Measuring the Relationships between Internet Geography and RTT

Raúl Landa, Richard G. Clegg, Joao T. Araújo, Eleni Mykonioti, David Griffin, Miguel Rio

Department of Electronic and Electrical Engineering
University College London

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreements 248565 and 318205.
Problem Statement

- How much information about a given variable can we infer, given knowledge of a set of other variables?
Basic Measurement Technique

- We reimplement TurboKing
  - Estimate RTT by inserting measurement point in the middle of recursive DNS queries

- ~54k DNS Servers
- ~5.5k Autonomous Systems
- ~189 Countries
- ~200M individual RTT measurements
- ~19M full RTT estimations, ~10 measurements each
- ~50GB in size

### Histogram Axes

<table>
<thead>
<tr>
<th>Variable</th>
<th>Interpretation</th>
<th>Bins used</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_p$</td>
<td>Pairs of /8 prefixes</td>
<td>14,189</td>
</tr>
<tr>
<td>$X_a$</td>
<td>Pairs of Top AS numbers</td>
<td>215,392</td>
</tr>
<tr>
<td>$X_z$</td>
<td>Pairs of subcontinental zones</td>
<td>66</td>
</tr>
<tr>
<td>$X_c$</td>
<td>Pairs of countries</td>
<td>2,648</td>
</tr>
</tbody>
</table>

#### Quantitative Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Interpretation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_l$</td>
<td>Common prefix length</td>
<td>32</td>
</tr>
<tr>
<td>$X_d$</td>
<td>Great circle distance</td>
<td>300</td>
</tr>
<tr>
<td>$X_t$</td>
<td>Round trip time (RTT)</td>
<td>300</td>
</tr>
</tbody>
</table>

**TABLE II: Definition of variables used**
Data Modelling

• We model our data as a 7-dimensional discrete random variable $X$ with PDF $\Phi(x)$, so that

$$X = \{X_p, X_a, X_c, X_z, X_l, X_d, X_t\}$$

• We are interested in the conditional entropy $\mathcal{H}(X_\cdot | Y)$ of each component $X_\cdot$ given that a subset of variables $Y \subset X$ has been observed. Then, we have that

$$\mathcal{H}(X_\cdot | Y) = \sum_{x_\cdot, y} \Phi(x_\cdot, y) \log \left( \frac{\Phi(y)}{\Phi(x_\cdot, y)} \right)$$
Quantitative Data Modelling

• For quantitative variables, we consider estimators $\hat{X}_\bullet$ that approximate the value of $X_\bullet$ given that the values of variables in $Y$ are known.

• An optimal $\hat{X}_\bullet(Y)$ minimizes the **Mean Square Error**

\[
\text{MSE} = E\left\{ (\hat{X}_\bullet(Y) - X_\bullet)^2 \right\}
\]

• This is called the **Minimum MSE**, or MMSE.

• A lower bound for the MMSE can be found as

\[
\text{MMSE} \geq \frac{1}{2\pi e} \exp \left( 2\mathcal{H}(X_\bullet|Y) \right)
\]
Categorical Data Modelling

- For categorical variables, we consider an estimator \( \hat{X}_\bullet = g(Y) \) that “guesses” the value of \( X_\bullet \) given that the values of variables in \( Y \) are known.

- We are interested in the probability that this “guess” is incorrect
  \[
P_e(X_\bullet) = \Pr\{X_\bullet \neq \hat{X}_\bullet\}
\]

- A lower bound for \( P_e(X_\bullet) \) can be estimated using Fano’s inequality:
  \[
P_e(X_\bullet) \geq \frac{\mathcal{H}(X_\bullet|Y) - 1}{\log N}
\]
Conditional Entropy (RTT)

- The single variable that gives most information about RTT is the *country pair* of the endpoints.
  - This gives an RMSE $\geq \sim 162$ ms.

- If we consider the /8 prefix pair, the country pair and the geodesic distance, the remaining uncertainty is close to the minimum achievable with 4 variables or less (2.7 bits).
  - This gives an RMSE $\geq \sim 12$ ms.
Conditional Entropy (Subcontinental Zone)

- The single variable that gives most information about the subcontinental zone pair is the /8 prefix pair of the endpoints.
  - This gives a $P_e \geq \sim 0.09$.

- If we consider the /8 prefix pair, the geodesic distance and the RTT, the remaining uncertainty is close to the minimum achievable with 4 variables or less (.35 bits).
Conditional Entropy (Country Pair)

- Both the AS pair and the subcontinental zone pair give close to the maximum information about country pair in a single variable.
  - This gives a $P_e \geq \sim .35$.

- If we consider the AS pair, the subcontinental zone pair and the geodesic distance, the remaining uncertainty is close to the minimum achievable with 4 variables or less (.98 bits).
Conditional Entropy (Geodesic Distance)

- The single variable that gives most information about the geodesic distance is the *country pair* of the endpoints.
  - This gives a RMSE $\geq \sim 1600$ km.

- The single network variable that gives most information about the geodesic distance is the *AS pair* of the endpoints.
  - This gives a RMSE $\geq \sim 4000$ km.
Conditional Entropy (Geodesic Distance)

- If we consider the AS pair, the country pair, the common prefix length and the RTT, the remaining uncertainty is close to the minimum achievable with 4 variables or less.
  - This gives an RMSE $\geq \sim 94$ km.

- If we consider only network variables (AS pair, common prefix length and RTT), the remaining uncertainty is 3.25 bits.
  - This gives an RMSE $\geq \sim 416$ km.
Conditional Entropy (AS Pair)

- The single variable that gives most information about the AS pair is *country pair* of the endpoints.
  - This gives a $P_e \geq \sim .5$.

- If we consider the country pair, the common prefix length, the geodesic distance and the RTT, the remaining uncertainty is close to the minimum achievable with 4 variables or less (3.5 bits).
  - This gives a $P_e \geq \sim .2$. 

Conclusions

- Large-scale analysis of RTT and its related geographic properties
  - Novel RTT dataset comprising 19 million measurements between 54 thousand measurement points.
  - RTT measurements as realisations of a random vector (7D histogram)
  - Analysis using conditional entropy tells us how much information we get about a given variable if we know other variables

- Want the dataset? Get in touch!
Backup Slides
RTT-Distance Distribution
Unfolded RTT/Distance Density

$R^2_\Phi = 0.937$

- Speed of Light in Vacuum
- Speed of Light in Fibre
- $x_t = 0.016x_d + 22.3$
Large-scale Circuitousness Measures

- Large scale routing distance excess $\sigma$

$$\sigma(X_z) = D_z - X_d = 2(X^*_d - X_d)$$

- Total distance ratio $\rho$

$$\rho(X_z) = \left(\frac{.65c}{2}\right) \left[\alpha_U \left(1 + \frac{\sigma(X_z)}{X_d}\right) + \frac{\beta_U}{X_d}\right]$$
### Large-scale Circuitousness Measures

<table>
<thead>
<tr>
<th>Zone 1</th>
<th>Zone 2</th>
<th>$\sigma$ (km)</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia Pacific</td>
<td>Western Europe</td>
<td>7,410</td>
<td>3.08</td>
</tr>
<tr>
<td>Asia Pacific</td>
<td>Eastern Europe</td>
<td>9,796</td>
<td>3.7</td>
</tr>
<tr>
<td>Oceania</td>
<td>Western Europe</td>
<td>2,702</td>
<td>1.98</td>
</tr>
<tr>
<td>S. A. (East)</td>
<td>Asia Pacific</td>
<td>973</td>
<td>1.79</td>
</tr>
<tr>
<td>Asia Pacific</td>
<td>Central Asia/Middle East</td>
<td>11,348</td>
<td>3.94</td>
</tr>
<tr>
<td>Oceania</td>
<td>Eastern Europe</td>
<td>5,685</td>
<td>2.33</td>
</tr>
<tr>
<td>Asia Pacific</td>
<td>Indian Subcontinent</td>
<td>4,110</td>
<td>3.14</td>
</tr>
<tr>
<td>Oceania</td>
<td>Central Asia/Middle East</td>
<td>7,623</td>
<td>2.57</td>
</tr>
<tr>
<td>Africa</td>
<td>Oceania</td>
<td>18,973</td>
<td>4.53</td>
</tr>
<tr>
<td>S. A. (East)</td>
<td>South America (West)</td>
<td>7,187</td>
<td>4.77</td>
</tr>
<tr>
<td>Oceania</td>
<td>South America (West)</td>
<td>3,608</td>
<td>2.15</td>
</tr>
<tr>
<td>Africa</td>
<td>South America (West)</td>
<td>11,208</td>
<td>3.41</td>
</tr>
</tbody>
</table>
Large-scale Circuitousness Measures

- Eastern Europe to Oceania

$\sigma \approx 5,700 km$
Large-scale Circuitousness Measures

- Western Europe to Asia Pacific

\[ \sigma \approx 10,000 \text{ km} \]

\[ \sigma \approx 4,000 \text{ km} \]
Large-scale Circuitousness Measures

- Africa to Oceania

$6,000 < \sigma < 24,000$ km
Geolocation Errors

• Mismatch between latitude/longitude and country
  – Spatial index of the $\sim123k$ cities with more than 1k inhabitants

• Since we are focusing in large-scale geography, small geolocation errors are unimportant
Least Squares Median Line

\[ \Phi(x \mid x_t) \]

- How well does a function \( f(x_d) \) approximate the median of the marginal distribution?

\[
E_{\Phi}^2(f(x_d)) = \int_0^\infty \left( \int_0^{f(x_d)} \Phi_{(d,t)}(x_d, x_t) \, dx_t - \int_0^{f(x_d)} \Phi_{(d,t)}(x_d, x_t) \, dx_t \right)^2 \, dx_d
\]

- We are interested in the special case where \( f(x_d) \) is linear. We formulate the following optimisation problem.

Minimise: \[ E_{\Phi}^2(\alpha x_d + \beta) \]
Least Squares Median Line

- We require a goodness of fit measure. We simply take the $R^2$ statistic and reformulate it. Let $m$ be the median RTT for a given $x_d$, then we have that:

$$R^2_\Phi = 1 - \frac{E^2_\Phi(\alpha x_d + \beta)}{E^2_\Phi(\hat{m})}$$

$$0 < R^2_\Phi < 1.$$

- Similarly to $R^2$, Moreover,

- $R^2_\Phi = 0$: The linear fit accounts for essentially no data variability

- $R^2_\Phi = 1$: The linear fit perfectly explains all median RTT variability for each $x_d$. 

26