The Large-Scale Geography of Internet Round Trip Times

Raul Landa, João Taveira Araujo, Eleni Mykoniati, Richard G. Clegg, David Griffin, Miguel Rio
Department of Electronic and Electrical Engineering, University College London
Email: {r.landa, j.araujo, e.mykoniati, r.clegg, d.griffin, m.rio}@ee.ucl.ac.uk

***Abstract*—One of the challenges that distributed systems designers face is that their performance is very sensitive to the characteristics of the underlying network. Hence, simple models that accurately describe some statistical properties of this substrate can be very helpful in the modelling, simulation and design of Internet-scale distributed systems.**

In this paper we present a model for the analysis of Internet round trip times (RTT) and its relationship with geolocation distance. This model is based on a novel RTT dataset comprising ∼19 million measurements between ∼54 thousand measurement points. This model can then be used to accurately predict median RTT for a given geographical distance.

Our main contribution is a procedure for the geographic analysis of RTT that allows the recovery of large-scale routing information. We accomplish this by investigating RTT on the basis of disjoint, large-scale geographic components. By applying a novel median-based, least-squares fitting algorithm to traffic flows between these components, we analyse their RTT\distance behaviour and compute a *large-scale routing excess* that quantifies the extra distance beyond the great circle that packets traverse when they flow between large-scale geographic areas.

I. Introduction

One of the difficulties in the design and modelling of Internet-scale distributed systems is that that their performance can be markedly affected by the performance characteristics of their underlying network substrate. Unfortunately, the sheer complexity of the Internet precludes a full model or simulation of these characteristics. It is however possible to use network measurements to develop scalable models based on specific statistical properties of the Internet. In this paper we focus on *Round Trip Time* (RTT) because it is both simple to measure and interpret, and widely used in the design of performance-sensitive applications such as network anycast [18], [30] or media content delivery [11], [34]. To address the relationship between geographical distance and RTT we collected one of the most comprehensive sets of Internet delay measurements available today, comprising more than 200 million individual RTT samples taken between ∼54 thousand measurement points. These measurements were then used to generate a model that can explain ∼94% of variability in median RTT.

Our work goes beyond previous contributions in the topic by explicitly considering the large-scale routing of Internet flows and its effect on RTT. For instance, many earlier works treat RTT as a topological path property to be estimated in isolation [15], [26], [27], and disregard additional information such as geolocation. Although other works have analysed the relationship between RTT and interdomain routing [33], [36], [33] or its geographical properties [21], [28], their dependency on geolocation-aided *traceroute* surveys has limited them to specific world regions (typically North America, Europe and Asia Pacific). We improve upon these works by significantly expanding the scope of data collection. Our dataset provides sufficient geographic diversity to to uncover geographic properties of RTT relevant for the entire Internet.

Our contributions are twofold. First, we present a data-centred model for the analysis of RTT and its large-scale geographic properties that treats measurements as realisations of a multidimensional random variable $\mathbf{X}$. We build an approximation for $\Phi(\mathbf{x})$, the density of $\mathbf{X}$, by collecting an RTT dataset that includes ∼200 million RTT samples between ∼54 thousand measurement points and creating a *contingency table* (the multidimensional equivalent of a histogram).

Second, we present a procedure for the geographic analysis of $\Phi(\mathbf{x})$ that allows the recovery of large-scale routing information. We achieve this by splitting $\Phi(\mathbf{x})$ into 66 individual contributions, each representing the RTT\distance behaviour between two large scale geographic areas denoted as *subcontinental zones*. By applying a novel median-based linear fitting algorithm to each one of these components, we characterise their RTT behaviour as either *increasing* or *decreasing* with great circle distance (i.e. the distance over the spherical surface of the Earth). Then, we hypothesise that those components with decreasing RTT behaviours are the result of routing paths that significantly deviate from the great circle segment connecting the relevant subcontinental zones, and propose a coordinate transformation that maps these components so that they manifest *increasing* RTT behaviour. This process yields a *routing distance* for every *decreasing* component that describes the actual distance that packets flowing between the relevant subcontinental zones need to traverse. We then compare the routing distances found for all *decreasing* components with information directly obtained using geolocated *traceroute* probes, and find agreement between the two. This allows us to present a purely linear model relating median RTT and routing distance, and to show that it explains ∼94% of the variability in median RTT as a function of geolocation distance that we observe in our dataset.

The structure of the paper is as follows. We commence in §II, where we detail our measurement methodology, and continue in §III by presenting our modelling and analysis. In §VI we present other research contributions that relate to ours, and our conclusions in §VII.

II. Measurements

A comprehensive study of the geography of Internet RTTs requires the measurement of round trips between a large number of measurement points in as many distinct geographical locales as possible. Due to their ubiquitous nature, DNS servers are ideal for this purpose; this led us to select TurboKing [24] as our main measurement technique.

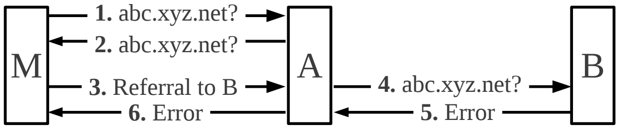


Fig. 1: Operation of TurboKing

The basic operation of TurboKing is presented in Fig. 1. Consider two DNS servers, **A** and **B**, along with the TurboKing measurement point **M**. Measurement proceeds as follows. First, **M** behaves as a DNS client, and sends a resolution query to **A** for a hostname for which **M** itself is registered as authoritative. **A** forwards the query to **M**, which allows **M** to estimate the RTT between itself and **A**. Measurement continues with **M** taking the role of a DNS server which informs **A** that **B** can resolve its original query. This causes **A** to forward the query to **B**. When **B** receives this request, since it pertains to a hostname for which it has no records, it responds with an error which is recursively sent to **M** via **A**. This allows **M** to estimate the RTT between itself and **B** via **A**. Along with the first measurement taken, the RTT between **M** and **A**, this second measurement allows **M** to calculate the RTT between **A** and **B**.

The RTT data set presented in this paper was obtained using this technique. Data collection took place from May 2011 to February 2012, and comprises ∼200 million individual RTT samples between ∼54 thousand recursive, non-forwarding DNS servers. This raw data was processed to remove short-term variations and caching effects (see §II-B), yielding ∼19 million RTT measurements. We now describe the two main phases of our data collection effort: the collection of a set of appropriate servers, and the use of this set to measure RTTs.

A. Collection of Measurement Servers

Our collection engine operates by taking IP addresses or hostnames as inputs, performing DNS queries for them, and extracting candidate DNS server hostnames from the **SoA** or **NS** records in the responses; we also make use of related **A** records if present [29]. Both forward and reverse queries were performed as required. All the candidate DNS server hostnames that our collector obtained were then tested for their usefulness as TurboKing measurement points, and their backup servers were used to further seed the set of input IP addresses or hostnames. Thus, every seed IP address or hostname became the starting point for a recursive traversal of the the graph of DNS severs pointing to each other as backups through **SoA** or **NS** records. Unnecessary queries were avoided by keeping a database of all IP addresses, hostnames and domains for which DNS queries had been performed; unnecessary load on the DNS system was avoided by rate-limiting.

To ensure adequate coverage, we provided diverse starting points to the aforementioned graph traversal procedure. We used the following as seeds for our collection algorithm.

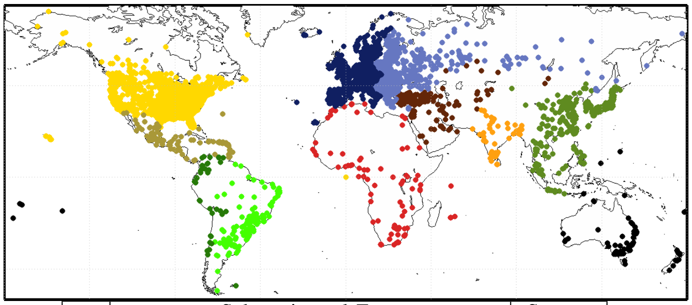
- **IP addresses**:
  - *Forward DNS*: Addresses obtained from forward DNS lookups of found hostnames.
  - *Random*: Addresses obtained by generating random 32-bit integers in the routable IP space.
  - *DNS Server Lists*: Addresses obtained from DNS server lists (e.g. among others, those included in DNS performance comparison applications).
  - *iPlane*: Addresses extracted from iPlane [26] data.
  - *Firewall and BitTorrent Blocklists*: Lists of addresses in specific categories (enterprise, residential, government, etc.) We ensured that we had representatives from all categories.
  - *BitTorrent ANNOUNCE Messages*: IP addresses of BitTorrent peers participating in the 500 most popular torrents in popular sites (e.g. The Pirate Bay, ISOHunt). We periodically queried all trackers for each swarm and eliminated all duplicates.
- **Hostnames**:
  - *Reverse DNS*: Hostnames obtained from reverse DNS lookups of found IP addresses.
  - *Web Search*: Hostnames obtained from search engines. This processed started with a set of ∼40 word lists, selected to span a large selection of topics and languages. Four-word sentences were created using words from the same language, and submitted to three different web search services. Webserver hostnames were then extracted from the URLs present in the search results.
  - *Webserver hostnames*: Hosts and domains present on various *top N* website lists.

This process was performed only for a single period of four weeks prior to the the beginning of the collection process. From that point on, the only process adding DNS servers into the database was the detection of forwarders; i.e. DNS servers that responded to queries that were originally issued to other DNS servers. The process yielded a total of ∼350,000 DNS servers, out of which ∼54,000 were non-forwarding, recursive servers useful as measurement points.

B. Collection and Processing of measurements

In order to ensure a representative sample, each endpoint of an RTT measurement was selected randomly from the available set of DNS servers; the only check made was to ensure that the two were distinct. Each RTT measurement consisted of 10 individual samples spaced 10 seconds apart. The median of these estimates is taken as the RTT measurement; this allowed the system to filter out any episodic RTT effects. Due to this procedure, our collected ∼200 million samples yielded a net ∼20 million RTT measurements. Additional processing was required to remove spurious results arising from DNS servers overriding the DNS Time-To-Live (TTL) [29] values present in the measurement queries with their own. Since TurboKing relies on the whole measurement path being traversed, its DNS messages must always have a TTL of zero. If this value is overridden by a server, the second hop of the measurement is eliminated, leading to RTT estimates close to zero. Although this behaviour was not reported in [24], we found that it affected ∼7% of our measurements, and could lead to a underestimation of the true RTT between two measurement points. We eliminated this error by making use of a very simple heuristic that relies on the fact that all our DNS servers were geolocated. If the RTT estimated for a given pair of DNS servers implied a signal propagation greater than the speed of light in vacuum (299.792 km/ms), the measurement was discarded as spurious. After removing these measurements, the dataset was reduced to ∼19 million RTT measurements.

C. *Dataset Coverage*



	Subcontinental Zone	Servers
🟥	Africa	519
🟫	Central Asia and the Middle East	1,490
🟩	Asia Pacific and China	7,730
🟧	Indian Subcontinent	449
🟨	North America (North)	21,276
🟫	North America (South) and the Caribbean	526
⬛	Oceania	1,116
🟩	South America (West)	270
🟩	South America (East)	1,333
🟦	Eastern Europe	6,798
🟦	Western Europe	12,953

TABLE I: Geographic distribution of measurement servers

Since we were interested in the relationships between RTT and geographic distance, we made extensive use of IP geolocation databases. Specifically, we used both MaxMind GeoLite and [5] and Neustar IP Geolocation [6]. To assign measurement points to subcontinental zones, we were interested not only in associating latitude and longitude data with IP addresses but also with the country in which servers were located. Unfortunately, we found that geolocation databases frequently associate an IP address not with the country in which it is physically located (as defined by its latitude and longitude), but with the country associated with its administrative aspects (e.g. the country of the entity that owns it). This led to gross errors at the country level, where IP addresses associated with country were placed in another one by their latitude/longitude. To avoid these errors, we loaded the GeoNames city names database [4] to a spatial index [20] and used it to resolve IP addresses to cities through their latitude and longitudes only. This resulted in 122,952 place bindings representing cities with 1,000 inhabitants or more, which we then used to obtain the country associated with a geolocated IP address. This greatly reduced these gross errors. Regarding errors in the latitudes and longitudes themselves, although these are non-negligible [32], we assume them to be accurate at least to the country level. Hence, we expect these errors to be relatively small when compared to the media propagation delay, which dominates RTT at a global scale.

As indicated by our geolocation and AS lookup sources, measurement endpoints for our dataset were present in 5,455 autonomous systems over 3,384 cities and 189 countries. According to [1], 99.6% of global Internet users belong to a country where there is at least one measurement point. To verify the IP coverage of our measurement endpoints, we compared it with a daily routing table snapshot obtained from the RouteViews server in WIDE [10]. Of the 216,344 prefixes received at that point, 20,881 included at least one measurement server. Overall, 476 million addresses (∼32% of 1.017 billion routable IP addresses) belong to network prefix where there is at least one measurement point.

III. Modelling Preliminaries

In order to better elucidate the large-scale effects that Internet routing has on RTT, we divided our measurement endpoints into 11 *subcontinental zones* consisting of geographically adjacent countries, as shown in Fig. I; this leads to a set of 66 distinct zone pairs. To map an IP address to a zone, we first map it to a country, which is then mapped to a subcontinental zone. Since RTT measurements are always performed between two distinct hosts, all our measurements are defined for distinct *pairs* of geolocated hosts. We will focus our analysis on two variables. The first one is the geodesic (*great circle*) distance in kilometres between the geolocated positions of both hosts; this will be denoted X_d. The second one is the Internet RTT (Round Trip Time) in milliseconds, denoted as X_t. To simplify the analysis, we discretised X_d and X_t to 300 bins, yielding a resolution of 67 km and 3.3 ms per bin for X_d and X_t respectively.

We model our data as a multidimensional discrete random variable $\mathbf{X} = \{X_d, X_t\}$. Each measurement in our data set is then one sample of $\mathbf{X}$, whose joint probability distribution will be denoted as $\Phi(\mathbf{x})$. We construct an empirical probability density for $\Phi(\mathbf{x})$ simply by considering a *contingency table* [17] with columns indexed on the distinct bins of X_d, rows indexed on those of X_t, and holding the frequency in which value pairs (X_d, X_t) are observed in the dataset. Since we choose the endpoints for each one of our measurements randomly, we expect this contingency table to converge in frequency to the appropriate density $\Phi(\mathbf{x})$.

A. *Least-Squares Median Line Fit*

Since Internet RTTs are prone to outliers, estimators based on medians tend to be much more robust than those based on means. Furthermore, prior work has shown improved linear fit when regressing the medians of the RTT distributions at given great circle distances [11]. Rather than using simple linear regression on the medians, for this paper we develop a lightweight, median-based least-squares fitting technique that approximates an underlying probability density.

To assess how well does a candidate function $x_t = f(x_d)$ that relates the RTT x_t and great circle distance x_d approximates the median of $\Phi(x_d, x_t)$ for each x_t, we propose the use of the *median distribution error* functional

$$E^2_\Phi(f(x_d)) = \int_0^\infty \left(\int_0^{f(x_d)} \Phi(x_d, x_t) dx_t \right.$$
$$\left. - \int_{f(x_d)}^\infty \Phi(x_d, x_t) dx_t \right)^2 dx_d.$$

Then, the *RTT\distance approximation problem* can be formulated as a variational problem in which we seek a function $x_t = f(x_d)$ that minimises $E^2_\Phi(f(x_d))$. Formally:

$$\underset{f(x_d)}{\text{Minimise:}}\ E^2_\Phi\left(f(x_d)\right). \tag{1}$$

Following [11], [22], we are interested in a constrained solution of (1) where $f(x_d)$ is linear; that, is, where $x_t = \alpha x_d + \beta$. In that case, (1) reduces to finding the optimal α and β that minimise the median squared error E^2_Φ.

Since we will be comparing different fits, we need a measure for the *goodness of fit* that a given α and β provide. To this end, we define a goodness of fit measure that is analogous to the *coefficient of determination* R^2 usually used in simple linear regression. This measure, which we denote as R^2_Φ, is defined as

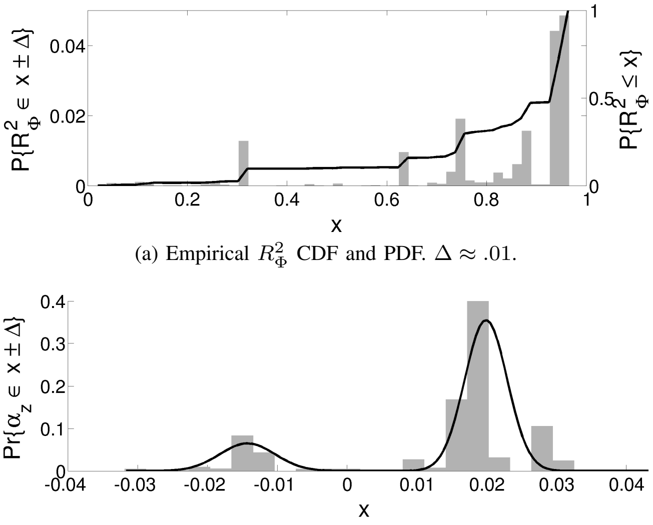
$$R^2_\Phi = 1 - \frac{E^2_\Phi(\alpha x_d + \beta)}{E^2_\Phi(\hat{m})}, \tag{2}$$

where $\hat{m}$ is the solution to (1) for a constant RTT $x_t = \hat{m}$. Hence, $E^2_\Phi(\hat{\beta}) \in (0, 1)$ represents the natural variability around the median RTT, irrespective of geodesic distance. Then, R^2_Φ represents the proportion of variability in the median that is accounted for by the median least squares fit $x_t = \alpha x_d + \beta$. Similarly to the R^2 statistic, $R^2_\Phi = 0$ implies that the linear fit accounts for essentially no variability in the data, and $R^2_\Phi = 1$ indicates that the linear fit perfectly explains all variability in the median RTT for each x_d as embodied by the underlying distribution $\Phi(x_d, x_t)$.

IV. RTT and Great Circle Distance

A visual representation of the relationship between RTT and great circle as revealed by $\Phi(x_d, x_t)$ is shown as a greyscale map in Fig. 4a, where pixel shading is proportional to the logarithm of the number of observations in the dataset for each (X_d, X_t) bin. As previously reported [18], [11], [22], we find a strong linear association between RTT and geolocation distance. This has been confirmed not only using PlanetLab [2], but also by the CAIDA macroscopic topology probing project [12]. However, as clearly visible in Fig. 5 of [18], Fig. 1 of [22] and Fig. 4a in this paper, $\Phi(x_d, x_t)$ also exhibits significant deviations from purely linear behaviour. We now show that these deviations arise due to large-scale routing behaviours between subcontinental zones. To this end, we introduce the subcontinental zone pair X_z of each measurement as an additional variable in $\Phi(\mathbf{x})$, and analyse the spatial properties of $\Phi(x_d, x_t, x_z)$ by considering the relationships between distance and RTT for each one of the 66 sub-components that arise between pairs of subcontinental zones X_z in Table I.

We proceed by isolating each subcontinental component $\Phi(x_d, x_t, x_z)$ for a given x_z and fitting a least squares median line $x_t = \alpha_z x_d + \beta_z$ to each one. To achieve this, we take the optimisation problem (1), discretise it, and solve it using a standard simplex search numerical optimisation technique [31]. We present some examples of these components in Fig. 3, along with their calculated least squares median lines and their goodness of fit measures R^2_Φ. At first glance, we note that although some components display strong linear behaviour (e.g. Figs. 3aand 3c), some exhibit very poor fit to a linear median model (e.g. Fig. 3b), suggesting that no underlying regularities are present. However, if we consider the proportion of measurements present on each component (and hence the number of measurement servers in the zones being considered), a significant pattern arises. First, in Fig. 2a we see that components accounting for ∼89% of all measurements have values of R^2_Φ of .62 or higher, thus suggesting that most measurements come from components with strong linear behaviour. Furthermore, in Fig. 2b we see that most measurements belong to components whose least squares median slope α_k is close to either +.017 or -.017.



(a) Empirical R^2_Φ CDF and PDF. $\Delta \approx .01$.

(b) Empirical PDF of α and 2-component GMM fitting. $\Delta \approx .0015$.

Fig. 2: Least-square median line slopes α_z and R^2_Φ for all 66 subcontinental zone components $\Phi_{\mathbf{S}(d,t,z)}$. In both graphs, a histogram bin width of 2Δ was used.

The fact that the α_z gave rise to a pair of symmetrically positioned distributions suggests that $|\alpha_z|$ may be strongly influenced by the same underlying process for every X_z, with the most important distinction between components being its sign. To explain the sign of α_z, we propose a simplified model of routing in which communication between subcontinental zones takes place through a single route which may deviate significantly from the relevant geodesic. Although this illustrative model is not realistic, it captures essential intuitions useful to develop techniques which can then be falsified using *real* Internet measurements.

Consider two subcontinental zones $\mathcal{Z}$ and $\mathcal{Z}'$, diagrammatically represented by circles in Figs. 3d, 3e and 3f. For our simplified one-link model, we posit that the sign of α_z will depend on the geographic relationship between the points at which the link connects the two zones. Consider two pairs of nodes $[a, a']$ and $[b, b']$, where $[a, a']$ denotes the two closest nodes in geodesic distance, and $[b, b']$ represents the two most further away. We now consider three different alternatives with respect to the position of the inter-zone route. In Fig. 3d we present the case in which the route between both zones connects them through points which are close in geodesic distance. Then, a $[a, a']$ will experience low RTT, and $[b, b']$ will experience high RTT. In this case, RTT will increase with great circle distance (as happens with Fig. 3a).). Consider now a situation like that presented in Fig. 3e. In that case, we see that although there may be significant differences in great circle distance between $[a, a']$ and $[b, b']$, their RTTs are very similar, and the result is an $\alpha_z \approx 0$, such as that presented in Fig. 3b. Finally, we consider a situation like that presented in Fig. 3f. In this case, we see that $[a, a']$ will experience high RTT due to its increased routing distance, and $[b, b']$ will experience low RTT. This pattern will lead to a general decrease of RTT with geolocation distance, and hence to a negative α_z such as presented in Fig. 3c. One particular way in which this can arise in practice is if the route taken reaches the destination zone by *wrapping around* the globe. As an illustrative example of how this might arise, consider a pair of nearly antipodal points which take a data path 180 degrees from the great circle arc direction. The path length is almost the same because the points are antipodes but a move which reduces the great circle distance increases the data path distance.

Of course, the question remains of what happens in the more realistic case in which there is more than a single link connecting the two zones. Then, each link will contribute its own RTT\distance profile, depending on which hosts use which links to reach the remote zone. Although this can lead to relatively simple behaviour such as that presented in Fig. 3b, it can also lead to more complex behaviours. However, as visible in Figs. 2a and 2b, for those zone pairs that have strong effects on the overall RTT\distance behaviour of the Internet, we find strongly linear behaviours of the types found in Figs. 3d and 3f. Hence, although actual routing behaviour is much more complex, we can approximate the Internet-scale relationship between RTT and distance by assuming that there is a single main large-scale geographic route connecting these most influential pairs of subcontinental zones. This will provide a simple algorithm to recover the geographic routing distance between them.

GMM Element	Weight	Mean	Variance
Increasing (C_I)	0.1544	-0.0143	3.129x10^{-5}
Decreasing (C_D)	0.8456	0.0199	1.841x10^{-5}

TABLE II: Gaussian Mixture Model parameters for α_z

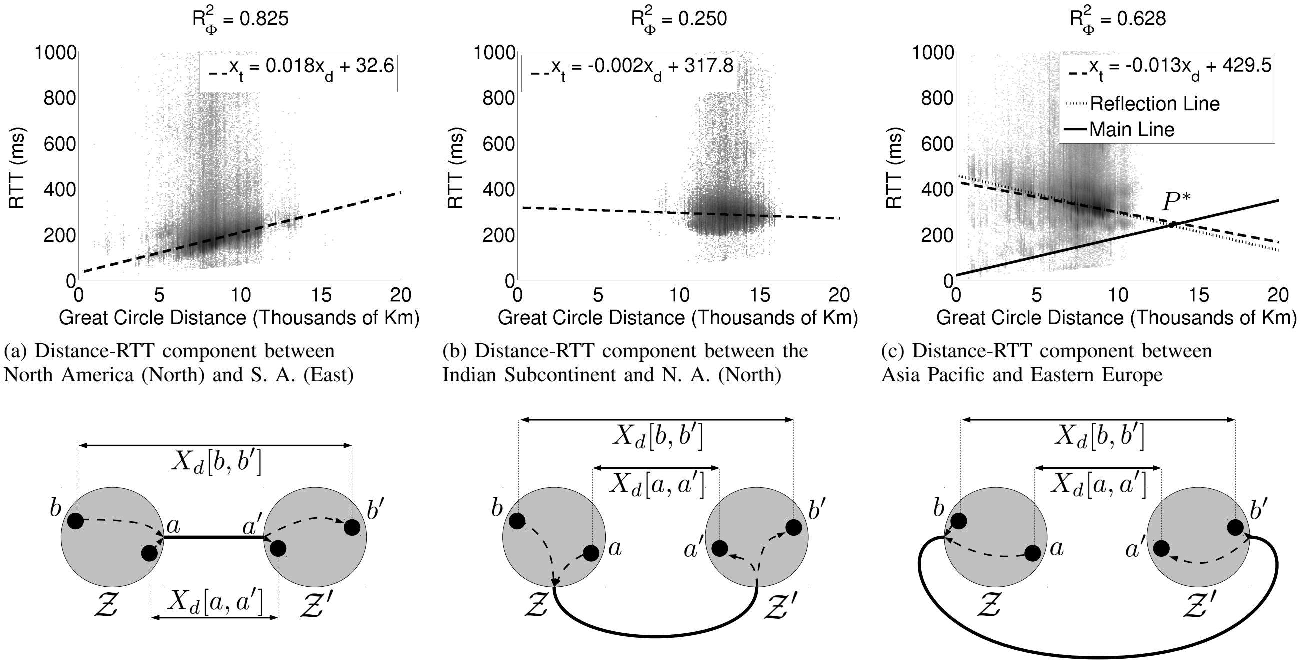
A. Estimating the Excess Routing Distance

We now propose a mechanism to recover the routing distance excess beyond the great circle that packets traverse when routed between specific pairs of subcontinental zones. To this end, we introduce a model which captures (and corrects) the fact that, in many cases, when points move closer along a great circle, the RTT increases. This is done using the equivalent of an *unfolding* on the globe. Taking Fig. 2b as a starting point, an explanatory model could create two sets of RTT measurements, one with best fit gradient α_F and one with $-\alpha_F$. This is suggested by Fig. 2, which shows that the dataset can be roughly decomposed in two sets with opposite signs but roughly the same absolute magnitude for their expected α. Let $\mathbf{S}_I$ be the set of components X_z associated with *increasing* (positive) slopes, and $\mathbf{S}_D$ the set of components X_z associated with *decreasing* (negative) slopes. In order to assign each component to each one of these sets, we performed a Gaussian Mixture Model fit with two components to the experimental α_z using Bregman soft clustering [13]. The two resulting Gaussian components C_I and C_D are shown in 2b, and their associated parameters are shown in Table II. Given the weights for each Gaussian component, we assigned components $\Phi(x_z, x_t, x_d)$ to either an *increasing* or a *decreasing* set, so that the ratio in their probability mass was as close as possible to that between the weights of GMM components C_I and C_D. This allowed us to robustly identify each component $\Phi(x_z, x_t, x_d)$ with either a positive α_F (for those in the *increasing* set) or a negative one (for those in the *decreasing* set). The *decreasing* set resulting from this assignment is shown in Table III.

To accurately estimate the value of α_F, we create a density $\Phi_I(x_t, x_d)$ that only contains those components associated with $\mathbf{S}_I$. We then perform a least-squares median line fit for this density, obtaining a slope $\alpha_F = 0.0164$, an intercept $\beta_F = 22.053$ ms and a goodness of fit $R^2_\Phi = .937$. Then, we assign a slope of $-\alpha_F$ to those components in $\mathbf{S}_D$, and use a restricted formulation of (1) to find an optimal fit to the marginal density $\Phi_D(x_d, x_t, x_z)$ where only the intercept β'_z is optimised. This allows us to associate the set $\mathbf{S}_I$ with a best approximation $x_t = \alpha_F x_d + \beta_F$ henceforth called the *main line*, and each element $X_z \in \mathbf{S}_D$ with a best approximation $x_t = -\alpha_F x_d + \beta'_z$ henceforth called the *reflection line*. These two are shown in Figs. 3c, along with the unrestricted best-fit median line $x_t = \alpha_z x_d + \beta_z$.

For each $X_z \in \mathbf{S}_D$, we find the intersection point $P^* = (X^*_d, X^*_t)$ between its reflection line and the main line (see Fig. 3c). Once P^* has been found, the density $\Phi_D(x_d, x_t, x_z)$ is *reflected* through the great circle distance X^*_d; that is, its great circle distance values are X_d mapped to a new *large-scale routing distance* value D_z so that

$$D_z = 2X^*_d - X_d. \qquad (3)$$



(a) Distance-RTT component between North America (North) and S. A. (East)

(b) Distance-RTT component between the Indian Subcontinent and N. A. (North)

(c) Distance-RTT component between Asia Pacific and Eastern Europe

(d) One-link model for $\alpha_z > 0$

(e) One-link model for $\alpha_z \approx 0$

(f) One-link model for $\alpha_z < 0$

Fig. 3: RTT\distance behaviour for 6 X_z example components, along with their respective one-link models

This operation allows us to calculate the appropriate routing distance such that the behaviour of the densities in $\mathbf{S}_D$ best matches the behaviour of those in $\mathbf{S}_I$. The intuitive meaning of this procedure can be seen in Fig. 3f: from the point of view of geodesic distance, the path taken by traffic from a to a' is *folded*, and can be transformed to resemble the path in Fig. 3d by *reflecting* each circle over the X_d axis and increasing the distance between them so that the total length of the link is preserved. Hence, (3) essentially *unfolds* the routing path so that distance along the path can be directly compared to an *increasing* (positive slope) path like that of Fig. 3d.

It may be noted that, since the reflection procedure was made on the basis of whole subcontinental zones rather than smaller routing units, there is some degree of noise introduced by the procedure. In particular, the "superluminal" density components in Fig. 4b (those that lie below the line marked as *speed of light in vacuum*) are artifacts of the reflection procedure that correspond to routes between Eastern Europe and Asia Pacific that follow a path which more closely resembles the geodesic. In fact, these components can be seen on Fig. 3c around the main line defined from $\mathbf{S}_I$. Although this noise decreases the quality of the resulting density, its probability mass is small, and its effect is limited.

The outcome of the unfolding procedure is shown in Fig. 4, along with its least squares median fit, which has a slope $\alpha_U \approx 0.016$ km/ms and an intercept $\beta_U \approx 22$ ms. As indicated by its R^2_Φ, this fit explains ~94% of the variability in median RTT as a function of routing distance, and hence constitutes a very accurate description of the structure of $\Phi(\mathbf{x})$.

V. Circuitousness of Internet Routing

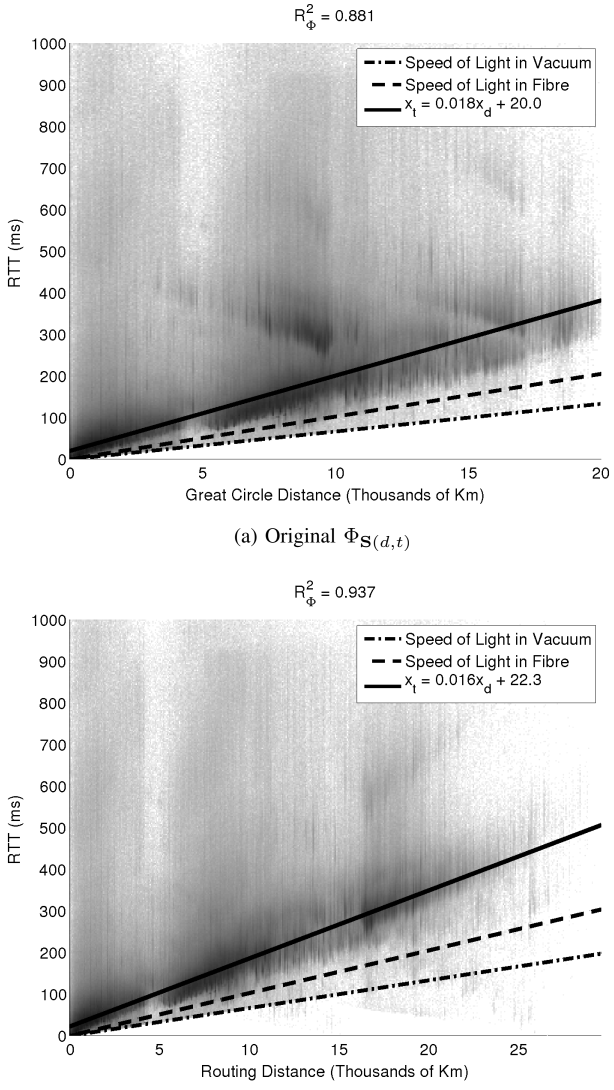
To better understand the effects of the unfolding procedure, we define two measures. The first one is the *large-scale routing distance excess* σ, defined as

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

which gives an estimate for the *additional* distance beyond the geodesic that a packet traverses when it is routed between two subcontinental zones $X_z \in \mathbf{S}_D$. The second measure, the *total distance ratio* ρ, is defined as the ratio between the distance that a packet could have traversed in half an RTT moving at a speed $v = .65c$, and the actual geodesic distance that is associated with that RTT in the least-squares median fit shown in Fig. 4b. Formally,

$$\rho(X_z) = \left(\frac{.65c}{2}\right)\left[\alpha_U(1 + \frac{\sigma(X_z)}{X_d}) + \frac{\beta_U}{X_d}\right], \qquad (4)$$

where $\sigma(X_z) \equiv 0$ if $X_z \in \mathbf{S}_I$ and $.65c$ approximates the speed of light in optical fibre. The values for σ and ρ for all $X_z \in \mathbf{S}_D$ are shown in Table III. The rationale behind these definitions for ρ and σ is that they quantify path *circuitousness*, the degree to which routing paths deviate from geodesic paths. However, whereas σ quantifies large-scale deviations at the subcontinental scale, ρ directly quantifies the intuition that Internet RTTs increase with distance significantly faster than what would be expected from propagation delay alone if they followed geodesic paths. A further benefit of these measures is to provide a tool to evaluate whether direct corroborating evidence of the values in Table III can be found in the current Internet. We address ρ first, then moving on to σ.



(a) Original $\Phi_{\mathbf{S}(d,t)}$

(b) Unfolded $\Phi_{\mathbf{S}(r,t)}$

Fig. 4: Relationship between RTT and distance

Most work addressing path circuitousness has required the explicit calculation of routing paths, normally using geolocated *traceroute* probes. Hence, most previous work has relied on C, the ratio between explicit routing path length and geodesic distance, as a circuitousness measure. Since we do not have full path length measures, we limit ourselves to qualitative comparisons between C and ρ. Kasiviswanathan et. al. [21] present evidence of very large values of $C \approx 10$ within the USA, with traffic *crisscrossing* between the East and West coasts. Although Subramanian et. al. [33] also find instances of large C within the USA, they show that it tends towards 1 for longer paths. However, since their dataset focused on the USA and Europe, their conclusions apply primarily between these zones. In [28], Mátray et. al. show that, for flows between Europe, the USA and the Asia Pacific regions, values of $C \geq 3$ are only frequent for geographic path lengths under 2000 km; for longer paths, they report $1 \leq C \leq 2$. Only connections between Europe and Asia Pacific experience $C > 1.5$.

Zone 1	Zone 2	σ (km)	ρ
Asia Pacific	Western Europe	7,410	3.08
Asia Pacific	Eastern Europe	9,796	3.7
Oceania	Western Europe	2,702	1.98
S. A. (East)	Asia Pacific	973	1.79
Asia Pacific	Central Asia/Middle East	11,348	3.94
Oceania	Eastern Europe	5,685	2.33
Asia Pacific	Indian Subcontinent	4,110	3.14
Oceania	Central Asia/Middle East	7,623	2.57
Africa	Oceania	18,973	4.53
S. A. (East)	South America (West)	7,187	4.77
Oceania	South America (West)	3,608	2.15
Africa	South America (West)	11,208	3.41

TABLE III: Large-scale circuitousness measures

By construction, (4) will have lower values for higher values of X_d; this is consistent with the properties of C reported in [21], [33], [28]. For traffic between North America (North) and Western Europe, (4) gives a value of $\rho \approx 1.8$, which is higher than the reported $C \approx 1$ because it includes effects such as store-and-forward, processing and queuing delays which are not directly attributable to geographic routing length. For traffic between Europe and Asia Pacific, (4) predicts higher values of ρ due to increased routing excess σ; this is in agreement with [28]. One important difference between our results and those of [33] is that, because of the higher geographical heterogeneity of our dataset, we could document many more instances of large-scale routing inefficiency.

We now move on to the evaluation of evidence for the found values of σ. The methodology that we followed was to perform a small number of geolocated traceroute measurements between hosts in the relevant areas, and use them to reconstruct the geographical routes taken by the packets. This was achieved by making use of public traceroute resources such as [9]. From this, we obtained representative values of σ that could be compared with those presented in Table III.

Although we found evidence of some traffic between Eastern/Western Europe and Asia Pacific being routed through the USA, implying routing distances D_z of ∼19,000 km, we also found evidence of much shorter routes using the TEA/CR2 China Telecom optical cable system [3], yielding D_z of ∼11,000 km. Although for Eastern Europe the observed value of $\sigma \approx$ 9,796 km is close to the $\sigma \approx$ 10,000 km expected from transit through the USA, the observed σ of 7,410 for Western Europe is too low to consider a situation where all traffic follows this route. However, the combined effect of transit traffic through the USA and through continental Asia via [3] might explain the lower observed σ.

In the case of traffic between Oceania and Western Europe, geodesic distances vary widely, spanning from ∼17,000 km to ∼19,000 km even if we only consider the highest population centres such as Eastern Australia and New Zealand. We found evidence of traffic between these zones going through the USA and crossing the Pacific through southern California. This results on routing distances D_z ranging between ∼20,000 km and ∼22,000 km. This implies a σ that varies between ∼1,000 km and ∼5,000 km, which is in broad statistical agreement with the observed value of $\sigma \approx$ 2,702 km. For traffic between

Eastern Europe and Oceania, we found evidence of routing via Western Europe, which imposes an additional $\sim$3,000 km of routing distance and leads to distance excess σ of $\sim$5,702 km. This is consistent with the observed $\sigma \approx$ 5,685 km.

With regards to traffic between the Asia Pacific and South America (East) zones, we found evidence of routing paths using the various cable systems crossing the Caribbean [7], entering the USA via Miami and then being routed to the Asia Pacific region through the southern USA. This leads to a geodesic routing distance of $\sim$19,000 km and a $\sigma \approx$ 1000 km, consistent with the low observed σ of $\sim$973 km. Conversely, we also confirmed the high σ between the Asia Pacific and the Central Asia / Middle East zones. In this case, we found evidence of traffic being routed across the USA and continental Europe, leading to a routed distance D_z of $\sim$21,000 km and a $\sigma \approx$ 12,000 km, which is compatible with the observed value of $\sim$11,348 km. Similarly, we found evidence of traffic between Oceania and Central Asia following routes through the USA and Canada, imposing a D_z in the order of $\sim$23,000 km and an excess distance of $\sim$8,400 km in broad statistical agreement with the observed value of $\sim$7,623 km.

Although we found evidence of very large RTTs between the Indian subcontinent and Asia Pacific zones due to routing through the USA, this was relatively unusual. More commonly, we found evidence of routing through both Mumbai and Singapore, which is indicative of the use of undersea cables like the SeaMeWe-3 or FEA [7]. This route through the Bay of Bengal and the China seas has an approximate length of $\sim$9000 km, which would imply a σ of $\sim$4000 km and is hence compatible with the observed value of $\sim$4,100 km for σ.

Due to the size of the African continent, traffic between itself and Oceania follows very diverse routes. We found evidence of traffic from Australia to West Africa being routed through the USA, London, and then Africa via Portugal, suggesting that infrastructure such as the West African Cable System (WACS) was used; for destinations in East Africa, we found evidence of traffic being routed through the Mediterranean [7]. Routes also exhibited increased variability in the Asia Pacific Region, sometimes passing through intermediate points (e.g. Tokyo) before being routed to the USA. Regarding the values of σ observed, they ranged from $\sim$6,000 km to $\sim$24,000 km depending on the specific sources and destinations, with values around $\sim$15,000 being more common. These values are statistically consistent with the observed $\sigma \approx$ 18,973 km, but further subdivision of Africa may be required to achieve more precise σ estimates.

Regarding traffic between South America (East) and South America (West), we found evidence that, possibly due to the presence of the Andes and the Amazon, traffic between these regions frequently travels to the USA and back, imposing very large worst-case routing distances. This usually happens by following the South American coast northward and reaching the USA through the Caribbean (crossing through the Panama Canal if necessary). For the areas with greater population densities, $\sigma \approx$ 7500 km. This is consistent with its observed value of $\sim$7,187 km. This same routing through the USA is responsible for the increased routing delay between Oceania and South America (West). In this case, we found evidence for a routing distance $\sim$19,000 km and a $\sigma \approx$ 4,000, which is in agreement with the observed value of $\sim$3,608 km.

As was the case with Oceania, traffic between South America (East) and Africa follows very diverse routes. In general, we found that flows were routed through the USA and Europe.If we consider traffic from the most populated areas in South America (East) on the Atlantic, we found values of σ ranging from $\sim$3,000 km to $\sim$12,000 km depending on the specific sources and destinations. Conversely, for highly populated areas in the Pacific, we found values of σ ranging from $\sim$6,000 km to $\sim$14,000 km. Overall, although these values do not seem to far from the observed value of $\sim$11,208 km, their dispersion indicates that further subdivision of Africa may be of interest for future work.

VI. Related Work

Many contributions have focused their attention on the scalable estimation of RTT, particularly in the context of specific applications (e.g. anycast services [18], [30] and content delivery overlays [11], [34]). A representative example of these efforts is Vivaldi [15], that embeds RTT measurements between end hosts in a coordinate model of the Internet delay space. The distance in this coordinate space is then used as an estimate of the RTT. Another example of a practical RTT estimation system is iPlane [26], that uses measurements to create an “atlas” of the Internet clustered on the basis of BGP atoms [14] (minimal elements experiencing equivalent routing paths). Given two end points, iPlane estimates an abstract view of the route that traffic will take between them. A database is collated and SQL-like queries are used to obtain data for a given pair of end hosts. iPlane NANO [27] is a “summarised” version of iPlane that requires smaller daily updates.

Other works have focused on the development of more accurate RTT models. In [22] Kaune et. al propose a scalable model to predict RTT that includes realistic delay jitter subject to the geographical positions of the sender and the receiver. Another example is [35], in which Zhang et. al. propose an RTT model that better preserves continental clustering and RTT triangle inequality variations (TIVs). [25].

An improved modelling of the relationship between RTT and geographic Internet properties has been at the heart of many works in network-centred host geolocation. In [16] Dong et. al. propose a model of the relationship between RTT and geographic distances using segmented polynomial regression and semidefinite programming. This system builds on the multilateration approach presented in [19], which transforms RTT measurements into geographic distance constraints to infer the location of Internet hosts.

Finally, there have been several efforts to map the geography of Internet resources. The seminal work of Lakhina et. al. [23] mapped routers to their geographical locations using both geolocation registries and DNS-based host naming heuristics. The authors presented an analysis of interface density across regions, with particular emphasis to its relationship with population density. In addition, they study the relationships between geographic distance and link density, and between the size and geographic extent of ASes. A more recent example is [21], where Kasiviswanathan et. al. investigate routing circuitousness, and show that more than 50% traffic volume has distance ratio higher than 2, and about 20% traffic volume has distance ratio higher than 4. Since they restrict their analysis to the U.S.A., this manifests as traffic taking long, *bouncing* detours between the East and West coasts before reaching its destination. A conceptually similar work is [28], in which Mátray et. al. present an analysis of the geography of routing paths. The authors present a frequency analysis of link lengths, quantify path circuitousness and explore the symmetry of end-to-end Internet routes. Although their dataset is significantly different from ours (*traceroute* probes from PlanetLab [2] nodes), their conclusions regarding route circuitousness are in line with our results.

VII. Conclusion

In this paper we presented a model for the large-scale analysis of RTT and geolocations distance based on a novel RTT dataset comprising $\sim$19 million measurements between $\sim$54 thousand measurement points. This model, although very simple, can account for $\sim$94% of the variability in median RTT for a given geolocation distance.

We approached the modelling of RTT by treating our measurements as realisations of a multidimensional random variable, whose distribution $\Phi(\mathbf{x})$ we estimated by constructing a two-dimensional histogram. Our main contribution was a procedure for the geographic analysis of $\Phi(\mathbf{x})$ that allows the recovery of large-scale routing information. We achieved this by dividing $\Phi(\mathbf{x})$ into components X_z on the basis of *subcontinental zones*. By applying a novel median-based linear fitting algorithm to each component, we synthesised a *large-scale routing excess σ* that quantifies the extra distance beyond the geodesic that packets traverse when flowing between the relevant subcontinental zones. This procedure can then be used to build a linear model for Internet RTT that, albeit very simple, is accurate and can be used to aid in the modelling and design of distributed systems.

This research has received funding from the Seventh Framework Programme (FP7/2007-2013) of the European Union, through the ENVISION project (grant agreement 248565).

References

[1] The World Factbook. Country comparison: Internet users. https://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html.

[2] PlanetLab. http://www.planet-lab.org, 2007.

[3] China Telecom Europe/China Cable Systems. http://en.chinatelecom.com.hk/cmsDt/HK/jsp/map_1.jsp, 2012.

[4] GeoNames. http://www.geonames.org, 2012.

[5] MaxMind GeoLite City. http://www.maxmind.com/app/geolitecity, 2012.

[6] Neustar IP Geolocation. http://www.neustar.biz/solutions/ip-geolocation, 2012.

[7] Submarine Cable Map. http://www.submarinecablemap.com/, 2012.

[8] Team Cymru Research: IP to ASN Mapping. http://www.team-cymru.org/Services/ip-to-asn.html, 2012.

[9] Traceroute.org. http://traceroute.org, 2012.

[10] WIDE Project. http://www.wide.ad.jp, 2012.

[11] S. Agarwal and J. R. Lorch. Matchmaking for online games and other latency-sensitive P2P systems. In *Proc. of ACM SIGCOMM*, pages 315–326, 2009.

[12] C. A. f. I. D. Analysis. CAIDA Macroscopic IP Topology Project, 2003.

[13] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman Divergences. *J. Mach. Learn. Res.*, 6:1705–1749, Dec. 2005.

[14] A. Broido and kc claffy. Analysis of RouteViews BGP data: policy atoms. In *Network Resource Data Management Workshop*, Santa Barbara, CA, May 2001.

[15] R. Cox, F. Dabek, F. Kaashoek, J. Li, and R. Morris. Practical, distributed network coordinates. *Proc. of SIGCOMM Comput. Commun. Rev.*, 34(1):113–118, 2004.

[16] Z. Dong, R. D. W. Perera, R. Chandramouli, and K. P. Subbalakshmi. Network measurement based modeling and optimization for IP geolocation. *Comput. Netw.*, 56(1):85–98, Jan. 2012.

[17] R. A. Fisher. On the Interpretation of Ξ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, Jan. 1922.

[18] M. J. Freedman, K. Lakshminarayanan, and D. Mazières. OASIS: anycast for any service. In *Proc. of ACM NSDI*, volume 3, 2006.

[19] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-based geolocation of internet hosts. *IEEE/ACM Trans. Netw.*, 14(6):1219–1232, Dec. 2006.

[20] A. Guttman. R-trees: a dynamic index structure for spatial searching. In *Proc. of the ACM SIGMOD*, 1984.

[21] S. Kasiviswanathan, S. Eidenbenz, and G. Yan. Geography-based analysis of the Internet infrastructure. In *Proc. of IEEE INFOCOM*, pages 131–135, april 2011.

[22] S. Kaune, K. Pussep, C. Leng, A. Kovacevic, G. Tyson, and R. Steinmetz. Modelling the Internet Delay Space Based on Geographical Locations. In *Proc. of the Euromicro PDP*, pages 301–310, feb. 2009.

[23] A. Lakhina, J. Byers, M. Crovella, and I. Matta. On the geographic location of Internet resources. *Selected Areas in Communications, IEEE Journal on*, 21(6):934–948, aug. 2003.

[24] D. Leonard and D. Loguinov. Turbo King: Framework for Large-Scale Internet Delay Measurements. In *Proc. of IEEE INFOCOM*, pages 31–35, 2008.

[25] C. Lumezanu, R. Baden, N. Spring, and B. Bhattacharjee. Triangle inequality variations in the internet. In *Proc. of IMC '09*, pages 177–183, New York, NY, USA, 2009. ACM.

[26] H. V. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, and A. Venkataramani. iPlane: An Information Plane for Distributed Services. In *Proc. of ACM OSDI*, pages 367–380, 2006.

[27] H. V. Madhyastha, E. Katz-Bassett, T. Anderson, A. Krishnamurthy, and A. Venkataramani. iPlane Nano: Path Prediction for Peer-to-Peer Applications. In *Proc. of ACM NSDI*, pages 137–152, 2009.

[28] P. Matray, P. Haga, S. Laki, I. Csabai, and G. Vattay. On the network geography of the Internet. In *Proceedings of IEEE INFOCOM*, pages 126–130, april 2011.

[29] P. Mockapetris. RFC 1035: Domain Names: Implementation and Specification, 1987.

[30] E. Mykoniati, L. Latif, R. Landa, B. Yang, R. Clegg, D. Griffin, and M. Rio. Distributed overlay anycast table using space filling curves. In *Proc. of the IEEE INFOCOM Global Internet Symposium*, 2009.

[31] J. A. Nelder and R. Mead. A simplex Method for Function Minimization. *Computer Journal*, 7:308–313, 1965.

[32] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. IP geolocation databases: unreliable? *SIGCOMM Comput. Commun. Rev.*, 41(2):53–56, Apr. 2011.

[33] L. Subramanian, V. N. Padmanabhan, and R. H. Katz. Geographic Properties of Internet Routing. In *Proceedings of the USENIX Annual Technical Conference (ATEC)*, pages 243–259, Berkeley, CA, USA, 2002.

[34] M. Szymaniak, D. Presotto, G. Pierre, and M. van Steen. Practical large-scale latency estimation. *Comput. Netw.*, 52(7):1343–1364, May 2008.

[35] B. Zhang, T. S. E. Ng, A. Nandi, R. Riedi, P. Druschel, and G. Wang. Measurement based analysis, modeling, and synthesis of the internet delay space. In *Proceedings of ACM IMC*, pages 85–98, 2006.

[36] H. Zheng, E. K. Lua, M. Pias, and T. G. Griffin. Internet Routing Policies and Round-Trip-Times. In *Proc. of PAM*, pages 236–250, 2005.