Geolocation of Internet hosts: accuracy limits through Cramér-Rao Lower Bound

Gloria Ciavarrini[a], Maria S. Greco[a], Alessio Vecchio[a]

[a]*Dip. di Ingegneria dell'Informazione, University of Pisa, 56122 Pisa, Italy*

Abstract

With active IP geolocation, the position of an Internet host is estimated by measuring the network delay from a number of other hosts with known position (usually called landmarks). In particular, after having converted delays into distances, geometrical techniques like trilateration are used to provide the estimated position on a global reference system. In this paper, we derive the Cramér-Rao Lower Bound (CRLB) of IP geolocation. The CRLB defines a bound on the minimum mean squared error that affects any unbiased estimator. From a practical point of view, the CRLB provides insights about the maximal theoretical accuracy that can be achieved by IP geolocalization methods. The CRLB also provides conceptual tools useful to understand how the position of landmarks and their distribution affect localization performance. Results show that to obtain accuracy levels in the order of a few tens of kilometers, the number of landmarks to be involved can be relevant and/or their distance from the target cannot be too large.

Keywords: IP geolocation, Cramér-Rao Lower Bound, localization accuracy.

1. Introduction

Knowing the geographical position of an Internet host is useful in a wide range of distributed applications and networking services. Examples include delivery of customized content based on users' positions, and localization of sources of illegal content [1]. Detailed knowledge about the position of hosts also provides the opportunity to include the geographical domain in the study of large scale networks (e.g. for inferring the topology of the Internet [2], or analyzing global routing policies). Unfortunately, the link between an IP address and its geographical position is in general very weak.

Nowadays, Internet hosts are mainly localized with the help of static sources of information, such as registries and databases. In detail, the IP address of the host is used to infer the identity of the organization responsible for such device, then an estimated position of the host is obtained via administrative data. However, the accuracy of these databases is in general not excellent, especially for very large institutions. Some studies made evident that errors in the order of several thousand kilometers are possible [3, 4]. In fact, when a very large organization is involved, the registry may reply with the administrative address of the company headquarter, which can be far away from the real location of the devices under its addressing umbrella. Another problem of this approach is related to the presence of possibly stale information, as database entries are frequently filled by hand.

For this reason, in the last years, research on active IP geolocation methods gained momentum [5, 6, 7, 8, 9, 10, 11, 12]. With active IP geolocation methods, the device to be localized (the target) is georeferenced through network delay measurements. First, the end-to-end delay between the target and a number of landmarks (hosts with known position) is measured using active probes (e.g. by using the ICMP protocol). Then, delays are converted into distances according to a previously defined delay-distance model. Finally, the coordinates of the target are inferred using geometrical techniques (e.g. trilateration) [13]. Nevertheless, also with active IP geolocation, achieving good accuracy is not straightforward, as the localization pro-

URL: gloria.ciavarrini@ing.unipi.it, maria.greco@unipi.it, alessio.vecchio@unipi.it (Alessio Vecchio)

cess is characterized by errors and approximations. In particular, some of these inaccuracies are introduced by time-dependent factors, such as background traffic and processing delay incurred at intermediate routers, whereas others are systematic, such as the assumption that the physical path traversed by probes is the shortest one. To improve localization accuracy, some studies tried to incorporate in the localization procedure also information about intermediate nodes [5, 14]. Other work also discussed the possibility of using smartphones as landmarks (thanks to their self-positioning capability) [15]. Localization error achieved by current techniques ranges from a few tens to a few hundreds of kilometers. Results obtained by different methods are not easily comparable because of the significantly different scenarios and conditions where evaluation has been carried out.

An aspect that has been scarcely investigated is the maximum theoretical accuracy that can be achieved by active IP geolocation methods. In this paper the Cramér-Rao Lower Bound (CRLB) of IP geolocation is derived and analyzed, as far as we know, for the first time. The CRLB represents the theoretical lower bound on the variance of any unbiased estimator of an unknown parameter (in this case the position of the host to be localized). Unbiased estimators that achieve this bound are said to be fully efficient. The CRLB has been extensively used during the last decades to evaluate numerous positioning techniques, as well as the impact of their parameters of operation on localization accuracy. For instance, the CRLB has been used to compute the maximum accuracy of localization systems for mobile devices [16, 17], indoor localization [18], and wireless sensor networks [19, 20, 21]. Gezici provides an overview of the theoretical limits, via CRLB, for a number of algorithms in the context of wireless positioning [22].

The aim of this paper is to determine the maximum theoretical accuracy of active IP geolocation. First, we characterize the delay-distance relationship using a large existing dataset. Our analysis shows that estimated distances are characterized by a Gaussian error with a standard deviation that linearly depends on the distance between the two endpoints. Then, we derive the CRLB for IP geolocation and we use it to compute the maximum accuracy that can be achieved in some reference scenarios. In particular, we evaluate the impact of the following factors: *i*) number of landmarks, *ii*) distance from the target, *iii*) number of probes used to measure the delay. Results show that to obtain relatively small localization errors the number of involved landmarks can be relevant and they cannot be too far from the host to be localized.

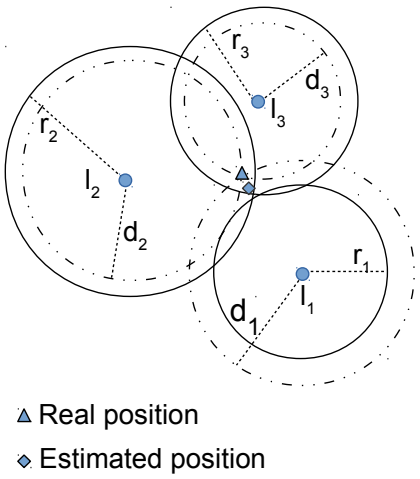


Figure 1: A simple localization scenario.

The remaining of the paper is organized as follows: in Section 2, the assumed localization procedure is presented together with the adopted system model; Section 3 describes the delay-distance relationship and characterizes the associated errors; the Cramér-Rao Lower Bound derived from system model and error characteristics is presented in Section 4; the influence of the above mentioned factors on localization accuracy is discussed in Section 5; Section 6 presents the most relevant work in the field; Section 7 discusses the impact of some parameters of operation and compares results with the ones of a well-known geolocalization method; Section 8 concludes the paper.

2. Localization procedure and system model

Many localization procedures consist of two phases. In the first phase, usually called *ranging*, the distances between the entity to be localized and a number of points with known position are estimated. In the second phase, three or more ranges are used to infer the position of the entity to be localized. Figure 1 shows a simple example: the first phase determines the radius of circles, whereas the second phase produces the estimated position.

In an IP geolocation scenario, an Internet host (the target, T) can be localized by estimating its distance from a number of landmarks. A landmark

is just an Internet host with known geographical coordinates that participates in the localization process. Let $\boldsymbol{l} = [l_1, ..., l_N]^\intercal$ be the set of landmarks. Also, let $\boldsymbol{\theta} = [x\, y]^\intercal$ be the coordinates of the target, and $\boldsymbol{\theta}_i = [x_i\, y_i]^\intercal$ the coordinates of l_i. The distance between a given landmark and the target can be estimated by measuring the one-way communication delay between the two hosts. Since the clocks of the two hosts are not synchronized, a simple method consists in measuring the Round Trip Time (RTT) and halving the result. To measure the RTT, each landmark sends a probe packet (usually an ICMP echo request) and measures the time needed to receive the reply from the target (usually an ICMP echo reply). Everything is done under the assumption that the reply follows the same route of the request (this is not always true in the Internet, but it is generally considered reasonable for the purposes of IP geolocation)[1].

To reduce the error that affects the ranging phase, the delay between a given landmark and the target is collected a number of times. In fact, the experienced delay depends on a number of factors such as background traffic, load of traversed routers, processing load of the target, etc. Let R be the number of values collected, and let $\boldsymbol{m}_i = [m_{i,1}, m_{i,2}, ..., m_{i,R}]$ be the set of one-way delay samples between l_i and the target. To limit the distortions introduced by cross traffic and variable load levels, each landmark selects the minimum observed value. Let us define $\hat{m}_i = \min(\boldsymbol{m}_i)$. Then $\hat{m}_i$ is used for computing the estimated distance r_i between l_i and the target using a function that models the delay-distance relationship.

Let us call $\hat{\boldsymbol{\theta}} = [\hat{x}\ \hat{y}]^\intercal$ the estimated position of T. Let us also define $\boldsymbol{r} = [r_1, r_2, ..., r_N]^\intercal$ the vector of measured distances between T and the landmarks, and $\boldsymbol{d} = [d_1, d_2, ..., d_N]^\intercal$ the vector of real distances between T and the landmarks. Ranging information can be modeled as

$$\boldsymbol{r} = \boldsymbol{d} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{e} = [e_1, e_2, ..., e_N]^\intercal$ is the vector of errors associated to the ranging phase. A circle with radius r_i is thus defined for the i-th landmark. In the ideal case of perfect measurements (i.e. when $\boldsymbol{r} = \boldsymbol{d}$), the N circles have a single intersection. In the real world, the presence of noise in measurements produces situations where the circles have no intersection or where their intersection is not unique (as shown in Figure 1). When $\boldsymbol{e}$ is not a zero vector, the estimated position ($\hat{\boldsymbol{\theta}}$) is generally different from the actual one ($\boldsymbol{\theta}$). The euclidean distance between the actual position of the target and the estimated position is used to evaluate the performance of a localization system.

In summary, we assume that the localization process operates as follows:

1. each landmark measures the communication delay towards the target R times;
2. each landmark l_i selects the minimum observed value $\hat{m}_i$;
3. the estimated distance between l_i and T is calculated as $r_i = g(\hat{m}_i)$, where $g()$ is the function that models the delay-distance relationship;
4. the position of T is estimated using the known positions of landmarks and previously computed distances.

Note that in the considered model the elements of $\boldsymbol{e}$ can be both positive and negative, as they represent the deviation from the average delay-distance behavior. The function $g()$ used to convert delays into distances is based, as better explained in Section 3, on a conversion factor that incorporates all the differences that can be observed for the existing end-to-end paths. For instance, let us consider the circuitousness of a path, i.e. its deviation from the shortest distance calculated along the surface of the Earth. A very circuitous path is characterized by a smaller delay-to-distance conversion factor with respect to a more rectilinear path. The $g()$ function can only incorporate the average level of circuitousness, as knowing the real value for the considered path would require to know the exact position of all intermediate routers. The same applies for other factors, such as the speed of the transmission medium or the number of intermediate nodes.

It is evident that characterizing the delay-distance model (function $g()$) and the nature of noise ($\boldsymbol{e}$) is extremely important, as they have a deep impact on localization accuracy.

3. Delay-distance model

As mentioned, distances between the target and landmarks are estimated by measuring delays and then using a conversion function. In this section

[1] This modus operandi has been adopted in all studies about active IP geolocation.

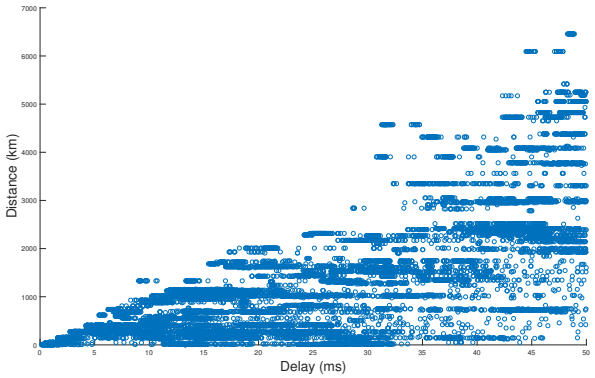


Figure 2: Scatterplot of real distance against observed delay ($\mathcal{D}$ dataset).

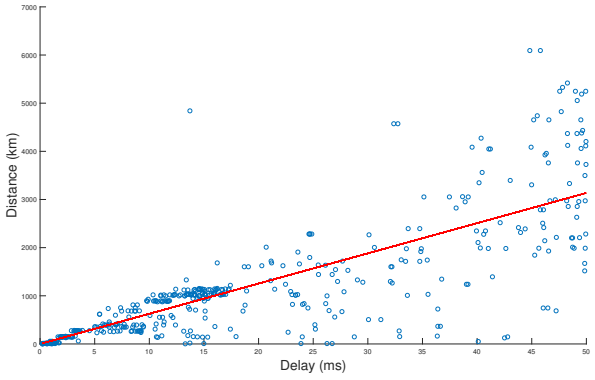


Figure 3: Scatterplot of real distance against observed delay ($\mathcal{D}'$ dataset). Each distance is present only once. The line represents the linear delay-distance model, with coefficient obtained through regression.

we describe how the conversion function has been calculated on a large existing set of end-to-end measurements. We also studied the variance that affects delay measurements for a given distance.

3.1. Experimental data

We built a dataset $\mathcal{D}$ using measurements collected by the PingER project, a measurement infrastructure aimed at studying the end-to-end delay on the Internet [23]. PingER comprises approximately 39 probing machines and 430 probed hosts (the number of both probing machines and probed hosts may slightly vary depending on the considered period). The position of both probing machines and probed hosts is known, thus it is possible to compute the real distance[2] between the two endpoints involved in delay measurements. The position of probing machines and probed hosts is shown in Figure 4.

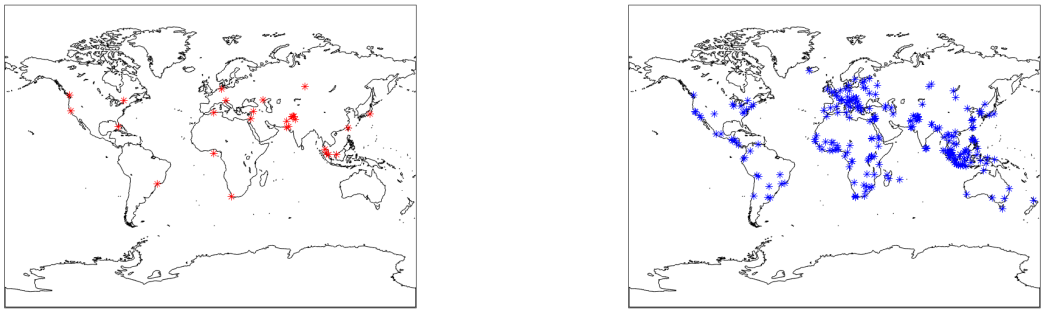
In PingER, each probing machine periodically sends a sequence of ten probes towards a set of probed hosts. Probes are based on ICMP. Collected RTTs are stored on a database and are publicly available through a Web interface. We extracted from the PingER database all measurements comprised in a one year interval.

For each sequence of ten measurements, we selected the minimum value as it corresponds to the measurement affected by the smallest error (as mentioned, errors due to queues and processing loads are additive). The $\mathcal{D}$ dataset consisted of more than 7 million measurements. A scatterplot of $\mathcal{D}$ is shown in Figure 2 (actually, Figure 2 shows a small fraction of all measurements, randomly sampled, for the sake of image clarity). Several horizontal bands are visible in the scatterplot. These bands originate from the large number of delay measurements associated to a single distance. It is evident that some couples of hosts are much more represented than others. This, in turn, means that the dataset is unbalanced in terms of distribution of distances. An unbalanced dataset could be source of possible distortions. For this reason we created a new dataset $\mathcal{D}'$ by purging over-represented distances from $\mathcal{D}$.

More formally, let us define the jth element in $\mathcal{D}$ as (δ_j, μ_j) where δ_j is the real distance between the two hosts involved and μ_j is the minimum observed delay in the sequence of ten measurements. In $\mathcal{D}$, it frequently happens that $\delta_j = \delta_k$ with $j \neq k$. In $\mathcal{D}'$, if $j \neq k$ then $\delta_j \neq \delta_k$. In addition, in $\mathcal{D}'$ only measurements where $\mu_j \leq 50$ ms have been included. We decided to remove all measurements characterized by "very large" delay values as they may be affected by significant errors. It has been shown that when the delay is large, the estimated distance scarcely contributes to the localization process [15]. A scatterplot of $\mathcal{D}'$ is shown in Figure 3.

To avoid possible biases depending on the specific subset of measurements extracted from $\mathcal{D}$ and included in $\mathcal{D}'$, we generated 500 instances of $\mathcal{D}'$. Each instance of $\mathcal{D}'$ is generated by extracting a randomly chosen set of measurement from $\mathcal{D}$ (but still preserving the above properties). Results presented in the remaining of the paper have been ob-

[2]The distance has been calculated as the *great circle distance*, i.e. the the shortest distance between two points on the surface of the Earth measured along the surface of the Earth.



(a) Probing machines

(b) Probed hosts

Figure 4: Position of probing machines and probed hosts.

tained averaging the output for all $\mathcal{D}'$ instances (but we refer to the generic $\mathcal{D}'$ instance for the sake of clarity).

3.2. Model

The relationship between distance and delay is assumed to be linear. This assumption is motivated by the fact that the only component of the end-to-end delay that is dependent on the physical distance is the propagation delay, which increases linearly with the distance. Thus, the delay-distance model we use can be defined by the following equation:

$$\boldsymbol{r} = p \cdot \hat{\boldsymbol{m}} \tag{2}$$

where $\boldsymbol{r}$ is the vector of estimated distances, $\hat{\boldsymbol{m}}$ is the vector of observed minimum delays, and p is a coefficient. We applied linear regression to the $\mathcal{D}'$ dataset, in order to compute p. Result of linear regression is represented by the red line in Figure 3. A value of p equal to $\sim$ 62.7 km/ms has been obtained. This value is in line with previous studies, such as [24]. As also discussed in Section 2, Equation 2 captures the average behavior of the delay-distance relationship. This means that Equation 2 can produce both over-estimations and under-estimations of real distances. For instance, if a given path is characterized by the presence of outdated communication technologies (say characterized by larger than average communication latency), then the estimated distance for a given observed delay will be an over-estimation of the real distance. On the contrary, if a path is characterized by a better than average communication latency, then the real distance is going to be underestimated.

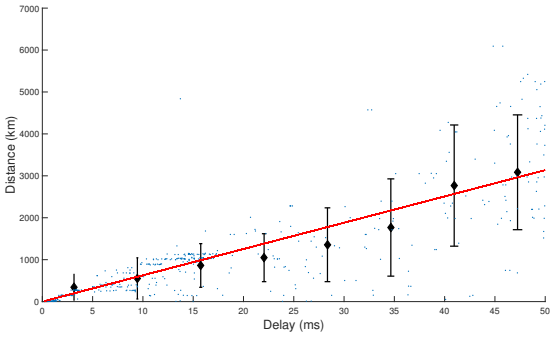


Figure 5: The set of data is divided in eight bins; for each bin the average value of the distances in the bin is depicted as a black diamond, together with the standard deviation of the distance estimation error obtained using the Equation 2.

Polynomial model	$\sigma(\mu) = c + s * \mu$
s	27.29 km/ms
c	32.46 km
R-square	0.9558

Table 1: Fitting results for standard deviation of e against observed delay.

3.3. Analysis of dispersion

It is evident from Figure 3 that the dispersion of measurements increases as the observed delay gets larger. To better study this phenomenon we divided $\mathcal{D}'$ in eight equally spaced bins, according to the value of the observed delay. The eight bins are

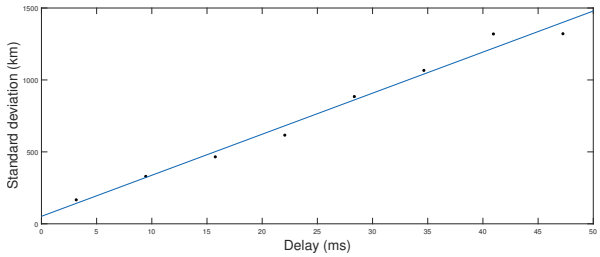


Figure 6: Linear regression of the standard deviation against observed delay (center of bins).

visible in Figure 5. For all elements of $\mathcal{D}'$, we computed the difference e_j between the real distance separating the two hosts involved in the jth measurement and the estimated distance computed using Equation 2 (i.e. $e_j = \delta_j - p \cdot \mu_j$). Figure 5 shows the mean value of δ for each bin, and the standard deviation of e values for the same bin. We used the Kolmogorov-Smirnov test to verify if the set of e values in each bin is compatible with the normal distribution (at 5% significance level). For $\sim 53\%$ of the bins the test does not reject the null hypothesis, whereas for the other bins the null hypothesis is rejected. For the sake of simplicity, we assume that the distribution of e is normal in all bins. We then performed weighted linear regression of the standard deviation of e against the delay corresponding to the center of bin, with weights inversely proportional to the variance of obtained samples. Figure 6 shows the data points used for regression and the line obtained as result. Parameters obtained from linear regression and the value of R-square are reported in Table 1. In the end, $e \sim \mathcal{N}(0, \sigma^2)$ where

$$\sigma = c + k \cdot d \tag{3}$$

with $c = 32.46$ km and $k = 0.44$ (k is obtained as s/p). In practice, this means that the standard deviation of the error that affects measurements increases linearly with distance.

4. Cramér-Rao Lower Bound for IP geolocation

The CRLB provides a lower limit for the covariance matrix of any unbiased estimator:

$$cov(\hat{\boldsymbol{\theta}}) \geq I(\boldsymbol{\theta})^{-1} \tag{4}$$

where $cov(\hat{\boldsymbol{\theta}})$ is equal to $\mathbf{E}_{\boldsymbol{\theta}}\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\intercal}\}$, and $I(\boldsymbol{\theta})$ is the Fisher Information Matrix (FIM). "$X \geq Y$" means that matrix $(X - Y)$ is non-negative definite, and $\mathbf{E}_{\boldsymbol{\theta}}\{\}$ is the expectation conditioned on $\boldsymbol{\theta}$. The inverse of FIM is also called CRLB matrix.

Measured distances are assumed to be statistically independent. The measured distance between T and the ith landmark is $r_i = d_i + e_i$, with $d_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$, and e_i a Gaussian error with zero mean and standard deviation equal to σ_i. According to our model, standard deviation linearly depends from the distance between the target and the landmark: $\sigma_i = c + k \cdot d_i$.

If N is the number of landmarks, the likelihood can be expressed as

$$f(\boldsymbol{r}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} exp(-\frac{(r_i - d_i)^2}{2\sigma_i^2}) \tag{5}$$

and the log-likelihood is equal to

$$\log f(\boldsymbol{r}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log f_i(r_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} (\log(\frac{1}{\sqrt{2\pi}\sigma_i}) - \frac{(r_i - d_i)^2}{2\sigma_i^2}) \tag{6}$$

The FIM $I(\theta)$ has elements defined as follows:

$$[I(\boldsymbol{\theta})]_{u,v} = -\mathbf{E}\left[[\frac{\partial \log f(\boldsymbol{r}|\boldsymbol{\theta})}{\partial\theta_u \partial\theta_v}]\right] \tag{7}$$

with u and $v \in 1..2$ for our bi-dimensional scenario (and thus $\theta_u = x$ and $\theta_v = y$).

The ith component of Equation 6 can be expanded as

$$\log f_i(r_i|\boldsymbol{\theta}) = -\frac{1}{2}\log\left(2\pi\left(c + k\sqrt{(x - x_i)^2 + (y - y_i)^2}\right)^2\right) - \frac{\left(r_i - \sqrt{(x - x_i)^2 + (y - y_i)^2}\right)^2}{2\left(c + k\sqrt{(x - x_i)^2 + (y - y_i)^2}\right)^2} \tag{8}$$

and its second partial derivative with respect to x is:

$$\frac{\partial^2 \log f_i(r_i|\boldsymbol{\theta})}{\partial x^2} = \frac{r_i - \sqrt{E}}{D^2\sqrt{E}} - \frac{A}{C} - \frac{k}{D\sqrt{E}} - \frac{A(r_i - \sqrt{E})}{4D^2E^{\frac{3}{2}}} + \frac{kA}{4DE^{\frac{3}{2}}} + \frac{kB}{D^3\sqrt{E}} + \frac{k^2A}{C} - \frac{3k^2AB}{4D^4E} - \frac{kA(r_i - \sqrt{E})}{D^3E} - \frac{kAB}{4D^3E^{\frac{3}{2}}} \quad (9)$$

where

$$A = (2x - 2x_i)^2$$
$$B = (r_i - \sqrt{E})^2$$
$$C = 4D^2E = 4\sigma_i^2 d_i^2$$
$$D = c + k\sqrt{E} = c + kd_i = \sigma_i$$
$$E = (x - x_i)^2 + (y - y_i)^2 = d_i^2$$

Considering that $\mathbf{E}[r_i] = d_i$ and $\mathbf{E}\left[r_i^2\right] = d_i^2 + \sigma_i^2$, we obtain $\mathbf{E}[B] = d_i^2 + {\sigma_i}^2 + d_i^2 - 2d_i^2 = {\sigma_i}^2$. Then it is easy to find that

$$\mathbf{E}\left[\frac{\partial^2 \log f_i(r_i|\boldsymbol{\theta})}{\partial x^2}\right] = -\frac{(x - x_i)^2(2k^2 + 1)}{{\sigma_i}^2 d_i^2} \quad (10)$$

Symmetrically, we can find that

$$\mathbf{E}\left[\frac{\partial^2 \log f(r_i|\boldsymbol{\theta})}{\partial y^2}\right] = -\frac{(y - y_i)^2(2k^2 + 1)}{{\sigma_i}^2 d_i^2} \quad (11)$$

The mixed partial derivative of the ith element of Equation 6 can be expressed as

$$\frac{\partial^2 \log f(r_i|\boldsymbol{\theta})}{\partial x \partial y} = \frac{\partial^2 \log f(r_i|\boldsymbol{\theta})}{\partial y \partial x} = \frac{kAB}{4EF^{\frac{3}{2}}} - \frac{AB}{C} + \frac{k^2AB}{C} - \frac{AB(r_i - \sqrt{F})}{4E^2F^{\frac{3}{2}}} - \frac{3k^2ABD}{4E^4F} - \frac{kAB((r_i - \sqrt{F}))}{E^3F} - \frac{kABD}{4E^3F^{\frac{3}{2}}} \quad (12)$$

where

$$A = 2y - 2y_i$$
$$B = 2x - 2x_i$$
$$C = 4E^2F = 4{\sigma_i}^2 d_i^2$$
$$D = (r_i - \sqrt{F})^2$$
$$E = c + k\sqrt{F} = \sigma_i$$
$$F = (x - x_i)^2 + (y - y_i)^2 = d_i^2$$

Again, considering that $\mathbf{E}[r_i] = d_i$ and $\mathbf{E}\left[r_i^2\right] = d_i^2 + \sigma_i^2$, we obtain $\mathbf{E}[D] = d_i^2 + {\sigma_i}^2 + d_i^2 - 2d_i^2 = {\sigma_i}^2$. Then it is easy to find that

$$\mathbf{E}\left[\frac{\partial^2 \log f(r_i|\boldsymbol{\theta})}{\partial x \partial y}\right] = \mathbf{E}\left[\frac{\partial^2 \log f(r_i|\boldsymbol{\theta})}{\partial y \partial x}\right] = -\frac{(x - x_i)(y - y_i)(1 + 2k^2)}{{\sigma_i}^2 d_i^2} \quad (13)$$

In the end, the elements of I are the following ones:

$$I_{x,x} = \sum_{i=1}^{N} \frac{(1 + 2k^2)(x - x_i)^2}{\sigma_i^2 d_i^2} \quad (14)$$

$$I_{y,y} = \sum_{i=1}^{N} \frac{(1 + 2k^2)(y - y_i)^2}{\sigma_i^2 d_i^2} \quad (15)$$

$$I_{x,y} = I_{y,x} = \sum_{i=1}^{N} \frac{(1 + 2k^2)(x - x_i)(y - y_i)}{\sigma_i^2 d_i^2} \quad (16)$$

The elements of I can also be expressed as follows:

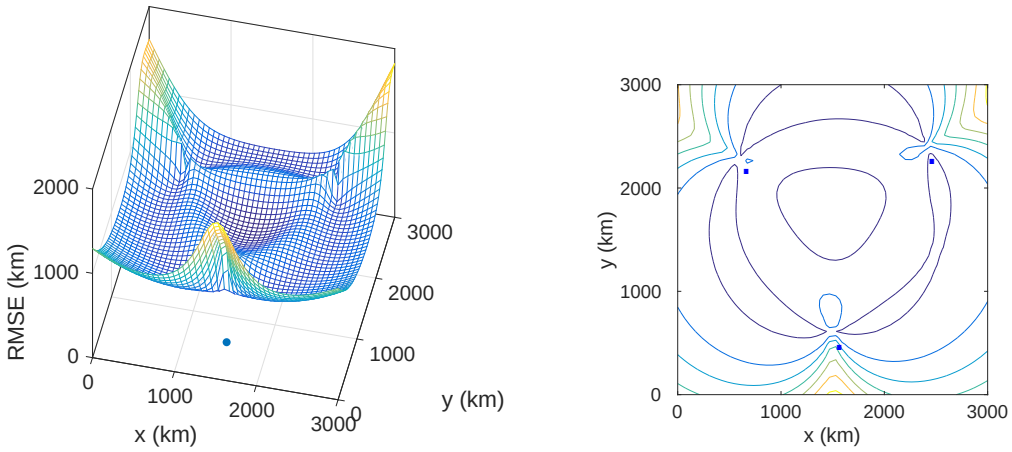
$$I_{x,x} = \sum_{i=1}^{N} \frac{(1 + 2k^2)\cos^2(\alpha_i)}{\sigma_i^2} \quad (17)$$

$$I_{y,y} = \sum_{i=1}^{N} \frac{(1 + 2k^2)\sin^2(\alpha_i)}{\sigma_i^2} \quad (18)$$

$$I_{x,y} = I_{y,x} = \sum_{i=1}^{N} \frac{(1 + 2k^2)\sin(\alpha_i)\cos(\alpha_i)}{\sigma_i^2} \quad (19)$$

where α_i is the angle between the target and the ith landmark.

The CRLB defines a lower bound on the mean square error (MSE) of a position estimate:



(a) Accuracy: 3D representation

(b) Accuracy: level curves

Figure 7: Accuracy when using 3 landmarks.

$$MSE = \mathbf{E}_{\boldsymbol{\theta}}\{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2\} = trace\{\mathbf{E}_{\boldsymbol{\theta}}\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\intercal}\}\} \geq trace\{I(\boldsymbol{\theta})^{-1}\} = MMSE \quad (20)$$

where MMSE is the minimum mean square error.

Localization accuracy can be defined as the square root of the minimum mean square error (RMSE).

5. Numeric analysis

In this section we present a numerical analysis of localization accuracy, as resulting from the CRLB previously found. In particular, we studied how localization accuracy is influenced by the most important parameters of operation: the number of landmarks, the distance between target and landmarks, and the number of probes

The area where both target and landmarks are located is supposed to be 3000 km x 3000 km (for a comparison we remind that the surface of USA is approximately 9.8 millions km^2 and the area of EU is approximately 4.5 millions km^2). The N landmarks are randomly scattered over the considered area.

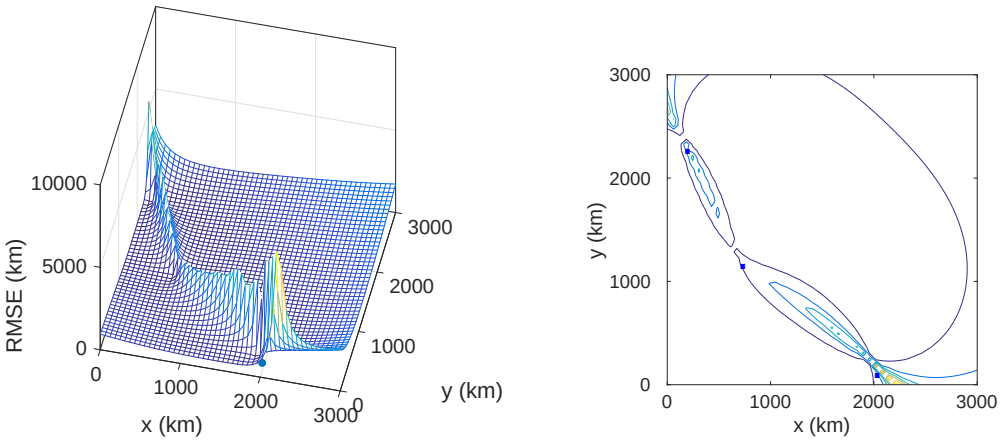
5.1. General considerations

Figure 7 shows the value of RMSE in a scenario with three landmarks. The two figures depict the same results as a 3D mesh and as a level plot, to make evident the location of the three landmarks and the accuracy levels. It is straightforward to notice that the area delimited by the three landmarks is characterized by better localization accuracy (approximately 470 km). The regions characterized by the worst RMSE values ($\sim$ 1800 km) are the ones in proximity of these points: (1500 km, 0 km), (0 km, 3000 km), (3000 km, 3000 km). The average value of accuracy on the whole region is $\sim$ 750 km. These values are due only to the geographical distribution of the considered landmarks and their distances from the considered points.

Inaccurate selection of landmarks may cause significant degradation of accuracy values. For instance, Figure 8 shows a scenario where the three landmarks are almost collinear. In this case, the maximum RMSE value, near the (3000 km, 0 km) corner, is approximately 7770 km. This value of RMSE is $\sim$ 4.2 times higher with respect to the maximum RMSE of the previously considered scenario (depicted in Figure 7). Also the average value of accuracy for the whole region worsens, as it becomes equal to 885 km.

RMSE values for the above three-landmark examples are rather high. This is due to the small number of landmarks and to the large distances between landmarks and possible positions of the target (the errors that affect ranges increase with distance).

These considerations may seem obvious and some of the guidelines for the selection of landmarks are somehow intuitive. Nevertheless, the use of the CRLB allows the designer of an IP geolocation system to evaluate quantitatively the impact of land-



(a) Accuracy: 3D representation (b) Accuracy: level curves

Figure 8: An example of inaccurate selection of landmarks.

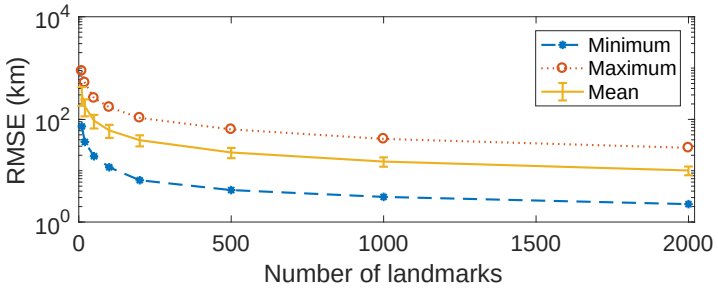


Figure 9: Accuracy when varying the number of landmarks.

marks selection. For example, given a set of landmarks, if the average localization accuracy is considered as not satisfactory, the localization system may use the CRLB to select those additional landmarks that provide the highest improvement (instead of enrolling additional landmarks that may bring only marginal gain).

5.2. Varying the number of landmarks

Increasing the number of landmarks that participate to the localization procedure usually provides better accuracy. Obviously, this comes at the cost of additional traffic and complexity of coordination. Figure 9 shows how RMSE is influenced by the number of landmarks. In particular, the mean value of RMSE is shown when the number of landmarks is in the $\{10, 20, 50, 100, 200, 500, 1000, 2000\}$ set. Results are averaged over 30 executions, where each execution is characterized by a different and randomly chosen placement of landmarks. Figure 9 also reports the average minimum and maximum values of RMSE that are achieved with a given number of landmarks. First, it can be noticed that increasing the number of landmarks, not only provides better results in terms of average RMSE, but it also reduces its variability. This is caused by the fact that as the number of landmarks gets larger, the area is covered more uniformly. Second, using a very large number of landmarks (say more than 500) provides a marginal gain in terms of accuracy (for the considered area). In fact, doubling the number of landmarks (to one thousand), causes the average RMSE to change from $\sim$ 22 km to $\sim$ 15 km. In this case, CRLB provides to the designer of an IP geolocation system the opportunity to quantitatively evaluate the possible benefits associated to the involvement of additional resources (landmarks). Third, even when using a very large number of landmarks, e.g. one thousand, the localization error can be still be in the order of a few tens of kilometers. These accuracy levels can be incompatible with some applications. In addition, they are reached using a number of landmarks that could be unpractical to manage in a real system.

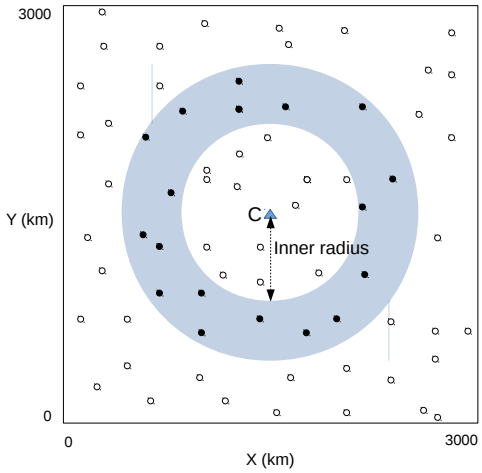


Figure 10: The W landmarks used in the localization process are the ones contained in the annulus; the inner radius of the annulus is progressively increased.

5.3. *Varying the distance between target and landmarks*

We studied how the CRLB is affected by the distance between target and landmarks. In the 3000x3000 km^2 area previously considered, 500 landmarks have been randomly placed. We then computed the RMSE value at point C with coordinates (1500 km, 1500 km), i.e. the center of the region. The RMSE value has been calculated using W landmarks belonging to an annulus, as depicted in Figure 10. The inner radius of the annulus has been varied to increase the distance between the W landmarks and C. The whole procedure has been repeated 100 times, each with a different initial placement of landmarks. Figure 11 shows three curves, representing the average value of RMSE, computed for three values of W, respectively equal to 10, 50, and 100. It is evident that when the distance between target and landmarks increases, the RMSE value increases as well. For the three curves, the standard deviation of obtained values is also reported.

5.4. *Varying the number of probes*

As mentioned, to measure the communication delay between the target and a landmark, the latter sends a sequence of probes (the number of probes is equal to R) and selects the minimum value ($\hat{m}_i$). This is done to remove, as far as possible, the additional delay introduced by queues and processing load. Increasing the number of probes may provide a better estimate of the distance between two

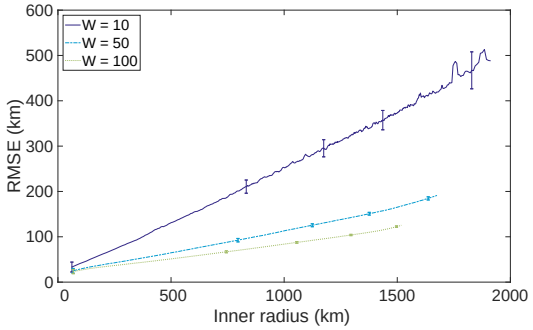


Figure 11: Accuracy when varying the distance between target and landmarks.

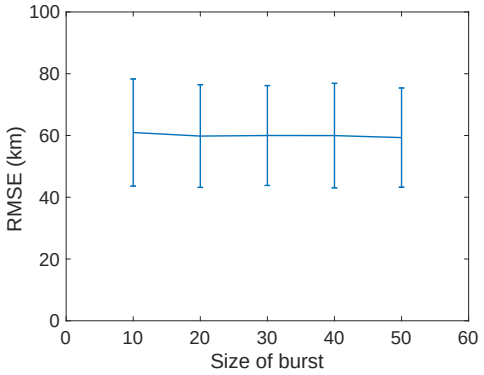


Figure 12: Accuracy when varying the size of bursts.

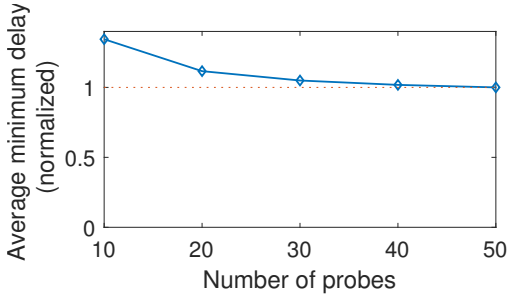


Figure 13: Average minimum delay when varying the number of probes; values are normalized against the average minimum value obtained when using 50 probes.

hosts. On the other hand, the use of large R values has negative effects as well: the duration of the localization process increases, the global amount of generated traffic gets larger, and the target has to reply to a possibly suspicious number of probes (the localization activity could be mistaken as a denial of service attack).

Figure 12 shows the accuracy of localization when the number of probes used by each landmark is

is varied between 10 and 50 (the number of landmarks has been kept fixed and equal to 100). Localization accuracy does not improve significantly when the number of probes is increased. Figure 13 shows the average minimum value obtained when using a number of probes in the 10-50 interval; values are normalized using the average minimum obtained when using 50 probes. Two considerations can be made. First, the curve is almost flat on the right-hand side of Figure 13: this means that 50 probes are sufficient to obtain a minimum value that is reasonably close to the real one (using a larger number would produce small changes in the observed delay, according to our dataset). Second, variations in the observed delay have a moderate effect on localization accuracy. The $\hat{m}$ obtained when using 10 probes is $\sim 30\%$ higher than the one obtained when using 50 probes, whereas the difference in RMSE obtained when using 10 and 50 probes is in the order of $\sim 3\%$. This is due to the fact that increasing the number of probes helps to alleviate the distortions introduced by run-time factors, such as queues and processing time. However, other sources of error in the distance estimation process are not affected by the use of a possibly larger number of probes. Examples include the different degrees of circuitousness of end-to-end paths and the heterogeneity in transmission technologies [25].

6. Related work

This section summarizes the most relevant approaches for geolocating an IP address using active measurements. Also some related studies about CRLB, but in other localization domains, are included.

Constraint-Based Geolocation (CBG) is a localization method based on geographical constraints [6]. Each landmark measures the RTT towards the target and estimates the distance between them. Then, a circular feasible region is determined for each landmark. Finally, the intersection of these regions is used for obtaining a small geographic area where the target must be located. CBG includes a calibration phase aimed at deriving an accurate model of the delay-distance relationship. During calibration, each landmark measures the RTTs with respect to all other landmarks and uses these measurements to define a landmark-specific linear distance estimator.

GeoBuD is another geolocation method based on constraints [14]. GeoBuD improves on CBG by estimating the amount of buffering occurring at intermediate routers, and thus obtaining a more accurate delay-distance model. This produces more stringent constraints and, as a consequence, a smaller intersection region. Experiments carried out on PlanetLab show that GeoBuD is able to obtain a smaller localization error with respect to CBG. For instance, in one of the considered datasets, the median error is 144 km for GeoBuD and 228 km for CBG. Another geolocation systems that uses information about intermediate routers is Topology-based Geolocation (TBG) [5].

The Octant framework uses both positive information, i.e. information related to where a node may be located, and negative information, i.e. information about where a node cannot be located [26]. Landmarks are calibrated by measuring the delays towards all other landmarks. The convex hull containing the delay-distance data points is used to calculate both upper and lower bounds in the delay to distance conversion. To cope with circuitousness of paths, Octant geolocalizes also intermediate routers and uses them as secondary landmarks. RTTs are measured by sending ten ICMP probes and selecting the minimum value. An experimental evaluation has been carried out using the PlanetLab network in North America.

A technique based on Maximum Likelihood Estimation is presented in [27]. The authors collected a set of delay-distance measurements using PlanetLab (limited to the North American region). Samples have been divided in bins, then the mean value for each bin has been computed to obtain a linear relationship between distance and latency. The geolocation algorithm estimates the position of the target by maximizing the likelihood function (on the base of measured delays). The maximum is found via exhaustive search on the considered area.

The relationship between delay and other geographic and network properties is discussed in [28]. The analysis is based on information theory and evaluates the amount of information that network and geographic variables provide about other properties. Data has been modeled as a 7-dimensional discrete random variable. The conditional entropy of a variable of interest with respect to observed variables has been evaluated. This has been done for both quantitative variables (such as distance and RTT) and categorical variables (such as countries or subcontinental zones). According to the authors, the strongest predictor for great circle distance is the pair of countries the hosts belong to.

This is however an information that cannot always be available via network measurements. The analysis is based on a dataset comprising $\sim$ 200M RTT samples between $\sim$ 54K DNS servers. The position of the hosts involved in the study has been extracted from IP geolocation databases.

Padmanabhan et al. [1] proposed three different techniques: GeoTrack, which uses DNS based information to infer the position of the target host, GeoPing, which is based on delay measurements collected from a number of landmarks, and GeoCluster, which groups IP addresses and then finds their positions using information from several sources.

GeoGet is a geolocation system that uses HTTP requests for evaluating the delay between the target and the set of landmarks [9]. In detail, the target plays the active role and initiates measurements towards Web servers with known locations. Multiple measurements towards a single server are executed, then the minimum value is used for further processing. The estimated position of the target is set to the landmark with smallest RTT (thus, in this case, there is no delay-to-distance conversion).

The CRLB has been extensively used for evaluating the efficiency of position estimators in different localization scenarios. Qi et al. evaluated via CRLB the accuracy of a number of positioning methods (based on time of arrival, time difference of arrival, angle of arrival, and signal strength) [16]. The properties of the CRLB for the time difference of arrival localization technique is also discussed by Yang et al. [29]. The CRLB of a localization method based on received signal strength has been refined to incorporate also the effects due to signal power and frequency [30].

In the context of wireless sensor networks, a geometric interpretation of the CRLB was given both for anchored localization and for anchor-free localization [20]. In the anchored localization model there are at least three nodes with known positions. In anchor-free localization no nodes have known positions and the only information available is the inter-node distance measurement. Authors show that local geometry can be used to predict localization accuracy. The use of CRLB for localization in wireless sensor networks is also discussed by Savvides et al. [21].

7. Discussion

As mentioned, circuitousness of paths is one of the "confounding factors" in the delay-distance relationship. It is known that the level of circuitousness is not constant across the different regions of the globe. For instance, some paths in the African Internet are characterized by higher-than-average levels of circuitousness [31]. This is due, in large part, to the fact that local ISPs are often not present at local exchange points or, when present, they frequently do not have peering agreements. This reduces the availability of direct paths between network sources and destinations. Round trip times are also influenced by the adopted policies for both intra- and inter-domain routing: Zheng et al. [32] presented cases of structural violations of triangle inequality caused, for instance, by hot potato routing (another source of violations is represented by the use of link weights that are strongly divergent from their geographical extension). In general, different levels of circuitousness can be observed depending on the economical, technological, and geographical characteristics of the considered region. Similar considerations can be made for the other factors influencing the delay-distance model, such as the adopted communication technologies. In the end, the imperfect homogeneity of all these influencing elements is the cause of possible geographically-dependent fluctuations of the coefficients used in Equations 2 and 3.

However, the impact of these fluctuations on the analysis presented in this paper is limited: an IP geolocation method is supposed to operate on a global scale, thus it makes sense to understand which can be its average localization accuracy in such global scenario. Obviously, in some regions the effectively reachable accuracy will be better than the average, whereas in others it will be worse than the average. Moreover, the main contribution of this work is, in our opinion, represented by the analytical expression of the CRLB presented in Section 4. The numerical analysis has then been carried out using the parameter values we found for the global scenario using the PingER dataset, as we felt this is the most common way of operation. If a regional analysis is of interests (e.g. for a system operating only in Europe or in Africa), it is sufficient to recompute the parameters (e.g. the values of p and k in Equations 2 and 3) for the considered region and substitute the values into the provided CRLB expression.

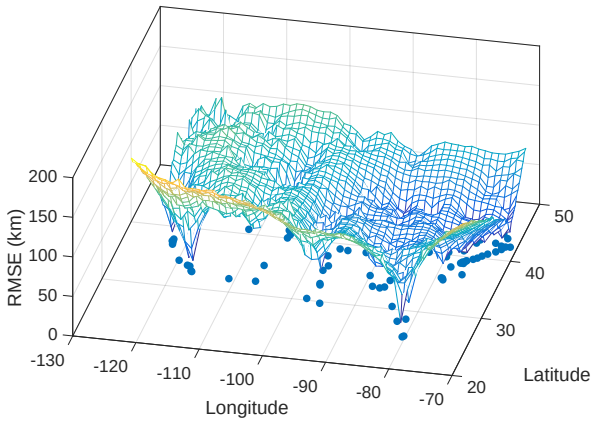


Figure 14: Accuracy of localization in the USA when using the hosts of NLARN AMP [33] as landmarks.

We compared the CRLB derived in this paper with the results obtained by CBG. We chose CBG because it has in turn been used as a comparison term in several subsequent works. The performance of CBG was originally evaluated using hosts with known positions. In particular, the NLARN AMP dataset [33] was used to compute the accuracy of CBG on 95 hosts in the USA. CBG geolocated each host one at a time, using the remaining hosts as landmarks. The mean error distance of CBG on such dataset is 182 km. The median error and the 80th percentile on the same dataset are 95 km and 277 km.

We evaluated the CRLB using the same locations of targets and landmarks used for evaluating CBG. In particular, we used Equations 17, 18, 19 with the great circle distance as d. The median, mean, and 80th percentile values of RMSE are equal to 44 km, 51 km, and 70 km, respectively. These values are smaller than the ones obtained by CBG. This is expected as the CRLB provides the maximum accuracy that can be obtained by an ideal non-biased estimator. These values suggest that some improvements are still possible.

Since the delay-distance model we used has been generated using only delays below 50 ms, we carried out the same analysis taking into account this factor. We re-evaluated the CRLB using only landmarks whose distance from the target is below the distance corresponding to the maximum considered delay. Results (median, mean, and 80th percentile of RMSE) changed minimally (fractions of km). This also confirms that distant landmarks are able to provide only limited information.

Finally, we evaluated the CRLB on the entire surface of the USA using the NLARN AMP hosts as landmarks. Figure 14 shows the results as a 3D mesh. The right-hand side of Figure 14 corresponds to the the East coast of the US. The position of some landmarks is visible, whereas others are hidden by the mesh. "Peaks" and "valleys" are just caused by the non uniform distribution of landmarks on the USA surface.

8. Conclusion

Literature discussing the importance of the CRLB for localization purposes is rather abundant, and attention generally focused on its adoption in the context of wireless sensor networks or for localizing mobile terminals. Somehow surprisingly, as far as we know, its use for a deeper understanding of IP geolocation has never been considered.

Since the CRLB provides a measure of the maximum theoretical accuracy that can be achieved by a localization system, we advocate its use as a baseline for assessing the performance of future and existing localization systems. If a real implementation obtains accuracy results that are close to the ones expressed by the CRLB, there is little reason to continue research in such direction, as significant improvements are difficult to achieve.

It is worthwhile to notice that the CRLB here presented has been derived from the observation of the minimum delay in a sequence of measurements, and that changing the observed variable may lead to different results. However, it is also important to stress out that using the minimum of a sequence of measurement is common practice in IP geolocation, as it reduces the impact of additional noise found at run-time.

Results show that obtaining a localization error below 20 km requires the use of a number of landmarks so large to be unpractical (when the area of operation has the size of a continent). Thus, measurement-based geolocation can be incompatible with applications with very stringent requirements in terms of accuracy. Results obtained when varying the distance between landmarks and target can be useful for the design of future IP geolocation systems. For instance, the analysis suggests that a multi-stage approach can be beneficial: first, a number of landmarks in proximity of the target are identified (e.g. using non measurement-based information), then the measurement-based phase takes

place using only these landmarks (which are characterized by reduced distance from the target).

Acknowledgment

This work was partially funded by the University of Pisa (project "PRA 2017_37 - IoT e Big Data: metodologie e tecnologie per la raccolta e l'elaborazione di grosse moli di dati").

References

[1] V. N. Padmanabhan, L. Subramanian, An investigation of geographic mapping techniques for Internet hosts, SIGCOMM Comput. Commun. Rev. 31 (4) (2001) 173–185. doi:10.1145/964723.383073.
URL http://doi.acm.org/10.1145/964723.383073

[2] E. Gregori, L. Lenzini, V. Luconi, A. Vecchio, Sensing the Internet through crowdsourcing, in: IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013, pp. 248–254. doi:10.1109/PerComW.2013.6529490.

[3] Y. Shavitt, N. Zilberman, A geolocation databases study, IEEE Journal on Selected Areas in Communications 29 (10) (2011) 2044–2056.

[4] S. S. Siwpersad, B. Gueye, S. Uhlig, Proceedings of the 9th International Conference on Passive and Active Network Measurement, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, Ch. Assessing the Geographic Resolution of Exhaustive Tabulation for Geolocating Internet Hosts, pp. 11–20.

[5] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, Y. Chawathe, Towards IP Geolocation Using Delay and Topology Measurements, in: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, IMC '06, ACM, New York, NY, USA, 2006, pp. 71–84. doi:10.1145/1177080.1177090.
URL http://doi.acm.org/10.1145/1177080.1177090

[6] B. Gueye, A. Ziviani, M. Crovella, S. Fdida, Constraint-Based Geolocation of Internet Hosts, IEEE/ACM Transactions on Networking 14 (6) (2006) 1219–1232. doi:10.1109/TNET.2006.886332.

[7] S. Laki, P. Mátray, P. Hága, I. Csabai, G. Vattay, A model based approach for improving router geolocation, Comput. Netw. 54 (9) (2010) 1490–1501. doi:10.1016/j.comnet.2009.12.004.
URL http://dx.doi.org/10.1016/j.comnet.2009.12.004

[8] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, Y. Zhang, Mining the Web and the Internet for Accurate IP Address Geolocations, in: INFOCOM 2009, IEEE, 2009, pp. 2841–2845. doi:10.1109/INFCOM.2009.5062243.

[9] Y. Tian, R. Dey, Y. Liu, K. W. Ross, Topology Mapping and Geolocating for China's Internet, IEEE Transactions on Parallel and Distributed Systems 24 (9) (2013) 1908–1917. doi:10.1109/TPDS.2012.271.

[10] D. Feldman, Y. Shavitt, N. Zilberman, A structural approach for PoP geo-location, Computer Networks 56 (3) (2012) 1029 – 1040. doi:http://dx.doi.org/10.1016/j.comnet.2011.10.029.
URL http://www.sciencedirect.com/science/article/pii/S1389128611004191

[11] Z. Hu, J. Heidemann, Y. Pradkin, Towards geolocation of millions of IP addresses, in: Proceedings of the 2012 ACM Conference on Internet Measurement Conference, IMC '12, ACM, New York, NY, USA, 2012, pp. 123–130. doi:10.1145/2398776.2398790.
URL http://doi.acm.org/10.1145/2398776.2398790

[12] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z. Zhang, Y. Zhang, IP-geolocation mapping for moderately connected Internet regions, IEEE Transactions on Parallel and Distributed Systems 24 (2) (2013) 381–391. doi:10.1109/TPDS.2012.136.

[13] A. Ziviani, S. Fdida, J. F. de Rezende, O. C. M. Duarte, Improving the accuracy of measurement-based geographic location of Internet hosts, Computer Networks 47 (4) (2005) 503 – 523. doi:http://dx.doi.org/10.1016/j.comnet.2004.08.013.
URL http://www.sciencedirect.com/science/article/pii/S138912860400249X

[14] B. Gueye, S. Uhlig, A. Ziviani, S. Fdida, Proceedings of the 5th International IFIP-TC6 Networking Conference, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, Ch. Leveraging Buffering Delay Estimation for Geolocation of Internet Hosts, pp. 319–330.

[15] G. Ciavarrini, V. Luconi, A. Vecchio, Smartphone-based geolocation of Internet hosts, Computer Networks 116 (2017) 22 – 32. doi:http://dx.doi.org/10.1016/j.comnet.2017.02.006.
URL http://www.sciencedirect.com/science/article/pii/S1389128617300440

[16] Y. Qi, H. Kobayashi, H. Suda, Analysis of wireless geolocation in a non-line-of-sight environment, IEEE Transactions on Wireless Communications 5 (3) (2006) 672–681. doi:10.1109/TWC.2006.1611097.

[17] C. Fritsche, A. Klein, Cramér-Rao lower bounds for hybrid localization of mobile terminals, in: Proceedings of the 5th Workshop on Positioning, Navigation and Communication, 2008, pp. 157–164. doi:10.1109/WPNC.2008.4510370.

[18] A. K. M. M. Hossain, W. S. Soh, Cramer-Rao bound analysis of localization using signal strength difference as location fingerprint, in: Proceedings of IEEE INFOCOM, 2010, pp. 1–9. doi:10.1109/INFCOM.2010.5462020.

[19] N. Patwari, J. Ash, S. Kyperountas, A. Hero, R. Moses, N. Correal, Locating the nodes: cooperative localization in wireless sensor networks, Signal Processing Magazine, IEEE 22 (4) (2005) 54–69. doi:10.1109/MSP.2005.1458287.

[20] C. Chang, A. Sahai, Estimation bounds for localization, in: Proceedings of the First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2004, pp. 415–424. doi:10.1109/SAHCN.2004.1381943.

[21] A. Savvides, W. Garber, S. Adlakha, R. Moses, M. B. Srivastava, On the error characteristics of multihop node localization in ad-hoc sensor networks, in: Proceedings of the 2nd International Conference on Information Processing in Sensor Networks, IPSN'03, Springer-Verlag, Berlin, Heidelberg, 2003, pp. 317–332.
URL http://dl.acm.org/citation.cfm?id=1765991.1766013

[22] S. Gezici, A survey on wireless position estimation, Wireless Personal Communications 44 (3) (2008) 263–

282.

[23] W. Matthews, L. Cottrell, The PingER project: active Internet performance monitoring for the HENP community, IEEE Communications Magazine 38 (5) (2000) 130–136. doi:10.1109/35.841837.

[24] S. Laki, P. Matray, P. Haga, I. Csabai, G. Vattay, A detailed path-latency model for router geolocation, in: Testbeds and Research Infrastructures for the Development of Networks Communities and Workshops, 2009. TridentCom 2009. 5th International Conference on, 2009, pp. 1–6. doi:10.1109/TRIDENTCOM.2009.4976258.

[25] P. Mátray, P. Hága, S. Laki, G. Vattay, I. Csabai, On the spatial properties of Internet routes, Comput. Netw. 56 (9) (2012) 2237–2248. doi:10.1016/j.comnet.2012.03.005.
URL http://dx.doi.org/10.1016/j.comnet.2012.03.005

[26] B. Wong, I. Stoyanov, E. G. Sirer, Octant: A comprehensive framework for the geolocalization of Internet hosts, in: Proceedings of the 4th USENIX Conference on Networked Systems Design and Implementation, NSDI, USENIX Association, Berkeley, CA, USA, 2007, pp. 23–23.
URL http://dl.acm.org/citation.cfm?id=1973430.1973453

[27] M. J. Arif, S. Karunasekera, S. Kulkarni, A. Gunatilaka, B. Ristic, Internet host geolocation using maximum likelihood estimation technique, in: Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications, IEEE, 2010, pp. 422–429.

[28] R. Landa, R. G. Clegg, J. T. Araújo, E. Mykoniati, D. Griffin, M. Rio, Measuring the relationships between Internet geography and RTT, in: Proceedings of the 22nd International Conference on Computer Communication and Networks (ICCCN), IEEE, 2013, pp. 1–7.

[29] B. Yang, J. Scheuing, Cramer-Rao bound and optimum sensor array for source localization from time differences of arrival., in: ICASSP (4), 2005, pp. 961–964.

[30] X. Zheng, H. Liu, J. Yang, Y. Chen, R. P. Martin, X. Li, A study of localization accuracy using multiple frequencies and powers, IEEE Transactions on Parallel and Distributed Systems 25 (8) (2014) 1955–1965.

[31] A. Gupta, M. Calder, N. Feamster, M. Chetty, E. Calandro, E. Katz-Bassett, Peering at the internet's frontier: A first look at isp interconnectivity in africa, in: Proceedings of the 15th International Conference on Passive and Active Measurement - Volume 8362, PAM 2014, Springer-Verlag New York, Inc., New York, NY, USA, 2014, pp. 204–213. doi:10.1007/978-3-319-04918-2_20.

[32] H. Zheng, E. K. Lua, M. Pias, T. G. Griffin, Internet routing policies and round-trip-times, in: Proceedings of the 6th International Conference on Passive and Active Network Measurement, PAM'05, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 236–250. doi:doi:10.1007/978-3-540-31966-5_19.

[33] NLARN AMP dataset, https://www.ripe.net/analyse/raw-data-sets, accessed: 2018-01-15.