE

Measuring networks for enhancing USer Experience

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

Table of contents

Creation of the Team: 2013 October 03

Keywords:

Computer Science and Digital Science:
1.2.3. - Routing
1.2.4. - QoS, performance evaluation
1.2.5. - Internet of things
3.1.7. - Open data
3.1.8. - Big data (production, storage, transfer)
3.3. - Data and knowledge analysis
3.5. - Social networks

Other Research Topics and Application Domains:
6.3. - Network functions
6.4. - Internet of things
6.5. - Information systems

1. Members

Research Scientists
Renata Cruz Teixeira [Team leader, Inria, Senior Researcher, HDR]
Vassilis Christophides [Inria, Advanced Research position]
Anna-Kaisa Pietilainen [Inria, Starting Research position, until Sep 2016]

Faculty Member
Timur Friedman [Univ. Paris VI, Associate Professor, until Aug 2016]

Engineer
Michail Katsarakis [Inria, University of Crete, until May 2016]

PhD Students
Maximilian Bachl [Univ. Paris VI, from Oct 2016]
Sara El Aouad [Inria/Technicolor, granted by CIFRE, from May 2014]
Diego Neves Da Hora [Inria/Technicolor, granted by CIFRE, from February 2014]
Giuseppe Scavo [Inria, granted by Alcatel-Lucent Bell Labs, from November 2013]

Post-Doctoral Fellow
Francesco Bronzino [Inria, from Dec 2016]

Administrative Assistant
Anna Bednarik [Inria]

Others
Arash Molavi Kakhki [Inria, Ph.D Student, Northeastern University, from Nov 2016]
Julio Adriazola Soto [Inria, Inria Chile, from Jan 2016 until Mar 2016]
Michele Pittoni [Inria, Univ. Paris VI, M2 Student, from Feb 2016 until Jul 2016]
Adhir Chakravarti [Inria, Univ. Paris VI, M2 Student, from Feb 2016 until Jul 2016]
Zied Ben Houidi [Bell Labs (Alcatel), from Apr 2016]

2. Overall Objectives

2.1. Overall Objectives

Muse's research is broadly in the area of network measurements. We focus on developing new algorithms and systems to improve user experience online. In particular, we are addressing two main problems of today's Internet users:

1. Technology is too complex. Most Internet users are not tech-savvy and hence cannot fix performance problems and anomalous network behavior by themselves. The complexity of most Internet applications makes it hard even for networking experts to fully diagnose and fix problems. Users can't even know whether they are getting the Internet performance that they are paying their providers for.
2. There is too much content. The proliferation of user-generated content (produced anywhere with mobile devices and immediately published in social media) along with the vast amount of information produced by traditional media (e.g., newspapers, television, radio) poses new challenges in achieving an effective, near real-time information awareness and personalization. For instance, users need novel filtering and recommendation tools for helping them to decide which articles to read or which movie to watch.

3. Research Program

3.1. Active probing methods

We are developing methods that actively introduce probes in the network to discover properties of the connected devices and network segments. We are focusing in particular on methods to discover properties of home networks (connected devices and their types) and to distinguish if performance bottlenecks lie within the home network versus outside. Our goal is to develop adaptive methods that can leverage the collaboration of the set of available devices (including end-user devices and the home router, depending on which devices are running the measurement software).

3.2. Passive monitoring methods

This part our research develops methods that simply observe network traffic to infer the performance of networked applications and the location of performance bottlenecks, as well as to extract patterns of web content consumption. We are working on techniques to collect network traffic both at user's end-devices and at home routers. We also have access to network traffic traces collected on a campus network and on a large European broadband access provider.

3.3. Inferring user online experience

We are developing hybrid measurement methods that combine passive network measurement techniques to infer application performance with techniques from HCI to measure user perception. We will later use the resulting datasets to build models of user perception of network performance based only on data that we can obtain automatically from the user device or from user's traffic observed in the network.

3.4. Filtering real-time Web streams

The Web has become a large-scale real-time information system forcing us to revise both how to effectively assess relevance of information for a user and how to efficiently implement information retrieval and dissemination functionality. To increase information relevance, Real-time Web applications such as Twitter and Facebook, extend content and social-graph relevance scores with "real-time" user generated events (e.g. re-tweets, replies, likes). To accommodate high arrival rates of information items and user events we explore a publish/subscribe paradigm in which we index queries and update on the fly their results each time a new item and relevant events arrive. In this setting, we need to process continuous top-k text queries combining both static and dynamic scores. To the best of our knowledge, this is the first work addressing how non-predictable, dynamic scores can be handled in a continuous top-k query setting.

3.5. Flexible online drift detection

Monitoring streaming content is a challenging big data analytics problem, given that very large datasets are rarely (if ever) stationary. In several real world monitoring applications (e.g., newsgroup discussions, network connections, etc.) we need to detect significant change points in the underlying data distribution (e.g., frequency of words, sessions, etc.) and track the evolution of those changes over time. These change points, depending on the research community, are referred to as temporal evolution, non-stationarity, or concept drift and provide valuable insights on real world events (e.g. a discussion topic, an intrusion) to take a timely action. In our work, we adopt a query-based approach to drift detection and address the question of processing drift queries over very large datasets. To the best of our knowledge, our work is the first to formalize flexible drift queries on streaming datasets with varying change rates.

4. Application Domains

4.1. Home Network Diagnosis

With the availability of cheap broadband connectivity, Internet access from the home has become a ubiquity. Modern households host a multitude of networked devices, ranging from personal devices such as laptops and smartphones to printers and media centers. These devices connect among themselves and to the Internet via a local-area network —a *home network*– that has become an important part of the "Interne experience". In fact, ample anecdotal evidence suggests that the home network can cause a wide array of connectivity impediments, but their nature, prevalence, and significance remain largely unstudied.

Our long-term goal is to assist users with concrete indicators of the causes of potential problems and—ideally—ways to fix them. We intend to develop a set of easy-to-use home network diagnosis tools that can reliably identify performance and functionality shortcomings rooted in the home. The development of home network diagnosis tools brings a number of challenges. First, home networks are heterogeneous. The set of devices, configurations, and applications in home networks vary significantly from one home to another. We must develop sophisticated techniques that can learn and adapt to any home network as well as to the level of expertise of the user. Second, there are numerous ways in which applications can fail or experience poor performance in home networks. Often there are a number of explanations for a given symptom. We must devise techniques that can identify the most likely cause(s) for a given problem from a set of possible causes. Third, even if we can identify the cause of the problem, we must then be able to identify a solution. It is important that the output of the diagnosis tools we build is "actionable". Users should understand the output and know what to do.

We are conceiving methods for two application scenarios: (i) when the end user in the home deploys our diagnostic tools either on the home gateway (the gateway often combines a DSL/cable modem and an access point; it connects the home network to the ISP) or on devices connected to the home network and (ii) when ISPs collect measurements from homes of subscribers and then correlate these measurements to help identify problems.

Assisting end users. We are developing algorithms to determine whether network performance problems lie inside or outside the home network. Given that the home gateway connects the home with the rest of the Internet, we are designing an algorithm (called *HoA*) that analyzes traffic that traverses the gateway to distinguish access link and home network bottlenecks. A measurement vantage point on the gateway is key for determining if the performance bottleneck lies within the home network or the access ISP, but we also need to deploy diagnosis tools in end-devices. First, some users may not want (or not know how) to deploy a new home gateway in their homes. Second, some problems will be hard to diagnose with only the vantage point of the gateway (for example, when a device cannot send traffic or when the wireless is poor in certain locations of a home). We can obtain more complete visibility by leveraging *multiple* measurement nodes around the home, potentially including the home gateway, all participating jointly in the measurement task. We have an ongoing project to realize a home network analyzer as a web-based measurement application built on top of

our team's recently developed browser-based measurement platform, *Fathom*. To integrate the home gateway in the analyzer, we plan to engage the BISmark Project. BISmark already provides a web server as well as extensive configurability, allowing us to experiment freely with both passive as well as active measurements. We must develop a home network analyzer that can first discover the set of devices connected to the home network that can collaborate on the diagnosis task. We will then develop tomography algorithms to infer where performance problems lie given measurements taken from the set of available vantage points.

Assisting Internet Service Providers (ISPs). Our discussions with several large access ISPs reveal that service calls are costly, ranging from $9–25 per call, and as many as 75% of service calls from customers are usually caused by problems that have nothing to do with the ISP. Therefore, ISPs are eager to deploy techniques to assist in home network diagnosis. In many countries ISPs control the home gateway and set-top-boxes in the home. We plan to develop more efficient mechanisms for home users to report trouble to their home ISP and consequently reduce the cost of service calls. This project is in collaboration with Technicolor and Portugal Telecom. Technicolor is a large manufacturer of home gateways and set-top-boxes. Portugal Telecom is the largest broadband access provider in Portugal. Technicolor already collects data from 200 homes in Portugal. We are working with the data collected in this deployment together with controlled experiments to develop methods to diagnose problems in the home wireless.

4.2. Quality of Experience

An increasing number of residential users consume online services (e.g., VoD, Web browsing, or Skype) in their everyday activities (e.g., for education or entertainment purposes), using a variety of devices (e.g., tablets, smartphones, laptops). A high Quality of Service (QoS) is essential for sustaining the revenue of service providers, carriers, and device manufactures. Yet, the perceived Quality of Experience (QoE) of users is far from perfect e.g., videos that get stalled or that take a long time to load. Dissatisfied users may change Internet Service Providers (ISPs) or the online services. Hence, the incentives for measuring and improving QoE in home networks are high while mapping network and application QoS to QoE is a challenging problem. In this work we have focused in measuring several network Quality-of-Service (QoS) metrics, such as latency and bandwidth, both in residential Wi-Fi as well as broadband networks, homes are using for connecting to the Internet.

The WiFi Context. Residential Wi-Fi performance, however, is highly variable. Competing Wi-Fi networks can cause contention and interference while poor channel conditions between the station and the access point (AP) can cause frame losses and low bandwidth. In some cases, the home Wi-Fi network can bottleneck Internet access. While problems in the Wi-Fi network may affect several network QoS metrics, users will typically only notice a problem when poor Wi-Fi affects the QoE of Internet applications. For example, a Wi-Fi network with low bandwidth may go unnoticed unless the time to load Web pages increases significantly. A user observing degraded QoE due to Wi-Fi problems may mistakenly assume there is a problem with the Internet Service Provider (ISP) network. Our discussions with residential ISPs confirm that often customers call to complain about problems in the home Wi-Fi and not the ISP network.

Prior work has focused on QoS metrics for some applications (e.g., on-line video, Web browsing, or Skype) with no attempt to identify when Wi-Fi quality affects QoE. We are particularly interested in assisting ISPs to predict when home Wi-Fi quality degrades QoE. ISPs can use this system to detect customers experiencing poor QoE to proactively trigger Wi-Fi troubleshooting. ISPs often control the home AP, so we leverage Wi-Fi metrics that are available on commercial APs. Detecting when Wi-Fi quality degrades QoE using these metrics is challenging. First, we have no information about the applications customers are running at any given time. ISPs avoid capturing per-packet traffic traces from customers, because of privacy considerations and the overload of per-packet capture. Thus, we must estimate the effect of Wi-Fi quality on QoE of popular applications, which most customers are likely to run. In this context, we study Web as a proof of concept, as a large fraction of home traffic corresponds to Web. Second, application QoE may be degraded by factors other than the Wi-Fi quality (e.g., poor Internet performance or an overloaded server). Although a general system to explain any QoE degradation would be extremely helpful, our monitoring at the AP prevents us from having the end-to-end view necessary for such general task. Instead, we focus on identifying when Wi-Fi quality

degrades QoE. Finally, Wi-Fi metrics available in APs are coarse aggregates such as the average PHY rate or the fraction of busy times. It is open how to effectively map these coarse metrics into QoE.

Predicting QoE. Clearly, different actors in the online service chain (e.g., video streaming services, ISPs) have different incentives and means to measure and affect the user QoE. Uncovering statistically equivalent subsets of QoS metrics across and within levels provides actionable knowledge for building QoE predictors. To achieve this goal, we leverage recent advances on feature selection algorithms to exploit available experimental evidence of the joint probability distributions of QoE/QoS metrics. This type of statistical reasoning will enable us to determine local causal relationships between a target QoE variable, seen as effect, and multiple QoS metrics across or within levels, seen as causes. Such data-driven analysis is justified by the multiplicity of dependencies that exist between network or application QoS metrics as different adaptation mechanisms (e.g., TCP congestion avoidance, HTTP bitrate adaptation) are activated at each level in real life. Building optimal predictors based on (eventually several) probabilistically minimal subsets of features opens the way for a principled comparison of the predictors.

4.3. Data Analytics for the Internet of Things

The Internet of Things (IoT) is rapidly transforming the physical world into a large scale information system. A wave of smart "things" smoothly disappear in our environment (aka *Pervasive Computing*), or be embodied in humans (aka *Wearable Computing*, and continuously produce valuable information regarding almost every living context and process. *Making sense of the data streams "things" produce and share* is crucial for disruptive IoT applications. From smart devices and homes, to smart roads and cities, IoT data analytics is expected to enable a resource-conscious automation of our everyday life in terms of operational efficiency, security, safety as well as of a lower energy footprint.

Multi-dimensional Usage Patterns. We have initially investigated how data analytics for Machine-to-Machine (M2M) data (connectivity, performance, usage) produced by connected devices in residential Intranet of Things, could support novel *home automation services* that enrich the living experience in smart homes. We have investigated new data mining techniques that go beyond binary association rule mining for traditional market basket analysis, considered by previous works. We design a multidimensional pattern mining framework, which collects raw data from operational home gateways, it discretizes and annotates the raw data, it produces traffic usage logs which are fed in a multidimensional association rule miner, and finally it extracts home residents habits. Using our analysis engine, we extract complex device co-usage patterns of 201 residential broadband users of an ISP, subscribed to a n-play service. Such fine-grained device usage patterns provide valuable insights for emerging use cases, such as adaptive usage of home devices (aka horizontal integration of things). Such use cases fall within the wider area of human-cognizant Machine-to-Machine communication aiming to predict user needs and complete tasks without users initiating the action or interfering with the service. While this is not a new concept, according to Gartner cognizant computing is a natural evolution of a world driven not by devices but collections of applications and services that span across multiple devices, in which human intervention becomes as little as possible, by analyzing past human habits. To realize this vision, we are interested in co-usage patterns featuring spatio-temporal information regarding the context under which devices have been actually used in homes. For example, a network extender which is currently turned off, could be turned on at a certain day period (e.g., evening) when it has been observed to be highly used along with other devices (e.g., a laptop or a tablet). Alternatively, the identification of frequent co-usage of particular devices at a home (say iPhone with media player), could be used by a things recommender to advertise the same set of devices at another home (say another iPhone user could be interested in a media player).

Time Series Motif. Furthermore, we are interested in extracting previously unknown recurring patterns (aka motifs) directly from traffic time series reported by residential gateways. Such motifs could help ISPs to reduce the cost for *serving and diagnosing remotely home networks*, or even help assist in *defining home-specific bandwidth sharing and prioritization policies*. More precisely, traffic motifs enriched with detailed home device information is a valuable input for root cause diagnosis and can be contrasted to the trouble description reported by users to the ISP. Moreover, in their majority, ISPs typically broadcast firmware and software updates to all gateways at nights (some operators even on a daily basis). This may cause service outages,

given that some gateways may exhibit an active network usage during night time. A fine-grained temporal characterization of residential bandwidth consumption will enable ISPs to differentiate RGWs firmware update policies according to the least cumbersome time window per home, thus, improving the overall QoE of residential users. Finally, home network resources (bandwidth) are shared not only among residents using an increasing number of on-line applications (e.g., social networking, gaming, uploading/downloading, etc.) and real time services (TV on-demand, teleconferencing), but also with guests, neighbors, or even the occasional passes by. Existing methods for bandwidth sharing and traffic prioritization are static and coarse. ISPs usually allocate a fixed percentage of home bandwidth to non-residential users, while traffic prioritization in commodity gateways is at best based on the network port on which traffic is sent or received. We believe that behavioural patterns extracted by gateway traffic time series can be used to support dynamic policies for sharing home bandwidth that consider the online habits of residential users. For example, in-home traffic congestion can be avoided by ordering the traffic patterns of different devices observed especially during afternoon and weekends. These patterns reveal the bandwidth consumption behavior of different groups of residential users (adults and children employ different devices during the same time-slots) while the comparison of traffic domination help us to distinguish between residents and guests (pattern-specific vs global traffic dominant devices).

4.4. Crowd-sourced Information Filtering and Summarization

With the explosion of the People-centric Web, there is a proliferation of crowd-sourced content either under the form of qualitative reviews (mainly textual) and quantitative ratings (as 5 star ratings) regarding diverse products or services or under the form of various "real-time" feedback events (e.g., re-tweets, replies, likes, clicks, etc.) on published web content (ranging from traditional news, TV series, and movies to specialized blogs and posts shared over social networks). Such content captures the wisdom of the crowd and is valuable information source for building collaborative filtering systems and text summarization tools coping with information overload. For example, they can assist users to pick the most interesting web pages (e.g. Delicious) or to choose which movie to watch next (e.g. Netflix).

Implicit Feedback in Communities of a Place. We are initially interested in addressing one of the main limitation of collaborative filtering systems namely, the strong user engagement required to provide the necessary input (e.g., regarding their friends, tags or sites of preference) which is usual platform specific (i.e., for a particular social network, tagging, or bookmark system). The lack of user engagement translates into cold start and data sparsity. To cope with this limitation, we are developing a system called WeBrowse that passively observes network traffic to extract user clicks (i.e., the URLs users visit) for group of people who live, study, or work in the same place. Examples of such communities of a place are: (i) the students of a campus, (ii) the people living in a neighbourhood or (iii) researchers working in the same site. WeBrowse then promotes the hottest and most popular content to the community members sharing common interests.

Personalized Review Summarization. Finally, we are interested in helping people to take informed decisions regarding their shopping or entertainment activities. The automated summarization of a review corpus (for example, movie reviews from Rotten Tomatoes or IMDB; or restaurant reviews from Yelp) aims to assist people to form an opinion regarding a product/service of interest, by producing a coherent summary that is helpful and can be easily assimilated by humans. We are working on review summarisation methods that combine both objective (i.e., related to the review corpus) and subjective (i.e., related to the end-user interests) interestingness criteria of the produced reviews. In this respect we are exploiting domain models (e.g., Oscar's merit categories for movies) to elicit user preferences and mine the aspects of products/services actually commented in the textual sentences of reviews. For example, different summaries should be produced when a user is more interested in the actors performance rather than the movie story. We are particularly interested in extracting automatically the signatures of aspects (based on a set of seed terms) and rank review sentences on their importance and relevance w.r.t. the aspects they comment. Last but not least we are optimizing the automatically constructed summary w.r.t. to a number of criteria such as the number of the length of included sentences from the original reviews, the polarity of sentiments in the described aspects, etc.

5. New Software and Platforms

5.1. Fathom

Fathom - browser-based network measurement platform
KEYWORDS: Internet access - Performance measure - Network monitoring
FUNCTIONAL DESCRIPTION

Fathom is a Firefox browser extension that explores the browser as a platform for network measurement and troubleshooting. It provides a wide range of networking primitives directly to in-page JavaScript including raw TCP/UDP sockets, higher-level protocol APIs such as DNS, HTTP, and UPnP, and ready-made functionality such as pings and traceroutes.

- Participants: Anna-Kaisa Pietilainen and Stephane Archer
- Contact: Anna-Kaisa Pietilainen
- URL: https://muse.inria.fr/fathom/

5.2. HostView

FUNCTIONAL DESCRIPTION

End-host performance monitoring and user feedback reporting

- Participants: George Rosca, Anna-Kaisa Pietilainen and Renata Cruz Teixeira
- Contact: Renata Cruz Teixeira
- URL: https://team.inria.fr/muse/

5.3. Online HoA

Online implementation of home and access throughput bottleneck detection algorithm 'HoA'
FUNCTIONAL DESCRIPTION

"Home or Access" (HoA) is a system that localizes performance problems in home and access networks. Originally, we implement HoA as custom firmware that collect traces from off-the-shelf home routers. HoA uses timing and buffering information from passively monitored traffic at home routers to detect both access link and wireless network bottlenecks. HoA runs offline on a server to locate last-mile downstream throughput bottlenecks based on the analysis of packet traces collected from home routers. Our attempts to run HoA online on commodity home routers, however, revealed the challenges with performing per-packet analysis on such resource-constrained devices. The online HoA resolves this issue. We design an access bottleneck detector based on lightweight pings of the access link, and a wireless bottleneck detector based on a model of wireless capacity using metrics that are easily available in commodity home routers such as the wireless physical rate and the count of packets/bytes transmitted.

- Contact: Renata Cruz Teixeira
- URL: https://github.com/inria-muse/browserlab

5.4. SimilarityExplanation

Prototype implementation for explaining a set of similar and recommended movies.
FUNCTIONAL DESCRIPTION

In this web-based prototype for similar movies explanation, we propose two types of browsing for: personalized browsing and non personalized browsing. In the non personalized browsing we suppose that we don't have the user profile. Similar movie sublists are ordered only according to their similarity to the selected movie. For the personalized browsing , we select users that have different profiles from our dataset. We give these users names of actors, according to the types of movies they watch. For each user, we compute the predicted ratings using the matrix factorization model. We select pairs of genres to display to each user based on the preferred genres for the user. In our prototype we identify the preferred genres per user based on the most frequent movie genre pairs that the user has already seen. We then organize the recommended movies with a high rating prediction in sublists, according to the user most preferred genre pairs. When a user selects a movie from the sublists of recommended movies, our application suggests the similar movies presented under four sublists with the added list of words. The sublists are personalized for each user by reordering the movies according to the users predicted ratings.

- Contact: Sara El Aouad
- URL: https://team.inria.fr/muse/

5.5. UCN

User-Centric Networking
FUNCTIONAL DESCRIPTION

The User-Centric Networking (UCN) project is seeking to understand how people consume various kinds of content when using computer networks. Within this project we are undertaking a detailed user study across a range of environments in order to understand the practices involved in consuming media and other content according to context.

- Participants: Renata Cruz Teixeira and Anna-Kaisa Pietilainen
- Contact: Anna-Kaisa Pietilainen
- URL: https://team.inria.fr/muse/

5.6. WeBrowse

FUNCTIONAL DESCRIPTION

WeBrowse is the first passive crowdsource-based content curation system. Content curation is the act of assisting users to identify relevant and interesting content in the Internet. WeBrowse requires no active user engagement to promote content. Instead, it extracts the URLs users visit from traffic traversing an ISP network to identify popular content. WeBrowse contains a set of heuristics to identify the set of URLs users visit and to select the subset that are interesting to users.

- Contact: Giuseppe Scavo
- URL: http://webrowse.polito.it/

6. New Results

6.1. Home Network or Access Link? Locating Last-mile Downstream Throughput Bottlenecks

Participants: Srikanth Sundaresan (ICSI), Nick Feamster (Princeton), Renata Teixeira

As home networks see increasingly faster downstream throughput speeds, a natural question is whether users are benefiting from these faster speeds or simply facing performance bottlenecks in their own home networks. We studied the problem whether downstream throughput bottlenecks occur more frequently in their home networks or in their access ISPs. We identified lightweight metrics that can accurately identify whether a throughput bottleneck lies inside or outside a user's home network and developed a detection algorithm that locates these bottlenecks. We validated this algorithm in controlled settings and characterized bottlenecks on two deployments, one of which included 2,652 homes across the United States. We found that wireless bottlenecks are more common than access- link bottlenecks—particularly for home networks with downstream throughput greater than 20 Mbps, where access-link bottlenecks are relatively rare.

6.2. Characterizing Home Device Usage From Wireless Traffic Time Series

Participants: Katsiaryna Mirylenka (IBM), Vassilis Christophides, Themis Palpanas (University Rene Descartes), Ioannis Pefkianakis (HP Labs), Martin May (Technicolor)

We conducted a thorough analysis of traffic dynamics of heterogeneous wireless (WiFi) devices connected to 196 real RGWs, which are subscribers of a major European ISP. We focus on a time-oriented analysis of continuous traffic data to extract previously unknown patterns recurring of internet consumption that happen within, or across homes. We also assess the impact of different types of devices, such as laptops, desktops (classiffied as fixed devices), and tablets, smartphones (classiffied as portables), on these patterns. Unsupervised learning techniques are used for patterns discovery as the ground truth data regarding home activities are not available. Rather than partitioning homes or devices into distinct behavioral clusters, we are looking to extract informative motifs of bandwidth consumption within or across homes. The main contributions of this work are:

- We propose a novel analysis framework for wireless home traffic data, namely: (a) a correlation-based similarity measure, which exploits the evolution characteristics, rather than the absolute traffic values, and is invariant to scaling;(b) a notion of strong stationarity that in addition to the similarity of data distributions imposes a correlation similarity across non-overlapping time windows; and (d) a definition of dominant devices based on the correlation similarity, that enables an intuitive and statistically grounded interpretation of the results.

- We evaluate the effectiveness of the proposed framework using real data of wireless traffic observations and report the main findings: (a) there are many repetitive patterns within and across RGWs which describe the intrinsic user behavior of users and valuable to ISPs; (b) as networking time series are not stationary certain aggregation should be performed in order to find statistically significant patterns. The best time windows to aggregate home traffic data is found to be 8 hours for weekly patterns and 3 hours for daily patterns; (c) frequent weekly patterns correspond to heavy bandwidth usage both during weekdays and weekends, and frequent daily patterns correspond to (mostly) evening usage, (d) weekend usage tends to rely on portable devices, weekday usage relies more on fixed devices, while discontinuous usage within a day (mostly active in the evening or the morning) is still due to portable devices; and (e) almost every RGW involves a device that dominates its overall traffic, thus the behavior of this device should be mainly considered by ISPs while planning the updates.

6.3. Towards a Causal Analysis of Video QoE from Network and Application QoS

Participants: Michalis Katsarakis, Renata Teixeira, Maria Papadopouli (Univeristy of Crete), Vassilis Christophides

We have exploited an original framework for mining causal relationships among a 5-start rating of user QoE and various QoS metrics at network and application level. In particular, we have analysed QoE scores provided by a set of users for YouTube video streaming applications under different network conditions. We found that optimal QoE predictors we can be build using a minimal signature of only three features from application or network QoS metrics compared to four when features from both layers are considered. A thorough comparative analysis of the prediction accuracy of three models build using minimal signatures composed of (i) only network QoS, (ii) only application QoS, and (iii) both QoS features demonstrated that we can predict the QoE using only network QoS metrics and more surprisingly, predicting the QoE from network QoS metrics is as accurate as when using application QoS metrics. This work is the first step towards our ambition to assess QoE directly from network QoS metrics obtained via passive measurements of real traffic generated by online users. We will rely on the extracted minimal QoE/QoS signatures to build real-time predictors and compare their accuracy when using only network, only application or both QoS metrics. Last but not least, we plan to extend our experimental setting for other online applications such as teleconferencing services.

6.4. Predicting the effect of home Wi-Fi quality on Web QoE

Participants: Diego Neves da Hora, Renata Teixeira, Karel Van Doorselaer (Technicolor), Koen Van Oost (Technicolor)

We developed a model that predicts the effect of Wi-Fi quality on Web QoE, using solely Wi-Fi metrics commonly available in commercial APs. We trained our predictor during controlled experiments on a Wi-Fi testbed and assess its accuracy through cross-validation, obtaining an RMSE of 0.6432 MO, and by applying it on a separate validation dataset, obtained on an uncontrolled environment, finding an RMSE of 0.9283. Finally, we apply our predictor on Wi-Fi metrics collected in the wild from 4,880 APs over a period of 40 days. We find that Wi-Fi quality is mostly good for Web—in more than 60% of samples Wi-Fi quality does not degrade Web QoE. When we consider average complexity Web pages, however, Wi-Fi quality degrades Web QoE in 11% of samples. Moreover, we saw that 21% of devices present more than 20% of poor Web QoE samples, with 5% of these showing highly intermittent QoE degradations, which are particularly hard to diagnose, indicating the need for a long-term monitoring approach to detect and fix problems.

6.5. Passive Wi-Fi Link Capacity Estimation on Commodity Access Points

Participants: Diego Neves da Hora, Karel Van Doorselaer (Technicolor), Koen Van Oost (Technicolor), Renata Teixeira, Christophe Diot (Safran)

We propose an algorithm to estimate the link capacity based on passive metrics from APs, which is ready to be deployed at scale. We show that it is possible to estimate the link capacity per PHY rate based on a limited set of parameters related to the particular AP instance. Then, we extend the initial model to estimate the link capacity when the PHY rate varies. We measured the link capacity in different link quality conditions and found that more than 90% of the estimations present error below 15% without prior parameter tuning, and more than 95% present estimation error below 5% with appropriate parameter tuning using fixed PHY rate tests.

6.6. Content-Based Publish/Subscribe System for Web Syndication

Participants: Zeinab Hmedeh CNAM, Harry Kourdounakis (FORTH-ICS, Vassilis Christophides, Cedric du Mouza (CNAM), Michel Scholl (CNAM), and Nicolas Travers (CNAM)

Content syndication has become a popular way for timely delivery of frequently updated information on the Web. Today, web syndication technologies such as RSS or Atom are used in a wide variety of applications spreading from large-scale news broadcasting to medium-scale information sharing in scientific and professional communities. However, they exhibit serious limitations for dealing with information overload in Web 2.0. There is a vital need for efficient real-time filtering methods across feeds, to allow users to effectively follow personally interesting information.

To efficiently check whether all keywords of a subscription also appear in an incoming item (i.e., broad match semantics), we need to index the subscriptions. Count-based (CI) and tree-based (TI) are two main indexing schemes proposed in the literature for counting explicitly and implicitly the number of contained key-words. The majority of related data structures cannot be employed for conjunctions of keywords (rather than attribute-value pairs) due to the space high-dimensionality. In this paper, we are interested in efficient implementations of both indexing schemes using inverted lists (IL) for CI and a variant for distinct terms of ordered tries (OT) for TI and study their behavior for critical parameters of realistic web syndication workloads. Although these data structures have been employed to evaluate broad match queries in the context of selective information dissemination and sponsored search or for mining frequent item sets, their memory and matching time requirements appear to be quite different in our setting. This is due to the peculiarities of web syndication systems which are characterized 1) by information items of average length (25?36 distinct terms) which are greater than advertisement bids (4?5 terms) and smaller than documents of Web collections (12K terms) and 2) by very large vocabularies of terms (up to 1.5M terms). Note also that due to broad match semantics, information retrieval techniques for optimizing ILs (e.g., early pruning) are not suited in our setting.

We present analytical models for memory requirements and matching time and we conduct a thorough experimental evaluation to exhibit the impact of critical parameters of realistic web syndication workloads. We found that for small vocabularies, POT matching time is one order of magnitude faster than the best IL (RIL), while for large vocabularies (like the one used on the Web), RIL outperforms the matching POT, which uses almost four times more memory space. The actual distribution of term occurrences has almost no impact on the size of the three indexing structures while it significantly affects the number of nodes that need to be visited upon matching something that justifies OT performance gains. The smaller the subscription length, the larger the OT factorization gain w.r.t. IL and the larger the rank of the term from which the OT substructure degenerates to an IL.

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. ANR

7.1.1.1. BottleNet: Understanding and Diagosing end-to-end communication problems

Type: ANR Project (N ANR-15-CE25-0013-01)

Instrument: PRCE

Duration: February 2016 - 36 mois

Coordinator: Renata Teixeira

Other partners: Inria Paris, LORIA, Ip-label SaS, ORANGE SA , TSP (SAMOVAR) Telecom SudParis, University of Lille 1 (CRIStAL)

Inria contact: Renata Teixeira

Abstract: As our lives become more dependent on the Internet, it is easy to understand peoples frustration when poor Internet performance prevents them from accomplishing ever-more important online activities. The Quality of Experience (QoE) when accessing the Internet is thus a key factor for todays society. When users experience poor Internet QoE, they are often helpless. The complexity of Internet services and of users local connectivity has grown dramatically in the last years with the proliferation of proxies and caches at the core and of home wireless and 3G/4G access. However, it is hard even for experts to diagnose the sources of performance bottlenecks. We argue that the new rules for regulating Internet access and plans to improve it should focus on the end-to-end Internet QoE - i.e., the user experience when performing her usual online activities on the Internet - and should provide means for all different players to accurately diagnose the sources of poor Internet QoE. The objective of BottleNet is to deliver methods, algorithms, and software systems

to measure Internet QoE and diagnose the root cause of poor Internet QoE. Our goal calls for tools that run directly at users devices. We plan to collect network and application performance metrics directly at users devices and correlate it with user perception to model Internet QoE, and to correlate measurements across users and devices to diagnose poor Internet QoE. This data-driven approach is essential to address the challenging problem of modeling user perception and of diagnosing sources of bottlenecks in complex Internet services. BottleNet will lead to new solutions to assist users, network and service operators as well as regulators in understanding Internet QoE and the sources of performance bottleneck.

7.1.2. CNRS Big Data Projects

7.1.2.1. BigGeoQUP: Big Geospatial Data Quality and User Privacy

Type: CNRS Mission Interdisciplinarité (MI) DEFI MASTODONS "La qualité des données dans les Big Data"

Instrument: AAP 2016

Duration: Mars 2016 - 12 mois

Coordinator: Dimitris Kotzinos (University of Cergy Pointoise)

Other partners: Inria Paris, IGN-COGIT

Inria contact: Vassilis Christophides

Abstract: Big Geo Data represents an important type of the crowd sourced data that are available today at a global scale. This kind of data refers to locations, i.e., Points of Interest (POIs), and is usually published in social media (e.g., Facebook, Google+) or in specialized platforms (e.g., Open Street Maps, Yelp). The quality (e.g., precision, accuracy, consistency) of geo-referenced crowd sourced content depends on the origin (machine vs. human generated), the level of detail of the extraction methods, as well as the obfuscation techniques used to protect users privacy. There is clearly a tradeoff between enhancing the quality of published geo data and the privacy risks entailed for the individuals, also known as geoprivacy, to uncover places visited, trajectories pursuit etc. Understanding the different aspects of geographic/geometric/geospatial quality involved in crowd-sourced geo data and assessing the privacy risks introduced by enhancing its quality in personal, social, and urban applications is a challenging topic. In this project we are interested in studying the following questions in concrete use-cases:

- How can we measure the quality of geographic/geometric/geospatial data involved in crowd-sourced content?
- How can we assess the privacy risks introduced by enhancing geospatial quality in personal, social and urban applications especially in the context of social media platforms?

In this project Inria (MUSE) is interested in the opportunities and privacy concerns of mobile location analytics supporting customers or travellers experience in venues of various types and sizes (e.g., retail stores, shopping malls, airports, theme parks, etc.). Such (indoor) location-based services in smart spaces presents new privacy risks as data continuously flows between visitors mobile devices, networked sensors embedded into the environment as well as the backend analytics platform in order to track users and anticipate their needs.

7.2. European Initiatives

7.2.1. FP7 & H2020 Projects

7.2.1.1. User-Centric Networking (UCN)

Type: FP7

Instrument: Specific Targeted Research Project

Duration: October 2013 - September 2016

Coordinator: Technicolor

Other partners: Eurecom, Fraunhofer FOKUS, Intamac, University of Cambridge, University of Nottingham, Martel, NICTA, Portugal Telecom

Inria contact: Renata Teixeira

Abstract: This project introduces the concept of User Centric Networking (UCN), which is a new paradigm leveraging user information at large to deliver novel content recommendation systems and content delivery frameworks. UCN recommendation and content delivery systems will leverage in-depth knowledge about users to help them find relevant content, identify nearby network resources and plan how to deliver the actual content to the appropriate device at the desired time. These systems will additionally account for influences from users' social networks on their content consumption. The goal of this project is to design a UCN system architecture for user-centric connected media services. We will build UCN upon three complementary research pillars:

1. understanding user context: This data can be broadly categorized into three groups. First, the physical and environmental context A second category of data is that which can be extracted from social network interactions. The third category of data is behavioural

2. profiling and predicting user interests: By gaining a deep understanding of the user, we may be able to cast a much wider net in the content ocean and locate a richer catalogue of interesting content for the user

3. personalizing content delivery: Rather than the user (or the service provider) having to worry about the mode of connectivity, device, service, location, etc., the network intelligently directs and adapts the transport stream, or perhaps pre-fetches and replicates content chunks, to the particular and immediate needs of the user.

See also: http://usercentricnetworking.eu/

7.3. International Initiatives

7.3.1. Inria Associate Teams Not Involved in an Inria International Labs

- BetterNet (Inria Project Lab): An observatory to measure and Improve Internet Service Access from User Experience. Project Coordinator: Isabelle Chrisment, Partners: Inria teams Spirals, Diana, Muse, Dionysos and Madynes, Inria Chile

BetterNet intends to build and deliver a scientific and technical collaborative observatory to measure and improve the Internet service access as perceived by users. We will propose new original user-centered measurement methods, which will associate social sciences to better understand Internet usage and the quality of services and networks. Our observatory can be defined as a vantage point, where:

- tools, models and algorithms/heuristics will be provided to collect data,
- acquired data will be analyzed, and shared appropriately with scientists, stakeholders and civil society, and
- new value-added services will be proposed to end-users.

This project will also allow Inria to become a key reference in the digital field, not only for scientist researchers but also for policy makers, rulers, and, for citizens in general, by giving them a more accurate and reliable basis for decisions making at an individual scale (privacy strategies) or at a collective scale (legal norms). An originality of our proposal is to ensure that researchers from different disciplines (exact sciences, applied sciences and social sciences) will collaborate in the design of this observatory, in the dissemination of results for the research community, executives and public at large.

7.3.2. Inria International Partners

7.3.2.1. Informal International Partners

- Princeton (Prof. Nick Feamster): We have a long-term collaboration on measuring the performance of residential broadband Internet access networks and more recently on home network diagnosis.
- ICSI (Dr. Srikanth Sundaresan, Dr. Christian Kreibich, Dr. Robin Sommer): With C. Kreibich, we have been developing Fathom, a browser-based network measurement platform. We are now adding home network diagnosis capabilities to Fathom. We are collaborating with S. Sundaresan on detecting last-mile bottlenecks. In addition, with Robin Sommer we are working on the potential of matching the profiles of a user across multiple online social networks.
- Northwestern University (Prof. Fabian Bustamante and his doctoral student Zachary Bischof): we are working on identifying user activity from network traffic.

7.4. International Research Visitors

7.4.1. Internships

- Arash Molavi Kakhki, Ph.D Student, Northeastern University, from Nov 2016
- Julio Adriazola Soto, M2 Student, Inria Chile, from Jan 2016 until Mar 2016
- Michele Pittoni, M2 Student, Univ. Paris VI, from Feb 2016 until Jul 2016
- Adhir Chakravarti, M2 Student, Univ. Paris VI, from Feb 2016 until Jul 2016
- Zied Ben Houidi Bell Labs (Alcatel), from Apr 2016

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events Organisation

8.1.1.1. Chair of Conference Program Committees

- R. Teixeira, Co-chair of the program committee of ACM IMC 2017,
- R. Teixeira, Co-chair of the program committee of ACM/ISOC ANRW 2017,
- V. Christophides, Tutorial co-chair of the 22nd IEEE Symposium on Computers and Communications (ISCC17) 03 - 06 July 2017, Heraklion, Crete, Greece (http://www.ics.forth.gr/iscc2017/)
- V. Christophides, Area Chair for "Semi-structured, Web, and Linked Data Management" in the 18th International Conference on Data Engineering (ICDE'16), May 16-20, 2016, Helsinki, Finland (http://icde2016.fi)
- V. Christophides, Workshop co-chair of the 1st International Workshop on Big Geo Data Quality and Privacy (BIGQP) with D. Kotzinos, C. Nikolaou, and Y. Theodoridis, March 21, 2017, Venice, Italy (http://www-etis.ensea.fr/BigGeoQ-UP/BIGQP2017/).
- V. Christophides, Workshop co-chair of the 2nd International Workshop on Preservation of Evolving Big Data, co-located with EDBT 2016 with G. Papastefanatos and S. Viglas March 15, 2016 Bordeaux, France (http://www.diachron-fp7.eu/2nd-diachron-workshop.html)

8.1.1.2. Member of the Conference Program Committees

- V. Christophides, PC member of the 2nd IFIP/IEEE International Workshop on Analytics for Network and Service Management (AnNet17) May 8, 2017 in Lisbon, Portugal.
- R. Teixeira, PC Member of 2016 Internet Measurement Conference (IMC) November 14-16, 2016 Santa Monica, California, USA.

- V. Christophides, PC member of the 1st International Workshop on Keyword-based Access and Ranking at Scale (KARS), collocated with EDBT2017, Venice, Italy.
- V. Christophides, PC member of the 25th ACM International Conference on Information and Knowledge Management (CIKM16), 24-28 October, Indianapolis, USA.

8.1.2. Invited Talks

- V. Christophides, "Web-scale Blocking, Iterative and Progressive Entity Resolution", ICDE 2017 Tutorial, with K. Stefanidis, and V. Efthymiou, April 19-22, 2017 in San Diego, California, USA.
- R. Teixeira, "HostView: Measuring Internet quality of experience on end-hosts", Dagsthul Seminar 16012, January 4 – 7, 2016.
- R. Teixeira, "The Challenges of Measuring Internet Quality of Experience", Invited Keynote at the Internet-QoE workshop co-located with SIGCOMM'16.
- R. Teixeira, "Residential Internet Performance: A view from the Gateway", Invited talk at the 1st Stakeholder Consultation Workshop on Mapping of Broadband Services in Europe (SMART 2014/0016).
- T. Friedman, A.-K. Pietilainen, R. Teixeira, "Internet Measurements: A Hands-on Introduction", Invited Tutorial at the ACM SIGCOMM 2016, Florianapolis, Brazil, August 22-26 2016.
- V. Christophides, "Resolving Entities in the Web of Data: Concepts, Algorithms and Systems", Invited Tutorial at the International Summer School on Web Science and Technology (WebST 2016) Bilbao, Spain, 18-22 July 2016, co-organized by Deusto and Rovira i Virgili University.

8.1.3. Leadership within the Scientific Community

- R. Teixeira, Vice-chair of ACM SIGCOMM (Since July 2013).
- R. Teixeira, Chair of the ACM SIGCOMM Industrial Liaison Board.
- R. Teixeira, Member of the steering committee of the ACM Internet Measurement Conference (Nov. 2009–Jun. 2016).
- V. Christophides, Member of the EDBT Association (since 2014).

8.1.4. Scientific Expertise

- Reviewer of H2020 and FP7 projects: CogNet, SUPERFLUIDITY, MONROE, MAMI.
- Technical advisory board of the project "Mapping of Broadband Services in Europe".

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Master: Renata Teixeira, "Methodology for research in networking", 10h eqTD, M2, UPMC, France.

Master: Timur Friedman, Renata Teixeira, "Network Metrology", 24h CM, M2, UPMC, France. Created this new master's level class.

Master: Vassilis Christophides, "Big Data Processing and Analytics", 25h M2, Ecole CentraleSupelec, France. Created this new master's level class in Spring 2016.

E-learning

Mooc: Timur Friedman, Renata Teixeira, "Network Metrology", preparation of a five-week Mooc (May 2016) in the platform FUN, supported by Inria.

8.2.2. Supervision

PhD: Giuseppe Scavo (December 2016), "Content Curation and Characterization in Communities of a Place", advisors: Z. Ben-Houidi and R. Teixeira

PhD: Sofia Kleisarchaki (November 2016): "Difference Analysis in Big Data: Exploration, Explanation, Evolution" advisors: S. Amer-Yahia and V. Christophides

PhD in progress:

- Maximiliam Bachl (Oct 2016-now), "Home network troubleshooting made easy feedback", advisor: R. Teixeira
- Sara el Aouad (May 2014-now): "Improving the quality of recommendation using semi-structured user feedback", advisors: C. Diot and R. Teixeira
- Diego da Hora (Feb 2014-now): "Crowdsourced Home Network Diagnosis", advisors: C. Diot and R. Teixeira
- Vassilis Efthimiou (Feb 2013-now): "Entity resolution in the Web of Data", advisor: V. Christophides

8.2.3. Juries

R. Teixeira: Member of Doctoral examining committee of Sheharbano Khattak (University of Cambridge, 2016),

R. Teixeira: Member of Doctoral examining committee of Zachary Scott Bischoff (Northwestern University, 2016).

V. Christophides: External evaluator of the Ph.D. Thesis of Maria Koutraki, University PARIS-SACLAY, September 2016

V. Christophides: Member of the appointment committee for the MCF 4234 position on Big Data at the University of Cergy-Pontoise, Mai 2016.

9. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] F. GEERTS, T. UNGER, G. KARVOUNARAKIS, I. FUNDULAKI, V. CHRISTOPHIDES. *Algebraic Structures for Capturing the Provenance of SPARQL Queries*, in "Journal of the ACM", March 2016, vol. 63, n⁰ 1, 7 p. [*DOI :* 10.1145/2810037], https://hal.inria.fr/hal-01411827

[2] Z. HMEDEH, H. KOURDOUNAKIS, V. CHRISTOPHIDES, C. DU MOUZA, M. SCHOLL, N. TRAVERS. *Content-Based Publish/Subscribe System for Web Syndication*, in "Journal of Computer Science and Technology", March 2016, vol. 31, pp. 359 - 380 [*DOI :* 10.1007/S11390-016-1632-8], https://hal.inria.fr/hal-01411848

International Conferences with Proceedings

[3] M. KATSARAKIS, R. TEIXEIRA, M. PAPADOPOULI, V. CHRISTOPHIDES. *Towards a Causal Analysis of Video QoE from Network and Application QoS*, in "ACM SIGCOMM Workshop on QoE-based Analysis and Management of Data Communication Networks (Internet-QoE 2016)", Florianopolis, Brazil, August 2016 [*DOI :* 10.1145/2940136.2940142], https://hal.inria.fr/hal-01338726

[4] K. MIRYLENKA, V. CHRISTOPHIDES, T. PALPANAS, I. PEFKIANAKIS, M. MAY. *Characterizing Home Device Usage From Wireless Traffic Time Series*, in "19th International Conference on Extending Database Technology (EDBT)", Bordeaux, France, March 2016, https://hal.inria.fr/hal-01249778

[5] G. POGHOSYAN, I. PEFKIANAKIS, P. LE GUYADEC, V. CHRISTOPHIDES. *Mining usage patterns in residential intranet of things*, in "The 7th International Conference on Ambient Systems, Networks and Technologies", Madrid , Spain, International Workshop on Big Data and Data Mining Challenges on IoT and Pervasive Systems (BigD2M 2016), May 2016, vol. 00, 6 p. , https://hal.inria.fr/hal-01411676

[6] S. SUNDARESAN, N. FEAMSTER, R. TEIXEIRA. *Home Network or Access Link? Locating Last-Mile Downstream Throughput Bottlenecks*, in "PAM 2016 - Passive and Active Measurement Conference", Heraklion, Greece, March 2016, pp. 111-123 [*DOI :* 10.1007/978-3-319-30505-9_9], https://hal.inria.fr/hal-01294924

Conferences without Proceedings

[7] D. NEVES DA HORA, R. TEIXEIRA, K. VAN DOORSELAER, K. VAN OOST. *Predicting the effect of home Wi-Fi quality on Web QoE*, in "Proceedings SIGCOMM Workshop on QoE-based Analysis and Management of Data Communication Networks (Internet-QoE)", Florianópolis, Brazil, ACM, August 2016 [*DOI :* 10.1145/2940136.2940146], https://hal.inria.fr/hal-01339522

[8] D. NEVES DA HORA, K. VAN DOORSELAER, K. VAN OOST, R. TEIXEIRA, C. DIOT. *Passive Wi-Fi Link Capacity Estimation on Commodity Access Points*, in "Traffic Monitoring and Analysis Workshop (TMA) 2016", Louvain-la-Neuve, Belgium, April 2016, https://hal.inria.fr/hal-01292633

Other Publications

[9] M. BACHL. *Collaborative Home Network Troubleshooting*, Université Pierre & Marie Curie - Paris 6 ; Inria, September 2016, https://hal.inria.fr/hal-01415767

[10] V. EFTHYMIOU, K. STEFANIDIS, V. CHRISTOPHIDES. *Minoan ER: Progressive Entity Resolution in the Web of Data*, March 2016, 19th International Conference on Extending Database Technology, EDBT 2016, Poster [*DOI :* 10.5441/002/EDBT.2016.79], https://hal.inria.fr/hal-01411910

[11] M. PITTONI. *Online Identification of Last-Mile Throughput Bottlenecks on Home Routers*, UPMC, September 2016, Supervisors: Renata Teixeira (Inria) , Promethee Spathis (UPMC)Advisors: Anna-Kaisa Pietilainen (Inria) , Srikanth Sundaresan (Samsara Networks/ICSI), Nick Feamster (Princeton University), https://hal.inria.fr/hal-01401856

[12] N. VOUZOUKIDOU, B. AMANN, V. CHRISTOPHIDES. *Continuous Top-k Queries over Real-Time Web Streams*, October 2016, working paper or preprint, https://hal.inria.fr/hal-01411893