The Inframetric Model for the Internet

Pierre Fraigniaud
CNRS
University Paris 7
France

Emmanuelle Lebhar
CNRS
University Paris 7
France

Laurent Viennot
INRIA
University Paris 7
France

***Abstract*—A large amount of algorithms has recently been designed for the Internet under the assumption that the distance defined by the round-trip delay (RTT) is a metric. Moreover, many of these algorithms (e.g., overlay network construction, routing scheme design, sparse spanner construction) rely on the assumption that the metric has bounded ball growth or bounded doubling dimension. This paper analyzes the validity of these assumptions and proposes a tractable model matching experimental observations.**

On the one hand, based on Skitter data collected by CAIDA and King matrices of Meridian and P2PSim projects, we verify that the ball growth of the Internet, as well as its doubling dimension, can actually be quite large. Nevertheless, we observed that the doubling dimension is much smaller when restricting the measures to balls of large enough radius. Moreover, by computing the number of balls of radius r required to cover balls of radius $R > r$, we observed that this number grows with R much slower than what is predicted by a large doubling dimension.

On the other hand, based on data collected on the PlanetLab platform by the All-Sites-Pings project, we confirm that the triangle inequality does not hold for a significant fraction of the nodes. Nevertheless, we demonstrate that RTT measures satisfy a weak version of the triangle inequality: there exists a small constant ρ such that for any triple u, v, w, we have $RTT(u, v) \leq \rho \cdot \max\{RTT(u, w), RTT(w, v)\}$. (Smaller bounds on ρ can even be obtained when the triple u, v, w is skewed). We call *inframetric* a distance function satisfying this latter inequality. Inframetrics subsume standard metrics and ultrametrics.

Based on inframetrics and on our observations concerning the doubling dimension, we propose an analytical model for Internet RTT latencies. This model is tuned by a small set of parameters concerning the violation of the triangle inequality and the geometrical dimension of the network. We demonstrate the tractability of our model by designing a simple and efficient compact routing scheme with low stretch. Precisely, the scheme has constant multiplicative stretch and logarithmic additive stretch.

I. Introduction

The quest for a better understanding of the Internet structure at the router level as well as at the AS level has yield a tremendous amount of work over the last decade, initiated by the pioneering contributions of Faloutsos et al. [1] identifying power laws in the Internet. A large amount of algorithms has also recently been designed for the Internet, including overlay network construction [2], routing scheme design [3], [4], sparse spanner construction [5], closest server selection [6]–[8], etc. The design of these algorithms assumes that the distance defined by the round-trip delay (RTT)[1] is a metric, and hence, in particular, that the triangle inequality

$$RTT(u, v) \leq RTT(u, w) + RTT(w, v)$$

is satisfied for any triple u, v, w. Moreover, the performance analysis of many of these algorithms relies on the assumption that RTT has bounded ball growth, i.e., for any $r > 0$, the size of any ball of radius r can be bounded by a constant times the size of the ball of radius $r/2$ centered at the same node (e.g., [7], [9]). Important contributions (e.g., [10], [11]) have relaxed this assumption to the bounded doubling dimension hypothesis, i.e., for any $r \geq 0$, any ball of radius r can be covered by a constant number of balls of radius $r/2$. Metrics of bounded doubling dimension have actually recently received a considerable attention because they provide a richer framework for the design and analysis of algorithms (cf., e.g., [10]–[15])

The bounded ball growth assumption can be well motivated intuitively [9] and is consistent with the transit-stub model [16]. Although it can be shown [17] that the RTT delays are poorly correlated to metrics such as physical distances, the formal verification of the bounded ball growth assumption has been statistically established in average by [1], [18]. For instance, [1] shows that the RTT distance over all pairs follows a power law, i.e., the number of pairs $P(h)$ at RTT distance h satisfies $P(h) \propto h^c$ for some constant c. As a consequence, $P(2h) \propto 2^c \cdot P(h)$. Nevertheless, the averaging over all pairs may hide the ball growth misbehavior for a large number of centers, and the assumption $P(2h)/P(h) \leq O(1)$ is not strong enough to enable algorithms designed under the bounded ball growth assumption to perform efficiently in a framework in which the bound only holds in average.

On the other hand, previous work tends to indicate that the basic metric assumption is questionable. In particular, [6], [18]–[20] already pointed out that the triangle inequality can be violated by the RTT latencies.

In this paper, we experimentally revisit the validity of the metric and geometrical assumptions made for the RTT latency distances. Significant contributions have been made by [6], [8], [21], [22] to better understand the RTT distance, but their

All authors are members the INRIA Projet-Team "GANG" between INRIA Paris Rocquencourt and LIAFA. Additional supports from the COST Action 295 "DYNAMO", the CRC "MARDI", and from the ANR projects "ALPAGE" and "ALADDIN".

[1]The RTT between two nodes u and v is the time taken to send a packet from u to v and to receive an acknowledgment back from v to u.

approach relies on extrapolating the missing data by using the triangle inequality, which is not necessarily satisfied by the RTT distance.

We propose a tractable analytical model for the RTT distances in the Internet, matching our experimental observations. We demonstrate the tractability of our model by showing how it can be used to design and analyze sophisticated algorithms. Our results are therefore complementary to the results in [18]. Indeed, [18] was (with [1]) among the first contributions considering experimentally the ball growth of the Internet, and modeling the violation of the triangle inequality. The objective of [18] was however quite different from ours. It developed a fine tuned and compact statistical tool box for the purpose of simulation and emulation of protocols. This tool box provides an artificial synthesis of a realistic delay space. In contrast we develop an analytical model for the purpose of design and analysis of algorithms. Consequently, we focus on worst case analysis of the experimental data, rather than on the global statistical distribution of the parameters.

A. On the Geometrical Dimension of the Internet

Based on Skitter data collected by CAIDA [23], we verify that the ball growth of the Internet is small for a vast majority of ball centers. However, this property does not hold for short radii or for specific placements of the ball centers (e.g., in New Zealand). The violation of the ball growth for small radii has some impact on the performances of algorithms designed by assuming that the bounded ball growth property holds at all scales, especially as our experiments demonstrate that the deviation to the bounded ball growth hypothesis at small scales can be significant. This implies that any analytical model must impose some threshold on the validity of the bounded ball growth asumption. It remains that, even for large enough balls, the bounded growth hypothesis is violated in some areas of the network. This latter fact leads us to relax the constraint, and to turn our attention toward the doubling dimension of the Internet.

Checking the doubling dimension of the Internet requires a complete distance matrix, and thus could not be performed using the Skitter measurements from CAIDA. Instead, we used King measurements coming from Meridian project [24] and P2PSim simulator [25]. Our experiments on these platforms demonstrate that one can hardly claim that the Internet has small doubling dimension in general. Indeed, as for the ball growth, the doubling dimension can be large for many balls for a large range of radii. Now, this observation is counterbalanced by the following fact. By computing the number of balls of radius r required to cover balls of radius $R > r$, we observed that this number grows with R much slower than what is predicted by a large doubling dimension.

B. On the Triangle Inequality in the Internet

Based on data collected on the PlanetLab platform by the All-Sites-Pings project [26], we confirm that the triangle inequality does not hold for a significant fraction of the nodes. Nevertheless, the same data also show that RTT measures satisfy a weak version of the triangle inequality. Namely, there exists a small constant ρ such that, for most of the triples u, v, w, we have

$$RTT(u, v) \leq \rho \cdot \max\{RTT(u, w), RTT(w, v)\},$$

to be contrasted with the triangle inequality. We call *inframetric* a distance function satisfying this latter relaxed inequality. Note that inframetrics subsume standard metrics ($1 \leq \rho \leq 2$) and ultrametrics ($\rho = 1$).

Better bounds on ρ can even be obtained when the triple u, v, w is *skewed*, i.e., when the side v, w of the triangle is much smaller than the side u, v (cf. Fig. 1). Our experiments demonstrate that the triangle inequality is violated more frequently for skewed triangles, but that the violation is less severe than for arbitrary triangles.

C. On the Inframetric Nature of the Internet

The observations summarized in the two previous subsections enable us to design an accurate analytical model for the RTT latencies in the Internet. Our model is tuned by two sets of parameters. The first set concerns the deviation to the triangle inequality. It consists of the parameter ρ defined above, and a pair of parameters (ρ_s, δ) allowing us to use tradeoffs between the deviation from the triangle inequality, and the skewness of the triangle involved in the inequality. Such tradeoffs will be shown helpful for the analysis of algorithms computing overlay data structures aiming at preserving distances approximatively.

The second set of parameters of our model concerns the geometrical dimension of the network, precisely its doubling dimension. It consists of three parameters α, β and τ. The parameters α and β enable to establish bounds on the doubling dimension at various scales. These bounds are however only valid for balls of radius at least τ. Note that we actually generalize the standard notion of doubling dimension by comparing balls of radius r with balls of radius r/ρ. While this modification in the definition of doubling dimension does not modify the nature of the geometrical dimension of the network, it simplifies the analysis a lot.

D. Low Stretch Compact Routing Schemes in Inframetric Spaces

Designing compact routing schemes is a case study that illustrates the tractability of our model. We design a simple and efficient compact routing scheme with low stretch, in the same spirit as the TZ algorithm [3], [4]. Precisely, we show how Slivkins routing algorithm [11] can be revisited to fit with the inframetric model, and how its complexity can be analyzed in this framework. We have chosen to consider Slivkins algorithm because it tackles the core of the problem. This algorithm is itself an extension of an algorithm by Chan et al. [27]. It has received considerable attention, and has been successively refined in [12], [14], and [15]. In these algorithms, routing in a metric often serves as a basis for constructing compact routing schemes in general networks. In the inframetric model, we present a compact routing scheme using routing tables of

polylogarithmic size at each node, with constant multiplicative stretch, and logarithmic additive stretch. I.e., the length of the route between any source s and target t computed by our scheme is at most $a \cdot RTT(s,t)+b$ where a is a small constant, and b grows logarithmically in the size of the network.

II. The Inframetric Model

This section describes the inframetric model for the Internet. This analytical model will be validated later in the text, via extensive experiments performed using data collected by CAIDA, PlanetLab, Meridian and P2PSim. As said before, the inframetric model is tuned by a small set of parameters concerning the violation of the triangle inequality and the geometrical dimension of the network (ball growth or doubling dimension). These parameters appear explicitly in Definitions 1, 2, and 3 summarizing the essence of our model.

A. Inframetrics

In this paper, we call *distance* any function aiming at capturing a notion of proximity between elements of a finite set V. Clearly the RTT latency falls into this category. Recall that a nonnegative (distance) function $d : V \times V \to \mathbb{R}$ is a *metric* if it satisfies:

- $d(u,v) = 0$ if and only if $u = v$;
- $d(u,v) = d(v,u)$ (symmetry property);
- $d(u,v) \leq d(u,w) + d(w,v)$ (triangle inequality).

As we mentioned in the introduction, and as our experiments presented later in the paper will demonstrate, the RTT latency merely satisfies the two first properties, but significantly violates the triangle inequality. We thus introduce a relaxed version of this latter property, yielding the notion of *inframetric*.

Definition 1: A distance $d : V \times V \to \mathbb{R}$ is a *ρ-inframetric* for $\rho \geq 1$ if it satisfies the two first axioms of metrics and the following relaxed triangle inequality: for any triple u, v, w in V,

$$d(u,v) \leq \rho \ \max\{d(u,w), d(w,v)\}. \qquad (1)$$

Note that, if d is a metric, then $\rho \leq 2$ by the triangle inequality, and that ultrametrics are defined as the metrics satisfying Inequality 1 with $\rho = 1$. Our experiments demonstrate that the RTT latency is a ρ-inframetric for some small $\rho > 2$. In fact, by restricting the measure of ρ to some specific types of triangles u, v, w, called skewed (cf. Fig. 1), one can obtain significantly smaller values for ρ.

Definition 2: Let $0 < \delta \leq 1$. A triangle u, v, w is *δ-skewed* if $d(w,v) \leq \delta \ d(u,v)$. An inframetric d is *(ρ_s, δ)-skewed* for $\rho_s > 0$ if for any δ-skewed triangle u, v, w, we have $d(u,w) \leq \rho_s \ d(u,v)$.

The notion of (ρ_s, δ)-skewness for a ρ-inframetric d is particularly interesting if ρ_s is significantly smaller than ρ. In particular, a detour via w when routing from u to v in a δ-skewed triangle u, v, w results in a stretch factor $\rho_s + \delta$ instead of $\rho + \delta$. The reason why it is interesting to restrict the measure of ρ to skewed triangles will be clear once we present our application to the design of compact routing schemes.

Note that for any $\varepsilon \in (0, 1]$, classical metrics are $(1+\varepsilon, \varepsilon)$-skewed, and that ultrametrics are $(1, \varepsilon)$-skewed.

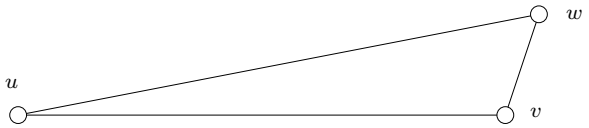

Fig. 1. A skewed triangle u, v, w

B. Ball Growth and Doubling Dimension

In this subsection, we generalize the notions of ball growth and doubling dimension to inframetrics. Given a distance function d on a set V, $B_u(r)$ denotes the ball of radius $r \geq 0$ centered at $u \in V$, i.e., $B_u(r) = \{v \in V \mid d(u,v) \leq r\}$. The standard definitions of ball growth and doubling dimension compare balls of radius $2r$ with balls of radius r. These notions can naturally be extended to ρ-inframetrics by comparing balls of radius ρr with balls of radius r. A ρ-inframetric d has *growth* $\gamma \geq 1$ if, for any $r \geq 0$ and $u \in V$,

$$|B_u(\rho r)| \leq \gamma \, |B_u(r)|.$$

Metrics of bounded growth are special cases of metrics of bounded doubling dimension, defined as follows. A ρ-inframetric is γ-doubling if, for any $r \geq 0$ and $u \in V$,

$$B_u(\rho r) \subseteq \cup_{i \in I} B_{v_i}(r)$$

for some $v_i \in V$, $i \in I$, $|I| \leq \gamma$. As for usual metrics, inframetrics of bounded growth are special cases of inframetrics with bounded doubling dimension. Specifically, we have:

Lemma 1: Let d be an inframetric. If d is of growth γ then it is γ'-doubling with $\gamma' \leq \gamma^4$.

Proof: Let d be a ρ-inframetric of growth γ. Let $r \geq 0$. From the inframetric property, we get $B_u(\rho r) \subseteq B_v(\rho^2 r)$ for any $v \in B_u(\rho r)$. On the other hand, since d is of growth γ, $|B_v(r/\rho)| \geq |B_v(\rho^2 r)|/\gamma^3$. Hence, $|B_v(r/\rho)| \geq |B_u(\rho r)|/\gamma^3$ for any $v \in B_u(\rho r)$. Thus

$$|B_v(r/\rho)| \geq |B_u(\rho^2 r)|/\gamma^4 \text{ for any } v \in B_u(\rho r). \qquad (2)$$

Let us now consider the following greedy process for covering $B_u(\rho r)$ with balls of radius r:

$k \leftarrow 0$;
while $B_u(\rho r) \setminus \cup_{1 \leq i \leq k} B_{v_i}(r) \neq \emptyset$ **do**
  select an arbitrary $v_{k+1} \in B_u(\rho r) \setminus \cup_{1 \leq i \leq k} B_{v_i}(r)$;
  $k \leftarrow k + 1$;

The process stops when $B_u(\rho r) \subseteq \cup_{1 \leq i \leq k} B_{v_i}(r)$. Let $w \in B_{v_i}(r/\rho)$, and $j \neq i$. If $w \in B_{v_j}(r/\rho)$, then $d(v_i, v_j) \leq \rho \ \max\{d(v_i,w), d(v_j,w)\} \leq r$, a contradiction. Hence $d(v_j, w) > r/\rho$, and therefore the balls $B_{v_i}(r/\rho)$ are pairwise disjoint, for $i = 1, \ldots, k$. Thus, by Eq. 2, we get $|\cup_{1 \leq i \leq k} B_{v_i}(r/\rho)| \geq k \, |B_u(\rho^2 r)|/\gamma^4$. Since by the bounded growth property $B_{v_i}(r/\rho) \subseteq B_u(\rho^2 r)$ for any $i \in \{1, \ldots, k\}$, we get that $k \leq \gamma^4$, which completes the proof. ∎

In view of our experiments regarding the Internet, the aforementioned definitions of ball growth and doubling dimension have two drawbacks. First, we observe a limited growth of the ball sizes only for radii above a certain threshold. Second, the

definitions $|B_u(\rho r)| \leq \gamma\ |B_u(r)|$ or $B_u(\rho r) \subseteq \cup_{i\in I} B_{v_i}(r)$, $|I| \leq \gamma$, actually still yields relatively large γ's even for reasonably large balls. These two problems are handled by the following definition that reflects more accurately our observations on the Internet.

Definition 3: A ρ-inframetric is (α, β)-doubling with threshold τ if, for any $u \in V$, any $r \geq \tau$, and any $R \geq \rho r$,

$$B_u(R) \subseteq \cup_{i\in I} B_{v_i}(r)$$

for some $v_i \in V$, $i \in I$, $|I| \leq \beta\alpha^{\log_\rho R/r}$.

The role of τ in the above definition is to limit the deviation due to too small balls, or balls of too small radius. The role of the pair (α, β) is more subtle. Note that $(\gamma, 1)$-doubling inframetrics are simply γ-doubling inframetrics. The two parameters α and β give more flexibility to the model. Indeed, if $R = \rho^i r$ for $i \geq 1$, then the usual definition of γ-doubling dimension implies that any ball of radius $\rho^i r$ can be covered by at most $f(i)$ balls of radius r with $f(i) \leq \gamma^i$. Our observations demonstrate however that $f(i) \propto \beta\alpha^i$ with $\beta\alpha = \gamma$ but $\alpha \ll \gamma$. In other words, even if some balls of radius ρr require a large number $\beta\alpha$ of balls of radius r to be covered in the Internet, the number of balls of radius r required to cover balls of radius $\rho^i r$ grows rather slowly with i, like $\beta\alpha^i$, and not like $(\beta\alpha)^i$.

III. Internet Latencies as an Inframetric

In this section, we present our experimental study of Internet latencies in terms of triangle inequality violations, and analyze how they fit with the inframetric model.

Internet latencies are basically measured through round trip times (RTT for short). The RTT between two nodes u and v is the time taken to send a packet from u to v and to receive an acknowledgment back from v to u. Such a measurement is typically made using the `ping` command. Note that under stable network conditions, this measure is inherently symmetric. Large dense matrices of estimations can also be obtained through the King method [28] using recursive DNS queries. However, we prefer PlanetLab measurements (from All-Sites-Pings project [26]) for testing triangle inequalities because it provides better accuracy while still containing many triangle measurements. The All-Sites-Pings project consists in taking regular snapshots of RTT measurements between various pairs of sites of PlanetLab. In our study, we use snapshots taken during one hour (where approximately 15 pings are made between each pair of sites in PlanetLab). The data consists in the minimal, average and maximal values of the measured latencies. A snapshot typically contains between 200 and 300 nodes.

A. Methodology

RTT measurements may vary due to node overload, especially for PlanetLab nodes which serve as an open testbed for many research projects. Since we are interested in modeling network distances under steady network load, we filter out measures with high overload (on a node or in the network). For this purpose, we use the following heuristic. We measure the difference between minimal, maximal and average values on the measures given by one snapshot between a fixed pair of nodes. We then filter out the pairs for which the three values differ from each other by more than 10%. For the same purpose, we ignore nodes answering to less than 10 ping requests, nodes having more than 30% of their measures filtered out by the first test, and nodes having a difference between their average RTT to other nodes higher than 1 second. Indeed, we consider that such behaviors correspond to overloaded nodes. After such filtering, the data is almost symmetrical, i.e., $RTT(u, v)/RTT(v, u)$ is always very close to one.

Finally, a significant fraction of the nodes does not provide any data in a snapshot (but they are probed by other nodes) and a significant fraction of the entries is missing. A snapshot is typically a 130×220 matrix with 13 % of the entries missing. After our filtering process, we end up with a matrix with size 100×150 approximately and the same proportion of valid entries. We also make use of the matrix obtained by averaging our 161 filtered snapshots made in 2007 in one single matrix in order to get a picture more homogeneous in terms of year period, we call that matrix *average matrix*.

We have studied 288 snapshots taken on each hour of twelve days spread from December 2005 to June 2007 (one in 2006, two in 2005 and nine in 2007). The notation 2006-12-10–03 will denote the snapshot made on December 2006, the 10th, from 3 am to 4am (GMT +8:00).

B. Triangle inequality

In order to test the triangle inequality, we compute the ratios $RTT(u, v)/(RTT(u, w) + RTT(w, v))$ for all triangles for which the measures are available (each snapshot contains several hundred thousands triangles). We study separately the triangles depending on their skewness: the parameter $\delta > 0$ for which they satisfy $RTT(w, v) < \delta\ RTT(u, w)$ (we authorize $\delta = \infty$ in this section for the sake of simplicity). The ratios depending on the skewness are illustrated in Figure 2 wich plots the cumulative distribution of these ratios on all snapshots. The figure should be read as follows: a point x, y on a curve indicates that a fraction y of the triangles considered in the curve has a ratio at most x. We can see that approximately 5 % of the triangles do not satisfy the triangle inequality. Interestingly, skewed triangles ($\delta = .5$ and $.1$) violations are more frequent (6 % and 13 % respectively).

In order to test the inframetric inequality, we compute the ratios $RTT(u, v)/\max\{RTT(u, w), RTT(w, v)\}$ on the same set of triangles. The results are illustrated on Figure 3 which plots the cumulative distribution of these ratios for various skew parameters δ. The curve $\delta = \infty$ (i.e., with all the triangles) indicates that Internet latencies are close to a 2-inframetric. Figure 4 provides the same distribution plot (on the average matrix) visible for ratios between 1 and 7. We see that the inframetric triangle inequality with $\rho = 7$ is almost never violated. Interestingly, we get similar results on the total matrix as on the average matrix.

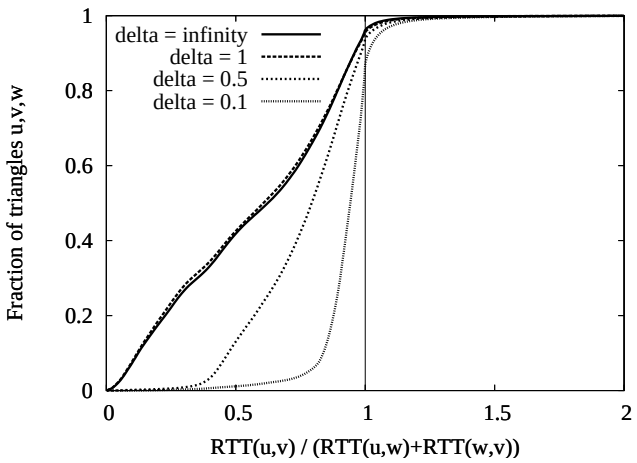

Fig. 2. Cumulative distribution over all snapshots of $RTT(u,v)/(RTT(u,w) + RTT(w,v))$ for all triangles such that $RTT(w,v) < \delta\ RTT(u,w)$ for $\delta = .1, .3, .5, 1, \infty$.

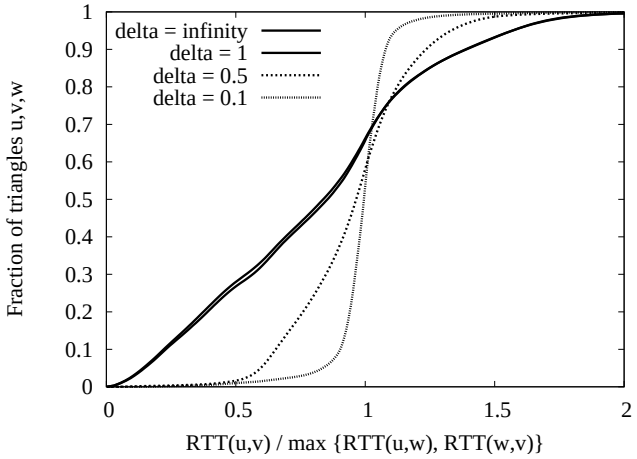

Fig. 3. Cumulative distribution over all snapshots of the ratio $RTT(u,v)/\max\{RTT(u,w), RTT(w,v)\}$ for all triangles such that $RTT(w,v) < \delta\ RTT(u,w)$ for $\delta = .1, , .5, 1$ and ∞.

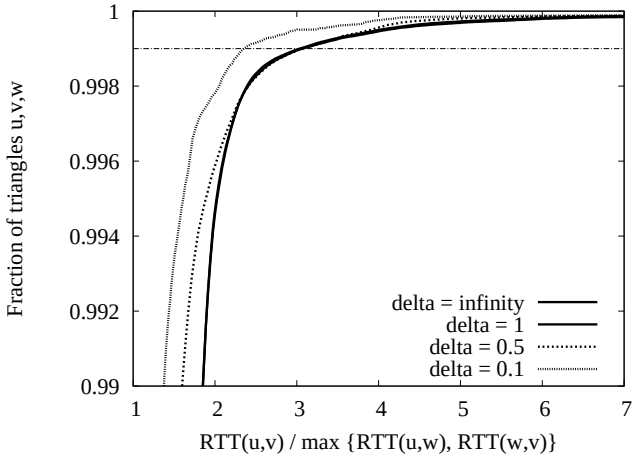

Fig. 4. Average matrix (over 2007).

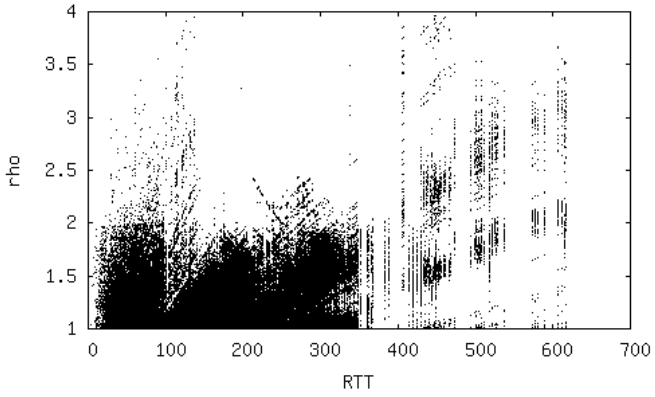

Fig. 5. A point (x, y) corresponds to a triangle u, v, w in the 2006-12-10–06 snapshot with $x = RTT(u,v)$ (in milliseconds) and $y = RTT(u,v)/\max\{RTT(u,w), RTT(w,v)\}$.

To give more insight on the nature of triangle inequality violations, we plot in Figure 5 a point $x = RTT(u,v)$, $y = RTT(u,v)/\max\{RTT(u,w), RTT(w,v)\}$ for all triangle u, v, w of the 2006-12-10–06 snapshot. Several alignments of points can be observed. Each vertical line corresponds to some pair of nodes u, v for which many nodes w with $RTT(u,w) < RTT(u,v)$ and $RTT(w,v) < RTT(u,v)$ give a point (x, y) where $x = RTT(u,v)$ and $y > 1$. Oblique lines are due to pairs u, w such that are several nodes v with $RTT(w,v) \leq RTT(u,w)$ for a wide range of values for $RTT(u,v)$. Each v yields a point $(RTT(u,v),\ RTT(u,v)/RTT(u,w))$. All these lines pass through the origin. Note how lines get steeper for small $RTT(u,w)$. We obtain similar plots for the average matrix. We have chosen to show the snapshot plot because the lines due to special pairs of nodes were more visible on that plot.

Figure 6 is made of points $x = \delta$, $y = RTT(u,v)/\max\{RTT(u,w), RTT(w,v)\}$ for all triangles u, v, w of the average matrix where $\delta = \min\{RTT(u,w), RTT(w,v)\}/\max\{RTT(u,w), RTT(w,v)\}$ is the skewness of the triangle. For most of the δ-skewed triangles, the skewed triangle inequality is roughly satisfied with ρ_s ranging from 1.2 to 1.8 for $0 \leq \delta \leq 0.7$. The inequality with $\rho_s < 1.2$ is violated quite frequently even with very small δ. We obtain similar plots when using snapshot matrices with a slightly better bounds on ρ_s.

IV. Ball Growth and Doubling Dimension of Internet Latencies

In this section we study the ball growth on Skitter data which provides the measurements with the largest set of destinations but only from few sources. We then study the doubling dimension on large complete distance matrices obtained with the King method.

A. Caida Skitter measurements

The Skitter project of Caida consists in few monitors regularly probing a fixed very large set of IP addresses. Each probe consists in a `traceroute` query, i.e., a sequence of

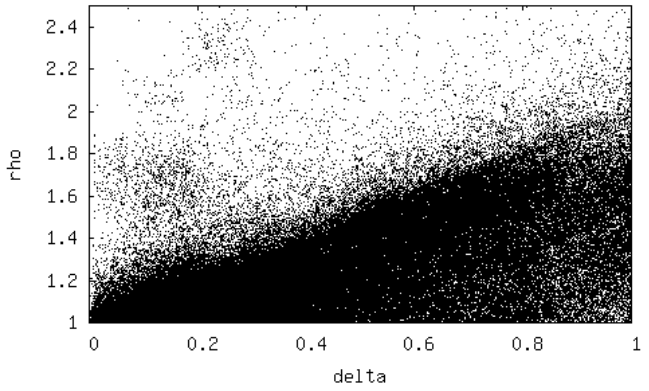

Fig. 6. A point (x, y) corresponds to a x-skewed triangle u, v, w in the average matrix with $y = RTT(u, v)/\max\{RTT(u, w), RTT(w, v)\}$.

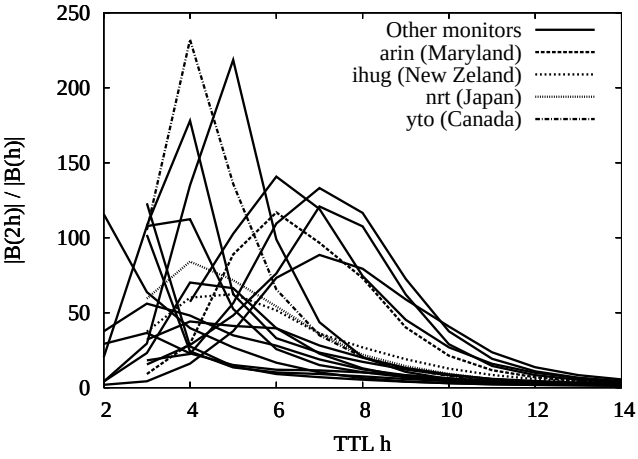

Fig. 7. Growth ratio $B_x^{TTL}(2h)/B_x^{TTL}(h)$ as a function of h (number of hops) for various monitors x. Plot begins when $B_x(h) \geq 20$.

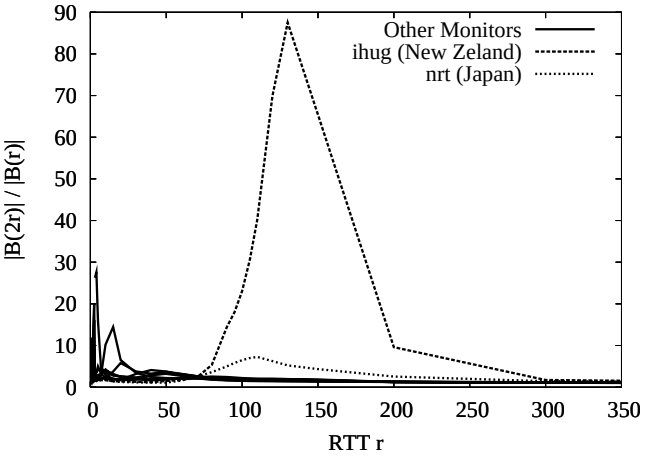

Fig. 8. Growth ratio $B_x^{RTT}(2r)/B_x^{RTT}(r)$ as a function of r (in milliseconds) for various monitors x. Plot begins when $B_x(r) \geq 20$.

pings with increasing TTL. (The TTL is a limit on the number of hops authorized for a ping packet). We extract from this data a RTT measurement from each monitor to each IP address probed as well as a hop distance measurement (obtained as the minimal cutoff TTL necessary to reach the destination). From this data we can indeed get a very precise idea of the ball growth of monitors. We have used the data of May 2004 because most of the probes were successful at that period (between $450\,000$ and $500\,000$ successful probes per monitor in the IPv4 list) whereas more recent traces have 30-40 % less successful probes. Due to space limitations, we only present May 2004 curves excluding intermediate routers. We get similar curves for other periods or when including routers.

B. Bounded growth

Figure 7 plots the growth of balls defined by hop counts as a function of the TTL hop count. There is clearly no low bound on this growth. This is not surprising since we count the number of leaves in the routing tree from a monitor at various depths. Figure 8 plots the growth of balls defined by RTTs as a function of the RTT radius. The RTT ball growth is generally smaller than 8 except for small radii and for two monitors. Interestingly, the growth of balls defined by RTT distance is generally much lower than the growth of balls defined by hop counts.

Now consider the highest growths observed. The highest peak is due to `ihug` monitor which is indeed an isolated monitor located in Auckland, New-Zealand. It has a peak of $300\,000/3500 \approx 85$ at 150 milliseconds (i.e., the size of the 300 milliseconds ball is $300\,000$ compared to 3500 for the 150 milliseconds ball). Few destinations are indeed reachable from New-Zealand within 150 milliseconds (only those in Australia and close Asia) whereas a large portion of destinations in America and Asia are reached within 300 milliseconds. Less strikingly, a similar phenomenon occurs at 110 milliseconds for `nrt` monitor which is located at Tokyo in Japan.

Figure 9 zooms in smaller RTTs. The highest peak is due to `arin` monitor which is located at Bethesda, Maryland (near Washington). It presents a peak of $9900/350$ at 4 millisecond. The peak is probably due to high density of the network (or of the destination set) in that region. In the same region, `iad` in Washington DC and `g-root` in Vienna, Virginia show rather high ratios for small RTTs. Indeed, almost all monitors appear to have two peaks, one for a very short radius and a second for a larger radius. Most monitors have smooth peaks. Notably, `yto` monitor, located in Ottawa, Canada, has a first peak of $720/35$ at 2.5 milliseconds and a second peak of $46\,500/3\,200$ at 15 milliseconds. The second peak could be explained by a threshold r where high speed international connectivity occurs. Many destinations are thus found within RTT less than $2r$ when significantly fewer destinations are found with RTT less than r. Indeed, second peaks at larger radii can be encountered when crossing broad physical deserts like oceans as illustrated by the striking peak of `ihug` located in Auckland.

Concerning short radii, the highest peaks are observed by considering the smallest possible radii. For instance, the radius r where the nearest destination from a monitor u is reached gives rise to a peak of $|B_u(2r)|/2$ which can be large. Such

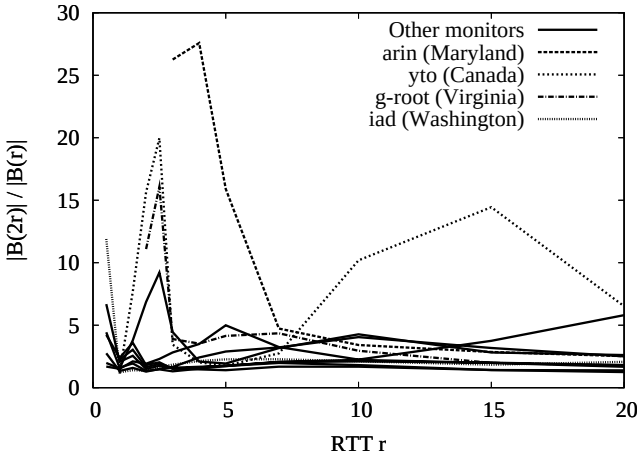

Fig. 9. Growth ratio $B_x^{RTT}(2r)/B_x^{RTT}(r)$ as a function of r (in milliseconds) for all monitors x. Plot begins when $B_x(r) \geq 20$.

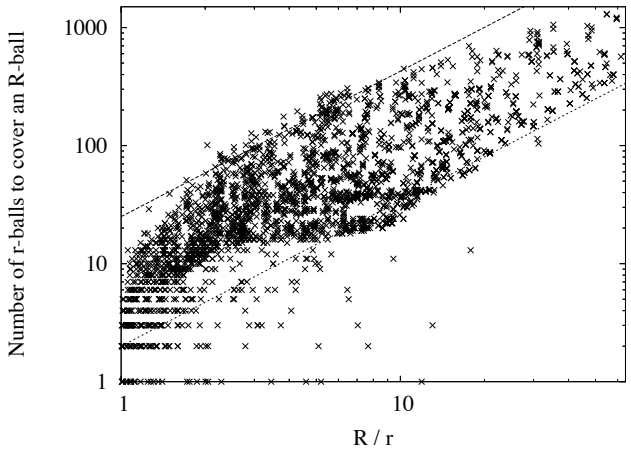

Fig. 10. Number of balls of radius r sufficient to cover a ball of radius R as a function of R/r in a complete 2303×2303 Meridian sub-matrix.

peak highly depends on the choice of the IPv4 list used for probing. To avoid such artifact, we only consider balls containing at least 20 nodes. Considering the hierarchical architecture of Internet it would not be surprising to find embedded trees at that scale. Complete regular trees are a notable example of structure with high ball growth and high doubling dimension.

To summarize, the ball growth appears to be bounded by 15 for radii greater than 7 milliseconds for all monitors but one and it is mostly lower than 5. We observe similar results for 2007 data. However the 2007 peaks are lower due to the fact that the destination set is less dense.

C. Doubling property

It is known that a bounded growth space is a particular case of low doubling metric. Randomly sampling nodes in a bounded growth space results in a bounded growth space [7]. However, sampling a bounded growth space results in general in a doubling metric. Including an isolated island of nodes (such as New-Zealand) can typically induce a high ball growth for these nodes. Considering the growths observed on Skitter data, doubling metrics seem a good candidate for modeling the geometry of latencies space. However, we must keep in mind that short radii may be problematic.

To test the doubling property on Internet latencies, we need a large set of nodes with all to all measurements (Skitter data only contains a small complete matrix between the monitors). To test larger scales, we have used two matrices obtained with the King method [28]. We have used the 1740×1740 matrix of the P2PSim simulator [25] and the 2500×2500 matrix of the Meridian project [24]. The King method consists in estimating the latency between two domain nameservers (DNS) u and v by making a recursive request for v through u. The delay for a direct request to u is then subtracted.

Given a complete matrix of distances, we test the (α, β)-doubling property as follows. We repeatedly select two random radii r, R with $r < R$ and run a heuristically optimized greedy algorithm to cover a randomly chosen ball B of radius R with balls of radius r. The number $U_{r,R}(B)$ of balls of radius r necessary to cover B is an estimation of $\beta\alpha^{\log_\rho R/r}$.

Figure 10 illustrates $U_{r,R}$ as a function of R/r on the Meridian matrix in log-log scale. More precisely, for each ball B of radius R and radius $r < R$ taken as input by our heuristic, we plot a point (x, y) where $x = R/r$ and $y = U_{r,R}(B)$ is the number of balls of radius r found to cover B. Most of the points lie between $2 \cdot (2.35)^{\log_2 R/r}$ and $25 \cdot (2.35)^{\log_2 R/r}$. As the matrix is not complete, we have extracted a complete sub-matrix of size 2303×2303. We also tried different ranges of radii, the highest values of $U_{r,R}(B)$ have been observed for all radii ranges, from $r = 1$ milli-seconds to $r = 200$ milliseconds. High R/r ratios can only be observed with small r. Observation on short radii cannot be conclusive since King matrices are made between nameservers and may miss high local density situations as those producing high ball growth on Skitter data for short distances.

We have obtained similar bounds for the P2PSim matrix which is a complete matrix of size 1740×1740. However, for higher R/r ratios, we observe lower $U_{r,R}$ values on that matrix. This is probably due to the fact that some nodes underestimate their distance to many others because of overload and the subtraction inherent to the method. This artifact was already pointed out by the authors of the P2PSim matrix [29]. It appears that such nodes have been better filtered out in the Meridian matrix.

Based on the observation of the largest complete latency matrices available, we can reasonably argue that Internet latencies satisfy the (α, β)-doubling property for a small value of α. The data sets investigated suggest $2 \leq \alpha \leq 3$ and $\beta \leq 30$. Note that points in a D dimensional grid with L_1 norm form a $(1, 2^D)$-doubling metric. We should thus compare α to 2 for a one-dimensional space and 4 for a bi-dimensional space. As Internet is embedded on a sphere, it is not surprising to find an α value between 2 and 4. The reason why β is significantly larger than α may come from the Internet nature

or from the inaccuracy of the data sets. Even moderately high values of β requires our fine grained definition of the doubling property in order to obtain low bounds in algorithm performances as illustrated in the next section. In particular, the classical definition of doubling dimension would result in an artificially high bound of $100^{\log_2 R/r}$ rather than the $25 \cdot (2.35)^{\log_2 R/r}$ bound observed on real data.

V. Compact routing in doubling inframetrics

In this section, we consider an n-node ρ-inframetric space (V, d) which is (α, β)-doubling with threshold τ. We also assume that the inframetric is (ρ_s, δ)-skewed with $\delta \leq 1$ (this assumption is relaxed in the last corollary).

A. Doubling inframetrics geometrical properties

Doubling metrics properties in computer science algorithmic problems are often analyzed trough a central decomposition tool: the *r-nets* [10]. This tool can be extended to the inframetrics setting: we say that a subset S of V is an r-net if for any $u, v \in S$, $d(u, v) > r$, and, for any $w \in V$, there exists $u \in S$ such that $d(u, w) \leq r$. An r-net always exists for any $r > 0$ and can be constructed greedily in a similar way as the greedy algorithm in Lemma 1 proof.

Note that if S is an r-net in V, then for any $u \in V$, $B_u(r/\rho)$ contains at most one node of S. Indeed, assume that $v, w \in S \cap B_u(r/\rho)$, then $d(v, w) \leq \rho \cdot r/\rho$, contradicting the r-net definition. Moreover, for any $i > 0$ and $t \geq \rho^i \tau$, $B_u(t)$ can be covered by less than $\beta \cdot \alpha^i$ ball of radius t/ρ^i. We get that:

Lemma 2: For any $r > \rho\tau$, any ball of radius $t \geq \rho r$ contains at most $\beta\alpha^{1+\log_\rho(t/r)}$ nodes of an r-net in V.

B. Low stretch compact routing

Let Δ be the aspect ratio of d, i.e. the maximum distance in (V, d) over the minimum distance in (V, d) . We say that a routing scheme has stretch (a, b) if the path length of the routing is at most $aD + b$ between any two nodes at distance D. Each node has an $O(\log n)$ bits ID.

Theorem 1: For any positive $\varepsilon < \delta/\rho$, there exists a compact routing scheme in (V, d) with stretch $(\rho_s/(1 - \rho\varepsilon), \tau\rho^2 \log n)$, table size $\beta\alpha^{3+\log_\rho(\frac{1}{\varepsilon})} \log_\rho(\Delta/\tau) \log n$ bits per node, and node label size $O(\log_\rho(\Delta/\tau) \log n)$ bits.

Proof: We show how Slivkins routing scheme [11] can be extended to our setting. For the sake of simplicity, assume that the smallest non zero distance is 1.

For every $i \in \{1, \ldots, \lceil \log_\rho(\Delta/\tau) \rceil\}$, let S_i be a $\rho^i\tau$-net. We say that the nodes of S_i are of *level* i. We define an ancestor relationship in the hierarchy of nets as follows. For any node u and $i \geq 1$, we denote by $a_i(u)$ the ancestor of level i of u. Let $a_1(u)$ be the closest node to u in S_1. For any $i > 0$, $a_i(u)$ is the node in S_i which is the closest to $a_{i-1}(u)$. Applying the inframetric inequality repeatedly over the levels, one can see that for any $u \in V$ and $i \geq 1$, $a_i(u)$ is at distance at most $\rho^{i+1}\tau$ from u.

Let $\varepsilon < \delta/\rho$ be any positive real. For every $u \in V$ and each $i \in \{1, \ldots, \lceil \log_\rho(\Delta/\tau) \rceil\}$, let $R_i(u) = S_i \cap B_u(\rho^{i+2}\tau/\varepsilon)$. From Lemma 2, we get that $B_u(\rho^{i+2}\tau/\varepsilon)$ contains at most $\beta\,\alpha^{3+\log_\rho(1/\varepsilon)}$ nodes of S_i, since $\rho^{i+2}\tau/\varepsilon \geq \rho^{i+3}\tau$.

For every $u \in V$, we define $\textsc{Table}(u)$ and $\textsc{Label}(u)$ respectively as the routing table and the label of u in our routing scheme.

$\textsc{Table}(u)$ stores the IDs and IP addresses of all nodes in $R_i(u)$ for each $i \in \{1, \ldots, \lceil \log_\rho(\Delta/\tau) \rceil\}$, which is at most $\beta\,\alpha^{3+\log_\rho(1/\varepsilon)} \lceil \log_\rho(\Delta/\tau) \rceil$ IDs in total from the previous discussion. The table size of u is therefore at most $\beta\,\alpha^{3+\log_\rho(\frac{1}{\varepsilon})} \log(\Delta/\tau) \log n$ bits.

$\textsc{Label}(u)$ is the set of IDs of $a_1(u), \ldots, a_{\lceil \log_\rho(\Delta/\tau) \rceil}(u)$, yielding a label size at most $O(\log_\rho(\Delta/\tau) \log n)$ bits.

Suppose we route from a source s to a target t. Let $s = u_0$. The first phase of the routing scheme proceeds as follows: in each step, the current node u_k, $k \geq 0$, forwards the packet with header $\textsc{Label}(t)$ to the node u_{k+1} in $\textsc{Label}(t)$ having lowest level, and whose ID appears in $\textsc{Table}(u_k)$. The first phase of the routing scheme ends as soon as the current node is at distance at most $\rho\tau$ to $a_1(t)$.

Let us analyze this first phase of the routing. Let i be such that $\rho^{i-1}\tau \leq \varepsilon \cdot d(s, t) < \rho^i\tau$. We have $d(a_{i-1}(t), t) \leq \rho^i\tau \leq \rho\varepsilon \cdot d(s, t)$. Thus, $d(s, a_{i-1}(t)) \leq \rho^{i+1}\tau/\varepsilon$ and $a_{i-1}(t) \in R_{i-1}(s)$. The next node u_1 on the routing path is thus either $a_{i-1}(t)$ or a node of level $< i - 1$. It thus satisfies $d(u_1, t) \leq \rho^i\tau \leq \rho\varepsilon \cdot d(s, t)$. Since $\rho\varepsilon < \delta$, the triangle $s, t, u_1(t)$ is δ-skewed. Therefore, $d(s, u_1) \leq \rho_s \cdot d(s, t)$. Thus, the current distance to t has been multiplied by a factor of at most $\rho\varepsilon < \delta \leq 1$ with a hop of length at most $\rho_s \cdot d(s, t)$. Moreover $\varepsilon \cdot d(u_1, t) \leq \varepsilon\rho^i\tau \leq \rho^{i-1}\tau$. Repeating the same analysis until the routing path reaches $a_1(t)$, we get a total path length at most $\sum_{j \geq 0} \rho_s(\rho\varepsilon)^j \cdot d(s, t) \leq \frac{\rho_s}{1-\rho\varepsilon}\, d(s, t)$, hence yielding a multiplicative stretch $\rho_s/(1 - \rho\varepsilon)$.

The final phase of the routing scheme consists in reaching t from $a_1(t)$. It proceeds the same way under both assumptions (skewness or not). Each node knows the ID of its ancestor of lowest level, which is at distance less than $\rho\tau$. In order to route among the nodes with same lowest ancestor $a_1(t)$, we can use a constant degree DHT (e.g., [30]–[32]) allowing to route within $O(\log n)$ hops between two such nodes. Such a DHT requires only $O(\log n)$ extra bits of memory in each node. As all nodes of the DHT lie in the ball of radius $\rho\tau$ centered at $a_1(t)$, each hop is of length at most $\tau\rho^2$ in the inframetric, thus yielding an additional length $\tau\rho^2 \log n$ to the routing path.

■

Remarks. Note that $\rho \leq 2$ and $\rho_s \leq 1 + \delta$ in classical metrics for any $\delta > 0$. Setting $\delta = 2\varepsilon$, the stretch factor of our routing scheme becomes $(1+2\varepsilon)/(1-2\varepsilon) = 1+4\varepsilon+o(\varepsilon)$. The above routing scheme may besides result in congestion since many routes may use the same intermediate nodes of the $\rho^i\tau$-nets. Nevertheless, it can be adapted to balance the load by routing through a random node in $B_a(\rho^i\tau)$ for $a \in S_i$, $i \geq 1$, instead of routing to a directly. The routing table sizes are slightly increased by this change.

As any ρ-inframetric is $(\rho, 1)$-skewed, we obtain the following corollary if we relax the skewness assumption.

Corollary 1: For any positive $\varepsilon < 1/\rho$, there exists a compact routing scheme in (V, d) with stretch $(\rho/(1-\rho\varepsilon), \tau\rho^2 \log n)$, table size $\beta\alpha^{3+\log_\rho(\frac{1}{\varepsilon})} \log_\rho(\Delta/\tau) \log n$ bits per node, and node label size $O(\log_\rho(\Delta/\tau) \log n)$ bits.

VI. Discussion

Our approach could be extended in several directions. Our model fits both instantaneous PlanetLab snapshots as well as an average view of all the snapshots of a year. However, RTTs can significantly vary during a day. Studying, modeling and bounding the dynamics of Internet latencies would be a natural follow up of our work. Finding appropriate algorithmic properties could allow to alleviate theoretical limitations in the dynamic setting (as those pointed in [33]).

Another track is to consider routing delays rather than round trip times. This would result in an asymmetric distance function. Our inframetric model could be extended in such a context as long as asymmetry is bounded, e.g., there exists ρ_a such that $d(u, v) \leq \rho_a d(v, u)$ for all u, v.

VII. Acknowledgments

We are grateful to CAIDA for providing Skitter data [23], Chad Yoshikawa for maintaining All-Sites-Pings on PlanetLab [26], P2PSim [25] and Meridian [24] projects for providing their King matrices.

References

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *ACM SIGCOMM Conference*, 1999, pp. 251–262.

[2] I. Abraham, D. Malkhi, and O. Dobzinski, "LAND: stretch $1+\epsilon$ locality-aware networks for DHTs." in *15th ACM-SIAM Symposium on Discrete algorithm (SODA)*, 2004, pp. 550–559.

[3] M. Thorup and U. Zwick, "Approximate distance oracles," in *33rd ACM Symposium on Theory of Computing (STOC)*, 2001, pp. 183–192.

[4] D. V. Krioukov, K. R. Fall, and X. Yang, "Compact routing on internet-like graphs." in *24th Conference of the IEEE Computer and Communications Societies (INFOCOM)*, 2004.

[5] H. T.-H. Chan and A. Gupta, "Small hop-diameter sparse spanners for doubling metrics," in *17th ACM-SIAM Symposium on Discrete algorithm (SODA)*, 2006, pp. 70–78.

[6] P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang, "IDMaps: a global internet host distance estimation service." *IEEE/ACM Transaction on Networking*, vol. 9, no. 5, pp. 525–540, 2001.

[7] D. R. Karger and M. Ruhl, "Finding nearest neighbors in growth-restricted metrics." in *34th ACM Symposium on Theory of Computing (STOC)*, 2002, pp. 741–750.

[8] T. S. E. Ng and H. Zhang, "Predicting internet network distance with coordinates-based approaches." in *21st Conference of the IEEE Computer and Communications Societies (INFOCOM)*, 2002.

[9] C. G. Plaxton, R. Rajaraman, and A. W. Richa, "Accessing nearby copies of replicated objects in a distributed environment." *Theory of Computing Systems*, vol. 32, no. 3, pp. 241–280, 1999.

[10] A. Gupta, R. Krauthgamer, and J. R. Lee, "Bounded geometries, fractals, and low-distortion embeddings," in *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2003, p. 534.

[11] A. Slivkins, "Distance estimation and object location via rings of neighbors." in *24th ACM Symposium on Principles of Distributed Computing (PODC)*, 2005, pp. 41–50.

[12] I. Abraham, C. Gavoille, A. V. Goldberg, and D. Malkhi, "Routing in networks with low doubling dimension." in *26th Int. Conference on Distributed Computing Systems (ICDCS)*, 2006, p. 75.

[13] K. Hildrum, R. Krauthgamer, and J. Kubiatowicz, "Object location in realistic networks," in *16th ACM symposium on Parallelism in algorithms and architectures (SPAA)*, 2004, pp. 25–35.

[14] G. Konjevod, A. Richa, and D. Xia, "Optimal-stretch name-independent compact routing in doubling metrics." in *25th ACM Symposium on Principles of Distributed Computing (PODC)*, 2006, pp. 198–207.

[15] G. Konjevod, A. Richa, D. Xia, and H. Yu, "Compact routing schemes with relaxed guarantees for networks of low doubling dimension," in *26th ACM Symposium on Principles of Distributed Computing (PODC)*, 2007.

[16] E. W. Zegura, K. L. Calvert, and S. Bhattacharjee, "How to model an internetwork," in *14th Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 2, 1996, pp. 594–602.

[17] B. Huffaker, M. Fomenkov, D. Moore, and K. C. Claffy, "Macroscopic analyses of the infrastructure: Measurement and visualization of internet connectivity and performance," in *2nd Workshop on Passive and Active Measurements*. RIPE NCC, 2001.

[18] 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 *6th Internet Measurement Conference (IMC)*. ACM Press, 2006, pp. 85–98.

[19] S. Banerjee, T. Griffin, and M. Pias, "The interdomain connectivity of PlanetLab nodes." in *5th Workshop on Passive and Active Measurements*, ser. Lecture Notes in Computer Science, vol. 3015. Springer, 2004, pp. 73–82.

[20] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson, "The end-to-end effects of internet path selection," in *ACM SIGCOMM Conference*, 2002.

[21] J. M. Kleinberg, A. Slivkins, and T. Wexler, "Triangulation and embedding using small sets of beacons." in *45th Symposium on Foundations of Computer Science (FOCS)*, 2004, pp. 444–453.

[22] A. Slivkins, "Distributed approaches to triangulation and embeddings," in *16th ACM-SIAM Symposium on Discrete algorithm (SODA)*, 2004.

[23] "Skitter project." CAIDA.

[24] B. Wong, A. Slivkins, and E. G. Sirer, "Meridian: a lightweight network location service without virtual coordinates." in *ACM SIGCOMM Conference*, 2005, pp. 85–96, http://www.cs.cornell.edu/People/egs/meridian/.

[25] T. M. Gil, F. Kaashoek, J. Li, R. Morris, and J. Stribling, "P2psim: a simulator for peer-to-peeer protocols." http://pdos.csail.mit.edu/p2psim/.

[26] "All sites pings for planetlab." http://ping.ececs.uc.edu/ping/.

[27] H. T.-H. Chan, A. Gupta, B. M. Maggs, and S. Zhou, "On hierarchical routing in doubling metrics," in *16th ACM-SIAM symposium on Discrete algorithms (SODA)*, 2005, pp. 762–771.

[28] P. K. Gummadi, S. Saroiu, and S. D. Gribble, "King: estimating latency between arbitrary internet end hosts." *Computer Communication Review*, vol. 32, no. 3, p. 11, 2002.

[29] F. Dabek, R. Cox, M. F. Kaashoek, and R. Morris, "Vivaldi: a decentralized network coordinate system." in *ACM SIGCOMM Conference*, 2004, pp. 15–26.

[30] I. Abraham, B. Awerbuch, Y. Azar, Y. Bartal, D. Malkhi, and E. Pavlov, "A generic scheme for building overlay networks in adversarial scenarios," in *17th International Parallel and Distributed Processing Symposium (IPDPS)*, 2003.

[31] P. Fraigniaud and P. Gauron, "D2B: a de Bruijn based content-addressable network," *Theor. Comput. Sci.*, vol. 355, no. 1, pp. 65–79, 2006.

[32] M. Naor and U. Wieder, "Novel architectures for P2P applications: the continuous-discrete approach," in *15th ACM Symposium on Parallel Algorithms (SPAA)*, 2003, pp. 50–59.

[33] D. V. Krioukov, K. Claffy, K. Fall, and A. Brady, "On compact routing for the internet," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 3, 2007.

[34] P. Assouad, "Plongements lipschitzien dans $\mathbb{R}^n$," *Bulletin de la Société Mathématique de France*, vol. 111, pp. 429–448, 1983.