Metrics and models for Web performance evaluation

or, How to measure SpeedIndex from raw encrypted packets, and why it matters

QoE = f(QoS)

This talk

Dario Rossi
dario.rossi@huawei.com
Director, DataCom (*) Lab
(*) Data Communication Network Algorithm and Measurement Technology Laboratory
Metrics and models for Web performance evaluation
or, How to measure SpeedIndex from raw encrypted packets, and why it matters

QoE = f(QoS)

This talk

Dario Rossi
and, in alphabetical order, Alemnew Asrese, Alexis Huet, Diego Da Hora, Enrico Bocchi, Flavia Salutari, Florian Metzger, Gilles Dubuc, Hao Shi, Jinchun Xu, Luca De Cicco, Marco Mellia, Matteo Varvello, Renata Teixeira, Tobias Hossfeld, Shengming Cai, Vassillis Christophides, Zied Ben Houidi
Offering Good user QoE is a common goal
For ISPs/vendors encryption makes the inference harder. Detect/forecast/prevent QoE degradation is important!
Quality at different layers

- **User QoE**
- **Human influence factors**
- **System influence factors**
- **Context influence factors**

**Layers:**
- **Network**
- **Application**
- **User**
- **Context**
Quality at different layers

- **Network QoS**
  - Latency
  - Bandwidth
  - Wi-Fi quality

- **Application QoS**
  - Page load time (PLT)
  - SpeedIndex
  - Video bitrate
  - Mean opinion score (MOS)
  - User PLT (uPLT)

- **User QoE**
  - User PLT (uPLT)
  - Engagement metrics

**Context**
- Device type
- Activity
- Location

---

31/01/2020
Quality at different layers

User QoE influences Application QoS affects Network QoS

- Device type
- Activity
- Location

Context

User
- Mean opinion score (MOS)
- User PLT (uPLT)
- Engagement metrics

Application
- Load time
- Speed
- Interactivity
- Interactivity

Network
- Latency
- Packet loss
- Bandwidth
- Wi-Fi quality

HTTP/2, QUIC… (true for any other apps)
Metrics and models for Web performance evaluation
or, How to measure SpeedIndex from raw encrypted packets, and why it matters
Metrics and models for Web performance evaluation

or, How to measure SpeedIndex from raw encrypted packets, and why it matters

1. Data collection
   (Crowdsourcing campaign)

2. Models
   (Data driven vs Expert Models)

3. Browser metrics
   (Instant vs Integral vs Compound)

4. Method
   (From raw encrypted packets)
Mean opinion score (MOS)

"Rate your experience from 1-poor to 5-excellent"

User perceived PLT (uPLT)

"Which of these two pages finished first?"

User acceptance

"Did the page load fast enough?" (Yes/No)

Lab experiments
- Small user diversity, volunteers
- Web browsing, but artificial websites
- Artificial controlled conditions

Crowdsourcing (payed crowdworkers)
- Larger userbase, but higher noise
- Side-to-side videos ≠ Web browsing!
- Artificial controlled conditions

Experiments from operational website
- Actual service users
- Browsing in typical user conditions
- Huge heterogeneity (devices/browsers/nets)

Data collection: Crowdsourcing campaigns

https://webqoe.telecom-paristech.fr/data

(Award winning) dataset [PAM18]

Ongoing, with [WWW19]

Collab with
Models: Data driven vs Expert models

**Expert models**

- Fit predetermined $y = f(x)$
- $x =$ single scalar metric, generally Page Load Time (PLT)
- $f(.) =$ pre-selected by the expert

**Data-driven**

- Learn $y = f(x)$
- $x =$ vector of input features
- Optimal $f(.)$ selected & tuned by machine learning

---

**IQX Hypothesis**

$$QoE(x) = a e^{-\beta x} + \gamma$$


**Comparison of the two models in [QoMEX-18]**

**Weber Fechner**

$$MOS = \frac{4}{\ln(\text{Max} / \text{Min})} \left( \ln(\text{Session Time}) - \ln(\text{Min}) \right) + 5$$

Standard ITU-T G1030

[https://www.itu.int/rec/T-REC-G.1030/en](https://www.itu.int/rec/T-REC-G.1030/en)

**UserQoE**

[https://webqoe.telecom-paristech.fr/models](https://webqoe.telecom-paristech.fr/models)

**More flexible and (slightly) more accurate**

[INFOCOM19]

**Still room for improvement (see [WWW19] )**
Browser metrics: Time Instant vs Time Integral (1/2)
Browser metrics: Time Instant vs Time Integral (1/2)

www.iceberg.com

* Images by vstudio, vectorpocket, Ydlabs / Freepik
Browser metrics: Time Instant vs Time Integral (1/2)

www.iceberg.com

$1 - \int_0^t dx(t)$

$t=\text{ATF}, \text{ visible portion (aka Above the Fold) loaded}$

* Images by vstudio, vectorpocket, Ydlabs / Freepik
Browser metrics: Time Instant vs Time Integral (1/2)

$t = \text{ATF}$, visible portion (aka Above the Fold) loaded

www.iceberg.com

$\int (1 - x(t)) \, dt$

$x(t)$

* Images by vvstudio, vectorpocket, Ydlabs / Freepik
Browser metrics: Time Instant vs Time Integral (1/2)

www.iceberg.com

Mathematical expressions:

\[ \text{SpeedIndex} = \int (1 - x(t)) \, dt \]

Visual Progress

* Images by vvstudio, vectorpocket, Ydlabs / Freepik
Browser metrics: Time Instant vs Time Integral (2/2)

- **SpeedIndex, RUMSI, PSSI**
  - Processing intensive: ☒
  - Only at L7 (in browser): ☒
  - Visual progress metric: ☑

- **ObjectIndex, ByteIndex and ImageIndex**
  - Lightweight: ☑
  - ByteIndex also at L3 (in network): ☑
  - Highy correlated with SpeedIndex: ☑
  - Possibly far from user QoE?: ☑

where:

\[
X = \int_{0}^{t_{end}} (1 - x(t)) \, dt
\]
Browser metrics: Time Instant vs Time Integral (2/2)

- **SpeedIndex, RUMSI, PSSI**
  - Processing intensive
  - Only at L7 (in browser)
  - Visual progress metric

- **ObjectIndex, ByteIndex and ImageIndex**
  - Lightweight
  - ByteIndex also at L3 (in network)
  - Highly correlated with SpeedIndex
  - Possibly far from user QoE?

- \[ X = \int_{0}^{t_{end}} (1 - x(t)) \, dt \]
- \[ X^{\text{PLT}} \]
- Same PLT but slower loading
Browser metrics: Time Instant vs Time Integral (2/2)

- **SpeedIndex, RUMSI, PSSI**
  - Processing intensive ✗
  - Only at L7 (in browser) ✗
  - Visual progress metric ✓

- **ObjectIndex, ByteIndex, and ImageIndex**
  - Lightweight ✓
  - ByteIndex also at L3 (in network) ✓
  - Highly correlated with SpeedIndex ✓
  - Possibly far from user QoE? ?

![Graph showing SpeedIndex and ObjectIndex metrics]

**SpeedIndex** % of visual completeness (histogram, rectangles or SSIM)

**ObjectIndex** % of objects downloaded

**ByteIndex** % of bytes downloaded

**ImageIndex** % of bytes of images downloaded

Different cutoffs

$X = \int_0^{t_{end}} (1 - x(t)) \, dt$
Method: From raw packets to browser metrics (1/2)

Single session

Domain x.com

Individual objects

\[
\text{SpeedIndex} = \int (1 - x(t)) dt
\]
Method: From raw packets to browser metrics (1/2)

Domain x.com

Individual objects

SpeedIndex

Visual Progress

\[ \int (1 - x(t)) dt \]

Single session

Packets vs. time
Method: From raw packets to browser metrics (1/2)

Single session

Domain x.com

Individual objects

img1

img2
css
js
htm

Train ML models
(XGBoost, 1D-CNN)

Single session

Packets

time

1 − x(t) dt

SpeedIndex

Visual Progress

DOM

ATF

PLT

??!

??!

??!

??!
Method: From raw packets to browser metrics (2/2)

- Webpage rendering
  - User
  - 1 burst = 1 object
  - 1 color = 1 domain

- Browser (L7)

- Network (L3)
  - 1 burst = 1 packet
  - 1 color = 1 IP server
Method: From raw packets to browser metrics (2/2)

- Expert models
- Data-driven

User → Browser (L7) → Network (L3)

- Works with encryption
- Handle multi-sessions (not in this talk)
- Exact online algorithm for ByteIndex
- Machine learning for any metric
- Accurate on joint tests with Orange
- Accurate for unseen pages & networks
- Available soon into Huawei products
Aftermath (1/3): From raw packets to rough sentiments

- Expert-driven feature engineering
  - Explainable but inherently heuristic approach
  - Hard to keep in sync with application/network change

- Neural Networks
  - Less interpretable but more versatile
  - Downside: requires lots of samples....

- Feed NN with $x(t)$ signal
  - Still lightweight

- Feed NN using a filmstrip
  - More complex

- User feedback (e.g. MOS, user PLT, etc.)
- Smartphone sensors (e.g. happiness estimation via facial recognition)
- Brain signals acquired with sensors
- Activity of brain areas correlated with user happiness
Aftermath (2/3): Divide et impera

- World Wild Web
  - Huge diversity, not captured by single model

- Increase accuracy
  - Per-page QoE models
  - Inherently non scalable

- Increase accuracy & scalability
  - Per-page QoE models (eg Alexa top 100 pages)
  - Aggregate QoE models (eg 100 clusters top 1M)
  - Generic QoE model (for the tail up to 1B pages)
Aftermath (3/3): Keep collecting (and sharing) data

- Other applications/players are doing this already!

- Sustained continuous user QoE indication benefits
  - Useful samples for QoE management assessment, troubleshooting, regression detection, etc.
  - Get continuous stream of samples for improving QoE = f(QoS) models on the long run

- Very limited downsides (risk of annoying users if leveraging small panels)
https://webqoe.telecom-paristech.fr/

Documents

Datasets

Code

[SIGCOMM-19] Huet, Alexis and Houidi, Zied Ben and Cai, Shengming and Shi, Hao and Xu, Jinchun and Rossi, Dario, **Web Quality of Experience from Encrypted Packets** ACM SIGCOMM Demo, aug. 2019

[INFOCOM-19] Huet, Alexis and Rossi, Dario, **Explaining Web users QoE with Factorization Machines** IEEE INFOCOM Demo apr. 2019

[WWW-19] F. Salutari, D. Da Hora, G. Dubuc and D. Rossi **A large scale study of Wikipedia users’ quality of Experience** Proc. WWW, 2019


[QOMEX-18] Hossfeld, Tobias and Metzger, Florian and Rossi, Dario, **Speed Index: Relating the Industrial Standard for User Perceived Web Performance to Web QoE** 10th International Conference on Quality of Multimedia Experience (QoMEX 2018) jun. 2018

[PAM-18] D. da Hora, A. Asrese, V. Christophides, R. Teixeira and D. Rossi, **Narrowing the gap between QoS metrics and Web QoE using Above-the-fold metrics** Proc. PAM 2018, **Best dataset award ★**

[PAM-17] Bocchi, Enrico and De Cicco, Luca and Mellia, Marco and Rossi, Dario, **The Web, the Users, and the MOS: Influence of HTTP/2 on User Experience** Proc. PAM 2017

[SIGCOMM-QoE-16] E. Bocchi, L. De Cicco, D. Rossi, **Measuring the Quality of Experience of Web users**, ACM SIGCOMM Internet-QoE workshop 2016, **Best paper award ★**

60k+ real user grades

[Chrome plugin implementation]

9k real human grades

10k automated experiments

9k real human grades

60k+ real user grades

[Chrome plugin implementation]

[WebQoe](https://webqoe.telecom-paristech.fr/)
Thanks for lis