Constraint-Based Geolocation of Internet Hosts

Bamba Gueye, Artur Ziviani, *Member, IEEE*, Mark Crovella, *Member, IEEE*, and Serge Fdida, *Member, IEEE*

***Abstract*—Geolocation of Internet hosts enables a new class of location-aware applications. Previous measurement-based approaches use reference hosts, called landmarks, with a well-known geographic location to provide the location estimation of a target host. This leads to a discrete space of answers, limiting the number of possible location estimates to the number of adopted landmarks. In contrast, we propose Constraint-Based Geolocation (CBG), which infers the geographic location of Internet hosts using multilateration with distance constraints to establish a continuous space of answers instead of a discrete one. However, to use multilateration in the Internet, the geographic distances from the landmarks to the target host have to be estimated based on delay measurements between these hosts. This is a challenging problem because the relationship between network delay and geographic distance in the Internet is perturbed by many factors, including queueing delays and the absence of great-circle paths between hosts. CBG accurately transforms delay measurements to geographic distance constraints, and then uses multilateration to infer the geolocation of the target host. Our experimental results show that CBG outperforms previous geolocation techniques. Moreover, in contrast to previous approaches, our method is able to assign a confidence region to each given location estimate. This allows a location-aware application to assess whether the location estimate is sufficiently accurate for its needs.**

***Index Terms*—Delay measurement, geolocation, internet, multilateration, position measurement.**

I. Introduction

NOVEL location-aware applications could be enabled by an efficient means of inferring the geographic location of Internet hosts. Examples of such location-aware applications include targeted advertising on web pages, automatic selection of a language to display content, restricted content delivery following regional policies, and authorization of transactions only when performed from pre-established locations. Inferring the location of Internet hosts from their IP addresses is a challenging problem because there is no direct relationship between the IP address of a host and its geographic location [1].

Previous work on the measurement-based geolocation of Internet hosts [2], [3] uses the positions of reference hosts, called landmarks, with well-known geographic location as the possible location estimates for the target host. This leads to a discrete space of answers; the number of answers is equal to the number of reference hosts, which can limit the accuracy of the resulting location estimation. This is because the closest reference host may still be far from the target.

To overcome this limitation, we propose the Constraint-Based Geolocation (CBG) approach, which infers the geographic location of Internet hosts using multilateration. Multilateration refers to the process of estimating a position using a sufficient number of distances to some fixed points. As a result, multilateration establishes a continuous space of answers instead of a discrete one. We use a set of landmarks to estimate the location of other hosts. The fundamental idea is that given geographic distances to a given target host from the landmarks, an estimation of the location of the target host would be feasible using multilateration, just as the Global Positioning System (GPS) [4] does. However, to use multilateration in the Internet, geographic distances from the landmarks to the target host have to be estimated based on delay measurements between these hosts. This is a challenging task because delay measurements cannot always be transformed accurately to geographic distances, since network delay is not necessarily well correlated with geographic distance. This happens because the relationship between network delay and geographic distance in the Internet is perturbed by many factors, including queueing delays, violations of triangle inequality [5], and the absence of great-circle paths between hosts [6].

In recent years, several propositions, such as GNP [7], Virtual Landmarks [8], and Vivaldi [9], have addressed the evaluation of network proximity between Internet hosts using coordinate systems. Nevertheless, distance in the context of network proximity problems refers to the network delay between a pair of Internet hosts. In contrast, for geolocating hosts, distances refer to actual geographic distances between hosts. Therefore, to the best of our knowledge, CBG is the first effort to use multilateration for geolocating Internet hosts.

A key element of CBG is its ability to accurately transform delay measurements into distance constraints. The starting point is the fact that digital information travels along fiber optic cables at almost exactly 2/3 the speed of light in a vacuum [10]. This means that any particular delay measurement immediately provides an *upper bound* on the great-circle distance between the endpoints. The upper bound is the delay measurement divided by the speed of light in fiber. Thus, from the standpoint of a particular pair of endpoints, there is some theoretical minimum delay for packet transmission, dictated by the great-circle distance between them. Therefore, no matter the reason (*e.g.* queueing delays, violations of the triangle inequality, absence of great-circle paths between hosts, and so on), the actual measured delay between them involves only an *additive* distortion.

Manuscript received September 30, 2004; revised July 11, 2005 and October 27, 2005; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor R. Caceres. This work was supported in part by FAPERJ, CNPq, Euronetlab, the National Science Foundation under Grants CCR-0325701 and ANI-0322990, and by a grant from Intel. A short version of this paper was presented at the ACM/SIGCOMM Internet Measurement Conference (IMC'04), October, 2004.

B. Gueye and S. Fdida are with the Laboratoire d'Informatique de Paris 6 (LIP6/CNRS), Université Pierre et Marie Curie (Paris 6), Paris, France (e-mail: gueye@rp.lip6.fr; fdida@rp.lip6.fr).

A. Ziviani is with the National Laboratory for Scientific Computing (LNCC), Petrópolis, Brazil (e-mail: ziviani@lncc.br).

M. Crovella is with the Department of Computer Science, Boston University, Boston, MA 02215 USA (e-mail: crovella@cs.bu.edu).

Digital Object Identifier 10.1109/TNET.2006.886332

However, if CBG were to use simple delay measurements directly to infer distance constraints, it would not be very accurate. For accurate results, it is important to estimate and remove as much of the additive distortion as possible. CBG does this by self-calibrating the delay measurements taken from each measurement point. This is done in a distributed manner as explained in Section III. After self-calibration, CBG can more accurately transform a set of measured delays to a target into distance constraints. Self-calibration deals with the several reasons that contribute to deviate the measured delay from the theoretical minimum delay corresponding to the great-circle distance between hosts. Some of these reasons, such as circuitous routing, localized delay, and shared paths, are further discussed in Section IV-F. CBG then uses multilateration with these distance constraints to establish a geographic region that contains the target host. In our experimental results, this region always contains the target host; identifying this region is CBG's principal output. Given the target region, a reasonable "guess" as to the host's location is at the region's centroid, which is what CBG uses as a point estimate of the target's position.

In contrast to previous approaches, CBG is able to assign a confidence region to the location estimate. This allows a location-aware application to assess whether the estimate is sufficiently accurate for its needs. A location server that uses CBG may be queried by a host that wants to learn its own location as well as by a web server that desires to locate its clients, for instance. Thus, using CBG-based geolocation service, both server- and host-driven protocols are possible.

We evaluate CBG using real-life datasets and a PlanetLab [11] deployment with hosts that are geographically distributed through the continental U.S. and Western Europe. These datasets comprise 95 landmarks in the U.S. and 42 landmarks in Western Europe. Results for both datasets suggest that a certain number of landmarks, typically about 30, are needed to level off the mean error distance. Our experimental results are promising and show that CBG outperforms previous geolocation techniques. The median error distance is below 25 km for the Western Europe dataset and below 100 km for the U.S. dataset. For the majority of evaluated target hosts, the obtained confidence regions allow a resolution at the regional level. Furthermore, from the obtained results, we are also able to indicate some reasons that lead to inaccurate location estimates, including fixed delay and the sharing of paths by the measurements. Concerning the PlanetLab deployment, the median error distance is below 50 km for target hosts located in Europe and below 130 km for U.S. target hosts.

This paper is organized as follows. Section II discusses the main motivations for geolocating Internet hosts, reviews the related work on this field, and points out the contributions of CBG in contrast to previous approaches. In Section III, we introduce CBG and its methodology to use multilateration with geographic distance constraints based on delay measurements to infer the location of Internet hosts. Following that, we present results for datasets in Section IV and for PlanetLab experiments in Section V. We discuss some issues related to geolocation techniques in Section VI. Finally, we conclude and present some research perspectives in Section VII.

II. Geolocation of Internet Hosts

A. Motivation

We expect the wide availability of location information to enable the development of location-aware applications that can be useful to both private and corporate users. For example:

- *Targeted advertising on web pages:* Online consumers may have different regional preferences based on where they live. Being able to locally tailor products, marketing strategies, and contents confers a business advantage;
- *Restricted content delivery:* Following regional policies, a geographic location service can determine which client has access to content. Similarly, enforcement of localized regulation is enabled;
- *Location-based security check:* If authorized locations are known, an e-commerce transaction that is requested from elsewhere might generate warnings on atypical or unauthorized behavior of a customer.

A large range of location-aware applications may be envisaged based on an IP address to location mapping service, benefiting end users as well as network management. Furthermore, different location-aware applications may have different requirements for the accuracy of the location information. Our goal is thus to provide a methodology that is able to geolocate Internet hosts with reasonable accuracy while associating a confidence region with the given answer.

B. Related Work

An approach based on using additional DNS records to provide a geographic location service of Internet hosts is proposed by Davis *et al.* in RFC 1876 [12]. Nevertheless, the adoption of this approach has been limited since it requires changes in the DNS records and administrators have little motivation to register new location records. Tools such as [13] and [14] query Whois databases in order to obtain the location information recorded therein to infer the geographic location of a host. This information, however, may be inaccurate or stale. Moreover, if a large and geographically dispersed block of IP addresses is allocated to a single entity, the Whois databases may contain just a single entry for the entire block.

There are also some geolocation services based on an exhaustive tabulation between IP addresses ranges and their locations. This is the case of some projects [15], [16] or commercial services [17], [18]. It is hard to compare this approach with our work because the algorithms are proprietary. In any case, exhaustive tabulation is difficult to manage and to keep up to date.

Padmanabhan and Subramanian [3] investigate three different techniques to infer the geographic location of an Internet host. The first technique infers the location of a host based on the DNS name of the host or another nearby node. This DNS-based technique is the base of GeoTrack [3], GTrace [19], and SarangWorld Traceroute project [20]. Quite often network operators assign geographically meaningful names to routers, presumably for administrative convenience. For example, the name `bcr1 − so − 2 − 0 − 0.Paris.cw.net` indicates a router located in Paris, France. Nevertheless, not all names contain an indication of location. Since there is no standard, operators commonly develop their own rules for naming their routers

even if the names are geographically meaningful. Thus, parsing rules to recognize a location from a node name must be specific to each operator, imposing great challenges in the creation and management of such rules. Further, since the position of the last recognizable router in the path toward the host to be located is used to estimate the position of this host, a lack of accuracy is also expected.

The second technique splits the IP address space into clusters such that all hosts with an IP address within a cluster are likely to be co-located. Knowing the location of some hosts in the cluster and assuming they are in agreement, the technique infers the location of the entire cluster. An example of such a technique is GeoCluster [3]. This technique, however, relies on information that is partial and possibly inaccurate. The information is partial because it comprises location information for a relatively small subset of the IP address space. Moreover, such information may be inaccurate because the databases rely on data provided by users, which may be unreliable.

The third technique (GeoPing) is the closest to ours, as it is based on exploiting a possible correlation between geographic distance and network delay [3]. The location estimation of a host is based on the assumption that hosts with similar network delays to some fixed probe machines tend to be located near each other. This assumption is similar to the one exploited by wireless positioning systems such as RADAR [21] concerning the relationship between signal strength and distance. Therefore, given a set of landmarks with a well-known geographic location, the location estimate for a target host is the location of the landmark presenting the most similar delay pattern to the one observed for the target host.

In GeoPing-like methods, the number of possible location estimates is limited to the number of adopted landmarks, resulting in a discrete space of answers. As a consequence, the accuracy of this discrete space system is directly related to the number and placement of the adopted landmarks [2]. Thus, in order to increase the accuracy of techniques like GeoPing, it is necessary to add additional landmarks. In [22], a measurement-based geolocation technique with a discrete space of answers is evaluated with respect to methods for assessing the similarity among the gathered delay patterns. In Section IV-C, we compare CBG with DNS-based and GeoPing-like methods and show that CBG outperforms them.

C. Contributions

In this section, we summarize the contributions of CBG with respect to related work in geolocation of Internet hosts:

- CBG establishes a dynamic relationship between IP addresses and geographic location. This dynamic relationship results from a measurement-based approach in which landmarks cooperate in a distributed and self-calibration manner, allowing CBG to adapt itself to time-varying network conditions. This contrasts with previous work that relies on a static relationship by using queries on Whois databases, exhaustive tabulation, or unreliable information provided by users;
- A major contribution of CBG is to point out that delay measurements can be transformed to geographic distance constraints to be used in multilateration, potentially leading to more accurate location estimates of Internet hosts;
- By using multilateration with distance constraints, CBG offers a continuous space of answers instead of a discrete one as do previous measurement-based approaches;
- CBG assigns a confidence region to each location estimate, allowing location-aware applications to assess whether the location estimate has enough resolution with respect to their needs.

III. Constraint-Based Geolocation (CBG)

In this section, we describe how the CBG methodology is able to transform delay measurements to distance constraints and apply these in a multilateration process to geolocate Internet hosts. The use of distance constraints is the CBG insight to deal with the difficulty in accurately transforming delay measurements to geographic distances in the Internet.

A. Multilateration With Geographic Distance Constraints

The position of a point can be estimated using a sufficient number of distances or angle measurements to some fixed points whose positions are known. When dealing with distances, this process is called multilateration. Likewise, when dealing with angles, it is called multiangulation. Strictly speaking, triangulation refers to an angle-based position estimation process with three reference points. However, quite often the same term is adopted for any distance or angle-based position estimation. In spite of the popularity of the term triangulation, we adopt the more precise term multilateration in this paper.

The main problem that stems from using multilateration is the accurate measurement of the distances between the target point to be located and the reference points. For example, the Global Positioning System (GPS) [4] uses multilateration to some satellites to estimate the position of a given GPS receiver. In the case of GPS, the distance between the GPS receiver and a satellite is measured by timing how long it takes for a signal sent from the satellite to arrive at the GPS receiver. Precise measurement of time and time interval is at the heart of GPS accuracy. Each satellite typically has atomic clocks on board and receivers use inexpensive quartz oscillators. Therefore, in the case of GPS, multilateration is performed with "perfect" distances (i.e., with negligible errors) from time measurements and hence very accurate position estimations are feasible. In contrast to GPS, it is a challenging problem to transform Internet delay measurements to geographic distances accurately. This is likely to be the reason why direct multilateration has remained so far unexploited for the purposes of geolocating Internet hosts. Hereafter, we explain the CBG design principles that enable the multilateration with geographic distance constraints.

Consider a set $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ of K landmarks (reference hosts with a well-known geographic location). For the location of Internet hosts using multilateration, we tackle the problem of estimating the geographic distance from the target host to be located to these landmarks given the delay measurements to the landmarks. From a measurement viewpoint, the end-to-end delay over a fixed path can be split into two components: a deterministic (or fixed) delay and a stochastic delay

[23]. The deterministic delay is composed of the minimum processing time at each router, the transmission delay, and the propagation delay. This deterministic delay is fixed for any path. The stochastic delay comprises the queueing delay at the intermediate routers and the variable processing time at each router that exceeds the minimum processing time. Besides the stochastic delay, the conversion from delay measurements to geographic distance is also distorted by other sources as well. The effects of different sources of distortion on the relationship between network delay and geographic distance are further discussed in Section IV-F.

The fundamental insight for the CBG methodology is that, no matter the reason, delay is only distorted additively with respect to the time for light in fiber to pass over the great-circle path. Therefore, we are interested in benefiting from this invariant by developing a method to estimate geographic distance *constraints* from these additively distorted delay measurements. How CBG uses this insight to infer the geographic distance constraints between the landmarks and the target host from delay measurements is detailed in Section III-B. It is also shown that as a consequence of the additive delay distortion, the resulting geographic distance constraints are generally overestimated with respect to the real distances.

Fig. 1 illustrates the multilateration in CBG using the set of landmarks $\mathcal{L} = \{L_1, L_2, L_3\}$ in the presence of some additive distance distortion due to imperfect measurements. Each landmark L_i intends to infer its geographic distance constraint to a target host τ with unknown geographic location. Nevertheless, the inferred geographic distance constraint is actually given by $\hat{g}_{i\tau} = g_{i\tau} + \gamma_{i\tau}$, i.e., the real geographic distance $g_{i\tau}$ plus an additive geographic distance distortion represented by $\gamma_{i\tau}$. This purely additive distance distortion $\gamma_{i\tau}$ results from the eventual presence of some additive delay distortion. As a consequence of having additive distance distortion, the location estimation of the target host τ should lie somewhere within the gray area (cf. Fig. 1) that corresponds to the intersection of the overestimated geographic distance constraints from the landmarks to the target host.

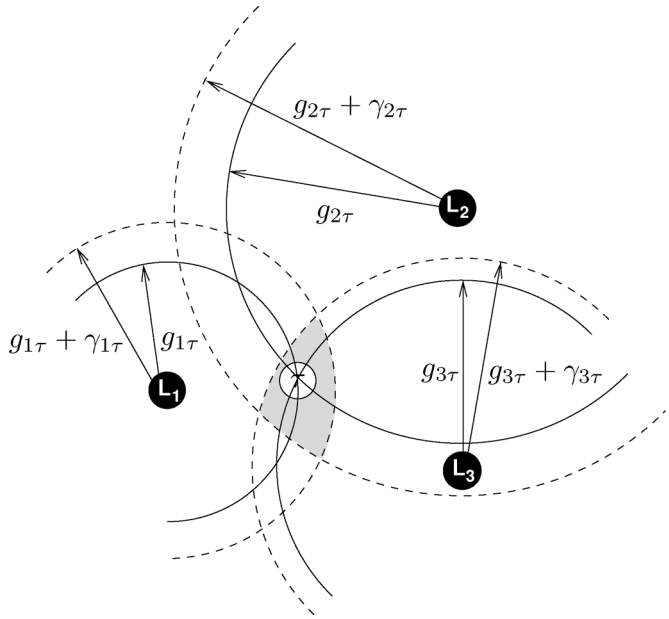


Fig. 1. Multilateration with geographic distance constraints.

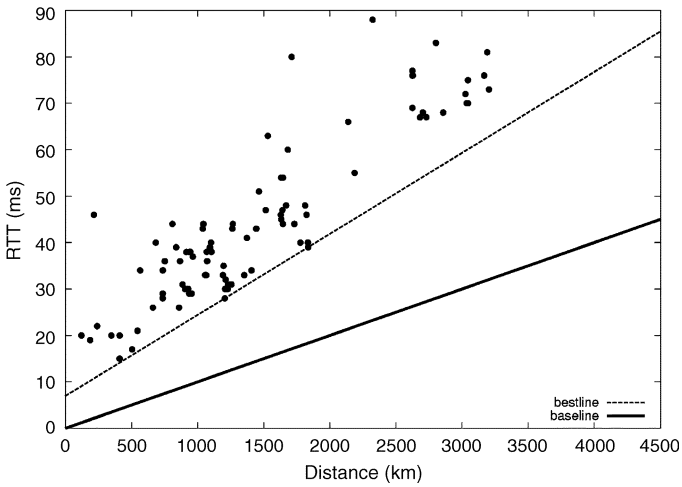


Fig. 2. Sample scatter plot of geographic distance and network delay.

B. From Delay Measurements to Distance Constraints

Before we introduce how CBG converts delay measurements to geographic distance constraints, let us first observe a sample scatter plot relating geographic distance and network delay. This sample, shown in Fig. 2, is taken from the experiments described in Section IV. The x axis is the geographic distance and the y axis is the network delay between a given landmark L_i and the remaining landmarks. Therefore, dots represent an observed relationship between geographic distance and network delay within the network as seen by landmark L_i with respect to each other landmark in the set. The meanings of "baseline" and "bestline" in Fig. 2 are explained in this section.

Recent work [2], [3] investigates the correlation coefficient found within this kind of scatter plot, deriving a least-squares fitting line to characterize the relationship between geographic distance and network delay. In contrast, we consider the *reasons* why points are scattered in the plot above, and argue that what is important is not the least-squares fit, but the tightest lower linear bound.

Based on these considerations, we propose a novel approach to establish a dynamic relationship between network delay and geographic distance. To illustrate this approach, suppose the existence of great-circle paths between the landmark L_i and each one of the remaining landmarks. Further, consider also that, when traveling on these great-circle paths, data are only subject to the propagation delay of the communication medium. In this perfect case, we should have a straight line comprising this relationship that is given by the slope-intercept form $y = mx + b$, where $b = 0$ since there are no fixed delays and m is only related to the speed bits travel in the communication medium. As already noted, digital information travels along fiber optic cables at almost exactly 2/3 the speed of light in vacuum [10]. This gives a very convenient rule of 1 ms round-trip time (RTT) per 100 km of cable. Such a relationship may be used to obtain an absolute physical lower bound on the RTT (or one-way delay) between sites whose geolocations are well known. This

lower bound is shown as the "baseline" in Fig. 2. In this idealized case, we could simply use this convenient rule to extract the accurate geographic distance between sites from delay measurements in a straightforward manner. Nevertheless, in practice, these great-circle paths rarely exist. Therefore, we have to deal with paths that deviate from this idealized model for several reasons, including queueing delay and lack of great-circle paths between hosts.

As stated in Section III-A, the main insight behind CBG is that the combination of different sources of delay distortion with respect to the perfect great-circle case only produces an additive delay to the theoretical minimum delay associated with the great-circle distance. We thus model the relationship between network delay and geographic distance using delay measurements in the following way. We define the "bestline" for a given landmark L_i as the line $y = m_i x + b_i$ that is closest to, but below, all data points (x, y) and has a non-negative intercept, since it makes no sense to consider negative delays. A positive intercept b_i in the bestline reflects the presence of some fixed delay. Note that each landmark computes its own bestline with respect to all other landmarks. Therefore, the bestline can be seen as the line that captures the least distorted relationship between geographic distance and network delay from the viewpoint of each landmark. The distance of each data point to the bestline corresponds to the presence of some source of extra additive distortion with respect to the best-observed case, i.e., the bestline. The region separating the bestline and the baseline (cf. Fig. 2) represents the observed gap between the current relationship of geographic distances and network delays within the network and the idealized case.

The finding of the bestline is formulated as a linear programming problem. For a given landmark L_i, there are the network delay d_{ij} and the geographic distance g_{ij} toward each landmark L_j, where $i \neq j$. We need to find for each landmark L_i the slope m_i and the intercept b_i that determine the bestline given by the slope-intercept form $y = m_i x + b_i$. The condition that the bestline for each landmark L_i should lie below all data points (x, y) defines the feasible region where a solution should lie:

$$y - m_i x - b_i \geq 0, \quad \forall i \neq j \tag{1}$$

The objective function to minimize the distance between the line with non-negative intercept and all the delay measurements is stated as

$$\min_{\substack{b_i \geq 0 \\ m_i \geq m}} \left(\sum_{i \neq j} y - m_i x - b_i \right) \tag{2}$$

where m is the slope of the baseline. We use (2) to find the solution m_i and b_i from (1) that determines the bestline for each landmark L_i. Each landmark L_i then uses its own bestline to convert the delay measurement to the target host into a geographic distance. Thus, the estimated geographic distance constraint $\hat{g}_{i\tau}$ between a landmark L_i and the target host τ is derived from the delay distance $d_{i\tau}$ using the bestline of the landmark L_i as follows:

$$\hat{g}_{i\tau} = \frac{d_{i\tau} - b_i}{m_i}. \tag{3}$$

If delays between landmarks are periodically gathered, this leads to a *self-calibrating* algorithm that determines how each landmark currently observes the dynamic relationship between network delay and geographic distance within the network.

C. Using Distributed Distance Constraints to Geolocate Hosts

CBG uses a geometric approach using multilateration to estimate the location of a given target host τ. Each landmark L_i infers its geographic distance constraint to the target host τ, which is actually the additively distorted distance $\hat{g}_{i\tau} = g_{i\tau} + \gamma_{i\tau}$, using (3). Therefore, each landmark L_i estimates that the target host τ is somewhere within the circumference of a circle $\mathcal{C}_{i\tau}$ centered at the landmark L_i with a radius equal to the estimated geographic distance constraint $\hat{g}_{i\tau}$ (similar to the example of Fig. 1). Given K landmarks, the target host τ has a collection of closed curves $\mathbf{C}_\tau = \{\mathcal{C}_{1\tau}, \mathcal{C}_{2\tau}, \ldots, \mathcal{C}_{K\tau}\}$ that can be seen as an order-K Venn diagram. Out of the possible 2^K regions defined by this order-K Venn diagram for the target host τ, we are interested in the unique region $\mathcal{R}$ that forms the intersection of all closed curves $\mathcal{C}_{i\tau} \in \mathbf{C}_\tau$ given by

$$\mathcal{R} = \bigcap_{i}^{K} \mathcal{C}_{i\tau}. \tag{4}$$

The region $\mathcal{R}$ corresponds to the gray area of Fig. 1 that hopefully comprises the real position of the target host τ. Note that $\mathcal{R}$ is convex, since the regions $\mathcal{C}_{i\tau}$ are convex, and the intersection of convex sets is itself convex. The conversion from the additively distorted delay measurements to geographic distance constraints will overestimate these distance constraints. The goal is to assure that since each landmark overestimates its geographic distance constraint toward the target host, there will be a region $\mathcal{R}$ determined by the intersection of all the curves with an overestimated radius. If the baseline were used for this conversion, the geographic distances would be strongly overestimated based on the delay measurements because these measurements are taken in a non-idealized case. This would potentially create a very large intersection region $\mathcal{R}$ for a given target host that would provide an inaccurate location estimation for this target host. In contrast, the bestline captures the best relationship between network delay and geographic distance as currently observed within the network. Therefore, the idea behind using the bestline is to minimize the overestimation of the geographic distances by taking into account the current network conditions as constraints. Using a certain number of landmarks intends to introduce some diversity into the bestline computation so that the bestline represents the best observed case for a set of different reference points given network conditions.

D. Effects of Over and Underestimation of Distance Constraints

When establishing the set of closed curves $\mathbf{C}_\tau$ for a given target host τ, there are three possible resulting situations: 1) the geographic distance constraints from all landmarks are overestimated; 2) the geographic distance constraints from all landmarks are underestimated; 3) the geographic distance constrains are overestimated for some landmarks and underestimated for

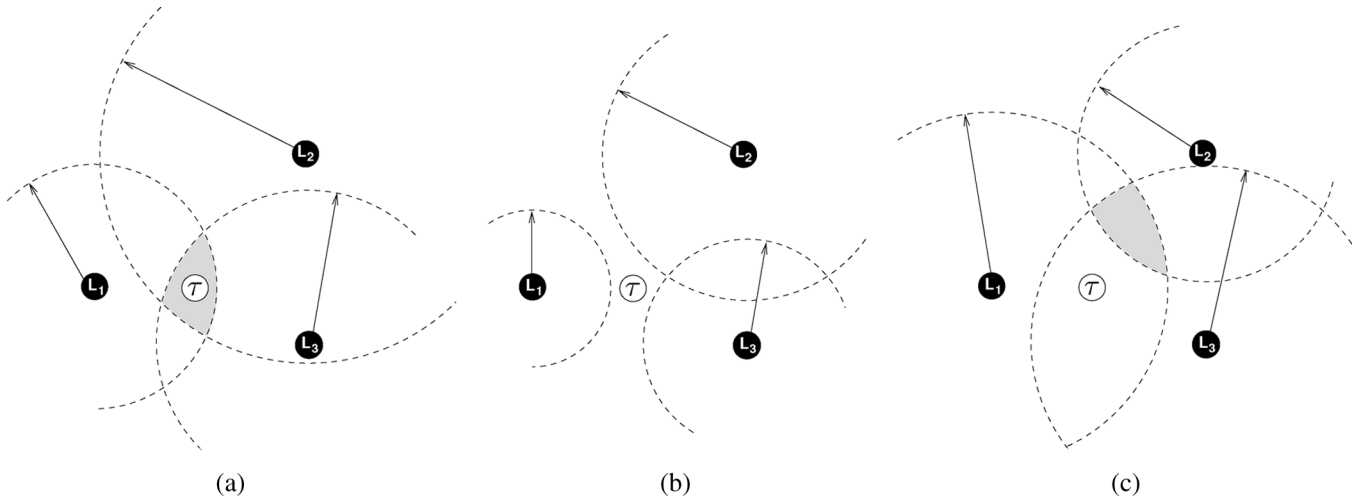


Fig. 3. Effects of the over and underestimation of the geographic distance constraints. (a) Overestimated distance constraints. (b) Underestimated distance constraints. (c) Mismatch.

the remaining landmarks, leading to a mismatch among landmarks. Fig. 3 depicts these three situations.

In Fig. 3(a), geographic distance constraints are overestimated. As a consequence, CBG can determine an intersection region $\mathcal{R}$ and use it to infer the location of the target host τ. We expect that this is the only likely situation to occur if a sufficient number of landmarks is used. Experimental results presented in Sections IV-B and V indeed confirm that distance constraints are overestimated for all considered target hosts.

If the geographic distance constraints to the target host τ from all landmarks are underestimated, as shown in Fig. 3(b), the region $\mathcal{R}$ is empty, i.e., there is no intersection region at all. This situation happens only if the target host presents, from the viewpoint of the landmarks, a smaller ratio of geographic distance to network delay than the one represented by the bestline, i.e., smaller than the one from *all* landmarks. This is clearly unlikely. In this case, based on the bestline approach, CBG will not find sufficient information to infer a location estimation. As a consequence, CBG declares that a location estimation is not possible for this specific target host τ. This is rather an important property of CBG because for several applications no location estimation at all may be better than blindly providing a geolocation estimate of the target host as other techniques would do.

In Fig. 3(c), we illustrate a situation where two landmarks, L_1 and L_3, overestimate their geographic distance constraints to the target host τ while the landmark L_2 underestimates its distance constraint. The mismatch in the distance constraints among the landmarks results in an intersection region that does not include the target host τ. This would defeat our methodology because the location estimation would be inferred as being inside the intersection region, away from the real position of the target host. We currently do not handle mismatches and this is left for future work, although we expect this mismatch situation to be unlikely. First, consider two groups of landmarks: one whose members overestimate their geographic distance constraints to the target host and another group wherein this distance constraint is underestimated. The mismatch situation happens when the observed relationship between geographic distance and network delay from these two groups toward the target host is very unbalanced. Although we know that routing asymmetry (and, as a consequence, capacity asymmetry) is somewhat usual in the Internet, we believe that the differences in capacity are unlikely to be enough to result in the mismatch situation. Moreover, the self-calibrating nature of the CBG method incorporates in the construction of each bestline the current network condition as seen by the whole set of landmarks. Therefore, each landmark has an unilateral viewpoint to the remaining landmarks, thus incorporating eventual asymmetries in the network conditions.

In summary, the CBG's method of transforming delay measurements to distance constraints is a constrained distance overestimation. This constrained overestimation results in an intersection region, whereby CBG estimates the location of the target host. In the case that a target host presents underestimated geographic distance constraints to the landmarks, CBG is able to detect this situation and then decline to provide a location estimation. The self-calibrating nature of CBG elegantly avoids a mismatch situation where the system would be defeated. We indeed confirm that the geographic distance constraints are overestimated in all our experiments (see Sections IV and V) and that a consistent location estimation has been always feasible in these experiments.

IV. Experimental Results Using Datasets

A. Datasets

To validate our methods, we need datasets with hosts whose geographic locations are well known. Unfortunately, datasets that provide the geolocation of the involved hosts are uncommon. For our experiments, we then use two datasets:

- RIPE: data collected in the Test Traffic Measurements (TTM) project of the RIPE network [24]. Each RIPE host generates nearly 300 kB per day toward every other RIPE host with an average of two packets sent per minute. The dataset we consider is composed by the 2.5 percentile of

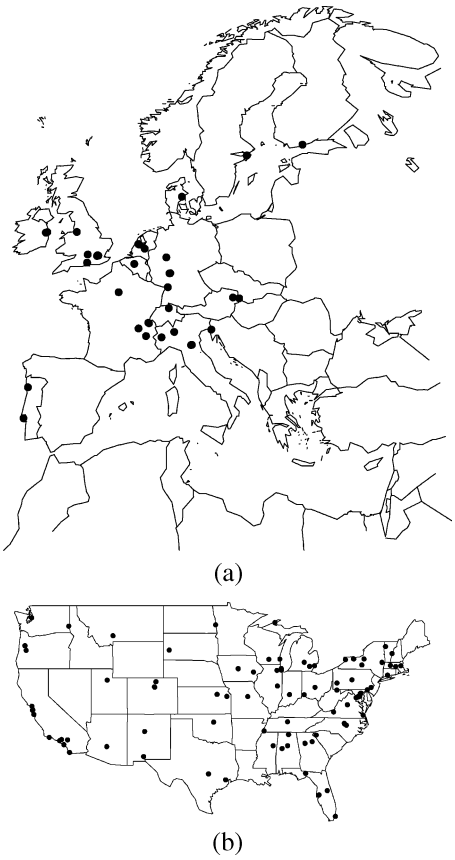


Fig. 4. Geographic location of landmarks (not to the same scale). (a) 42 landmarks in Western Europe (RIPE dataset). (b) 95 landmarks in the continental U.S. (AMP dataset).

the one-way delay observed from each RIPE host to each other host in the set during a period of 10 weeks from early December 2002 until February 2003. Most RIPE hosts are located in Europe and they are all equipped with GPS cards, thus allowing their exact geographic position to be known. We then use the 42 RIPE hosts located in Western Europe (W.E.) to compose our W.E. landmark dataset. Fig. 4(a) shows the geographic distribution of the W.E. dataset.

- NLANR AMP: data collected in the NLANR Active Measurement Project (AMP) [25]. Delay is sampled on average once a minute. This leads to an average measurement load of about 144 kB per day sent by each AMP host toward each other AMP host. The dataset we consider is composed by the 2.5 percentile of the RTT delay between all the participating nodes located in the continental United States (U.S.), in a total of 95 hosts. This data was collected on January 30, 2003 and is symmetric. The exact location of each participating node (in pairs of latitude and longitude) is also available. These 95 AMP hosts compose our U.S. landmark dataset. Their geographic distribution is illustrated in Fig. 4(b).

The minimum RTT taken on the measurements is used since it is more likely to be reflective of actual propagation delay and may filter out some of the effects of queueing and local delay. We note that no correctly-measured RTT can be an underestimate of the minimum RTT. Nevertheless, we indeed consider that some RTTs may be erroneously measured. We thus use the 2.5 percentile to avoid erroneous under-measurements of the minimum RTT.

The experimental datasets comprise hosts in United States and Western Europe. The main reason for this restriction is that the datasets we have had correspond to hosts located in these regions. Nevertheless, we indeed believe that the results we report in this paper are interesting and promising in spite of being limited to the U.S. and Western Europe.

Using the gathered delays in each dataset, we construct two delay matrices $\mathbf{D}_{\mathrm{ripe}}$ and $\mathbf{D}_{\mathrm{amp}}$ with dimensions (42×42) and (95×95), respectively. We consider all hosts in each dataset as landmarks, leading to two sets of landmarks: $\mathcal{L}_{\mathrm{ripe}} = \{L_1, L_2, \ldots, L_{42}\}$ and $\mathcal{L}_{\mathrm{amp}} = \{L_1, L_2, \ldots, L_{95}\}$. We then find the set of bestlines, as described in Section III-B, for each element belonging to each landmark dataset $\mathcal{L}_{\mathrm{ripe}}$ and $\mathcal{L}_{\mathrm{amp}}$. The bestline computation for each landmark is done considering only landmarks of the same dataset. The set of bestlines is determined by a slope vector $\mathbf{m} = [m_1, m_2, \ldots, m_i]^T$ and an intercept vector $\mathbf{b} = [b_1, b_2, \ldots, b_i]^T$ for each landmark dataset. After computing the bestline for each landmark in the landmark dataset, delays in each dataset are converted to geographic distance constraints applying (3). This results in two geographic distance constraint matrices $\mathbf{G}_{\mathrm{ripe}}$ and $\mathbf{G}_{\mathrm{amp}}$. These matrices comprise the additively distorted geographic distances between the landmarks that we use in our experiments for performance evaluation.

In our experiments, we geolocate each host one at a time using CBG. The remaining hosts in the same dataset are then considered as landmarks to perform the location estimation of a target host. The bestline of each landmark is computed using the set of landmarks of each scenario, thus excluding the target host. We stress that when we calculate the bestline for a particular experiment in which we are geolocating a given target host, we do not include this target host in the bestline calculation. We repeat this procedure to evaluate the resulting location estimation of each host in both U.S. and W.E. landmark datasets.

B. Location Estimation of a Target Host

From the geographic distance constraints in matrices $\mathbf{G}_{\mathrm{ripe}}$ and $\mathbf{G}_{\mathrm{amp}}$, CBG determines for each target host τ a set of closed curves $\mathbf{C}_\tau = \{\mathcal{C}_{1\tau}, \mathcal{C}_{2\tau}, \ldots, \mathcal{C}_{K\tau}\}$ (see Section III-C), where $\mathrm{K} = 42$ for the W.E. dataset and $\mathrm{K} = 95$ for the U.S. dataset. Each curve in $\mathbf{C}_\tau$ is centered at its respective landmark L_i and has as radius the estimated geographic distance constraint $\hat{g}_{i\tau}$. To illustrated the CBG methodology, Fig. 5 shows two example sets of closed curves extracted from our experimental study. Fig. 5(a) refers to the location estimation of a RIPE host in Brussels, Belgium. There are 41 curves corresponding to the viewpoints of the remaining landmarks in the W.E. landmark dataset. Similarly, Fig. 5(b) presents the set of 94 closed curves used to estimate the location of an AMP host located in Lawrence, Kansas, USA.

The gray areas in Fig. 5(a) and (b) represent the respective regions $\mathcal{R}$, i.e., the intersection of all closed curves in each case.

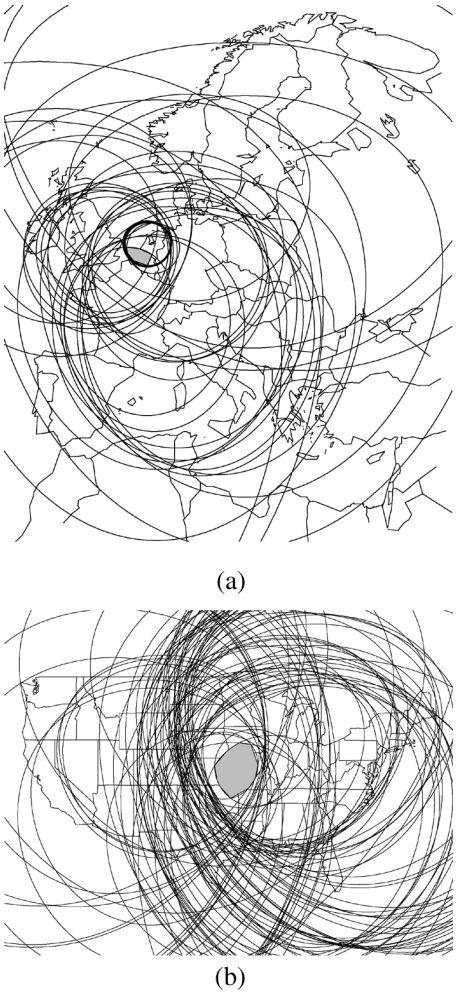


Fig. 5. Two location estimation examples (not to the same scale). (a) RIPE host in Brussels, Belgium. (b) AMP host in Kansas, U.S.

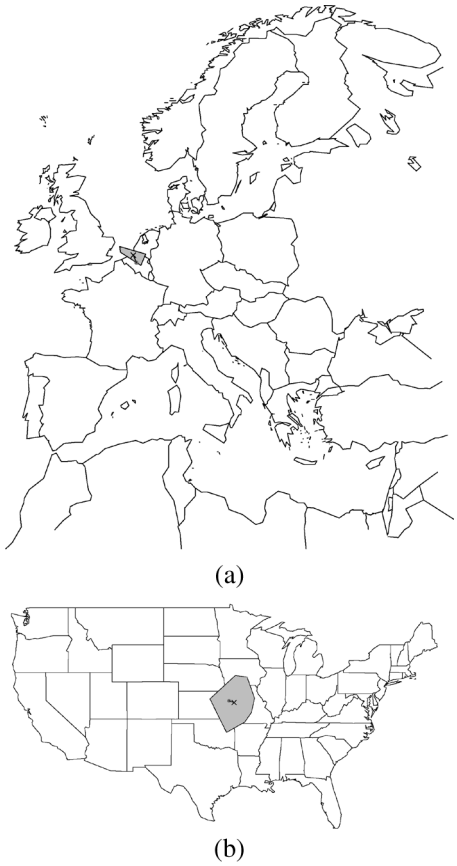


Fig. 6. Sample result from the polygon heuristic (not to the same scale). (a) Locating the RIPE host in Brussels, Belgium. (b) Locating the AMP host in Kansas, U.S.

In our experiments, we take all hosts in the datasets and use them one at a time to be target hosts. It is important to point out that for all the target hosts in both landmark datasets, there is always a region $\mathcal{R}$ that contains the target host. This means that CBG successfully overestimates the geographic distance constraints for all target hosts. Such a result verifies that the situation of Fig. 3(a) is indeed prevalent, at least in our experimental datasets, as postulated in Section III-D.

The area of the intersection region $\mathcal{R}$, i.e., the gray areas in Fig. 5(a) and (b), indicates the confidence region that CBG associates with each location estimate. Note that in most cases confidence regions have a relatively small area, not visible in similar plots with all closed curves (Sections IV-D and V present results on the sizes of confidence regions in our experiments). These two examples have larger confidence regions than are typical, but are chosen so that the region is sufficiently visible so as to illustrate the CBG methodology.

C. Geolocating Internet Hosts

The region $\mathcal{R}$ is the location estimate of CBG. Given this region, a reasonable "guess" as to the target host's location is at the region's centroid. Therefore, CBG uses the centroid of region $\mathcal{R}$ as a point estimate of the target's position.

We adopt the following heuristic to approximate the intersection region $\mathcal{R}$ by a polygon. The resulting polygon is used to approximately measure the area of the region $\mathcal{R}$ and provide an estimate of the point location of the target host. To form the polygon, we consider as vertices the crossing points of the circles $\mathcal{C}_{i\tau}$ that belong to all circles. Since the region $\mathcal{R}$ is convex, the polygon is an underestimate of the area of $\mathcal{R}$. For example, in Fig. 1, the vertices would be the crossing points of the dashed lines that touch the gray area, thus determining a polygon that approximates this area. Thus, we approximate the region $\mathcal{R}$ by a polygon made up of line segments between N vertices $v_n = (x_n, y_n)$, $0 \leq n \leq N-1$. The last vertex $v_N = (x_N, y_N)$ is assumed to be the same as the first, i.e., the polygon is closed. The area of a non-self-intersecting polygon with vertices $v_0 = (x_0, y_0), \ldots, v_{N-1} = (x_{N-1}, y_{N-1})$ is given by

$$A = \frac{1}{2} \sum_{n=0}^{N-1} \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix} \tag{5}$$

where $|\mathbf{M}|$ denotes the determinant of matrix $\mathbf{M}$. The centroid c of the polygon, i.e., the position estimate of the target host τ, is positioned at (c_x, c_y) given by

$$c_x = \frac{1}{6A} \sum_{n=0}^{N-1} (x_n + x_{n+1}) \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix} \tag{6}$$

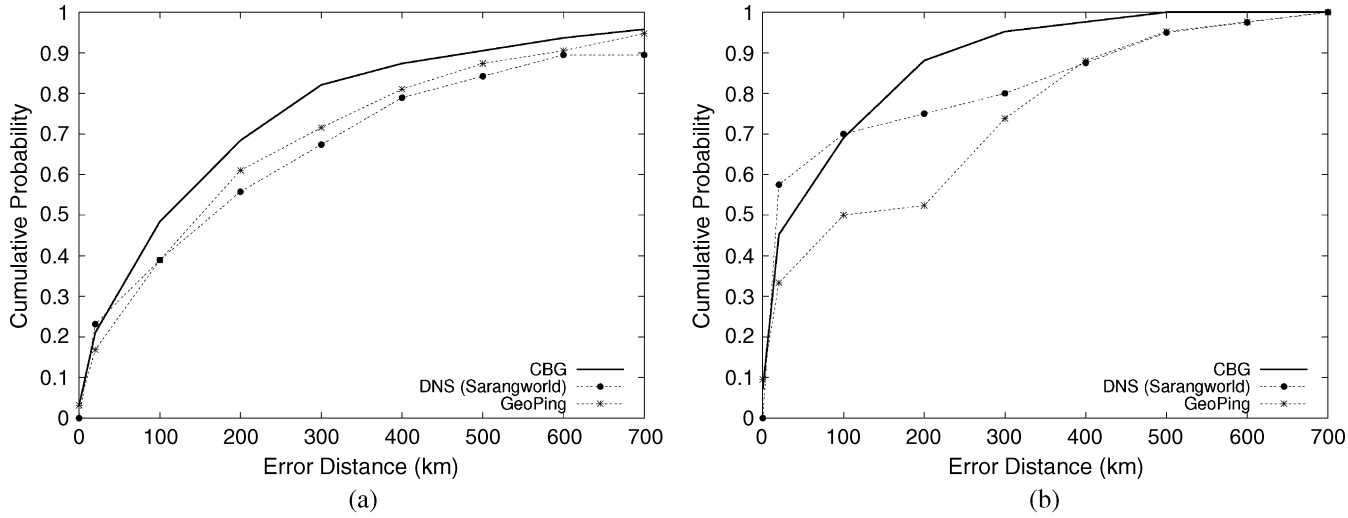


Fig. 7. Error distance for CBG, DNS-based, and GeoPing-like methods. (a) U.S. dataset. (b) Western Europe dataset.

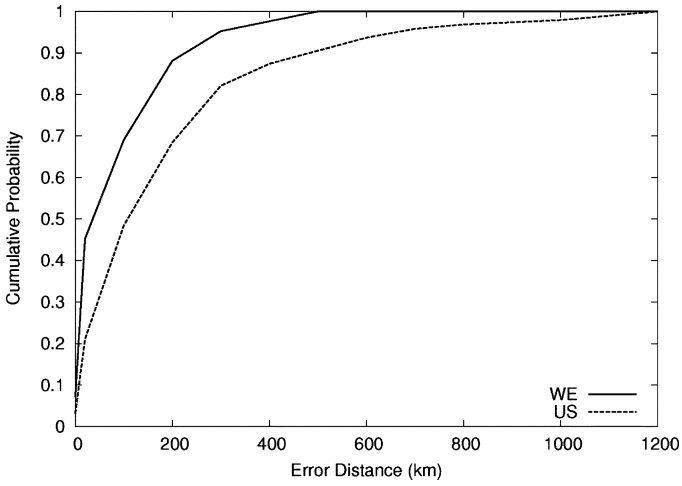


Fig. 8. Error distance for CBG in the U.S. and W.E. datasets.

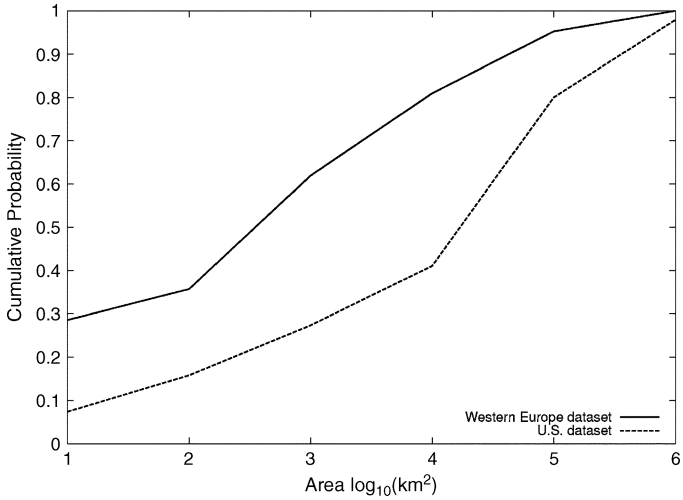


Fig. 9. Confidence regions provided by CBG in km^2.

and

$$c_y = \frac{1}{6A}\sum_{n=0}^{N-1}(y_n + y_{n+1})\begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix}. \qquad (7)$$

The point estimate of the target host and the estimate of the confidence region are the centroid (c_x, c_y) and the area A of the approximated polygon, respectively. Fig. 6 shows two sample polygons provided by this heuristic. The gray areas presented in Fig. 6 are the resulting polygon approximations of intersection regions shown in Fig. 5. Solid circles indicate the real location of each target host while crosses indicate the point estimate provided by the centroid of the polygon.

After inferring the point estimate for each considered target host, we compute the error distance, which is the difference between the estimated position and the real location of the target host τ. We compare our performance with the results obtained by a DNS-based method and by a GeoPing-like measurement-based geolocation system. The DNS-based method (i.e., SarangWorld Traceroute project [20]) performs traceroutes toward the target host and it infers the geolocation of intermediate routers from their DNS names. The inferred geolocation of the closest recognizable router with respect to the target host is used as a location estimate. The GeoPing-like method uses a measurement-based approach with a discrete space of answers [2], [3], i.e., where the location of the landmarks are used as location estimates.

Fig. 7 shows the cumulative distribution function (CDF) of the observed error distance using CBG, the DNS-based method, and the GeoPing-like approach with a discrete set of answers. CBG outperforms the DNS-based approach as well as the GeoPing-like method. The performance gap between the two measurement-based approaches is more significant in the Western Europe dataset. This is probably because this dataset presents fewer landmarks than the U.S. dataset. In the discrete space approach, since the number of possible answer is limited to the locations of the landmarks, the number and placement of landmarks is a key point to the performance [2]. In Section IV-E, we investigate the impact of the number of adopted landmarks on the performance of CBG.

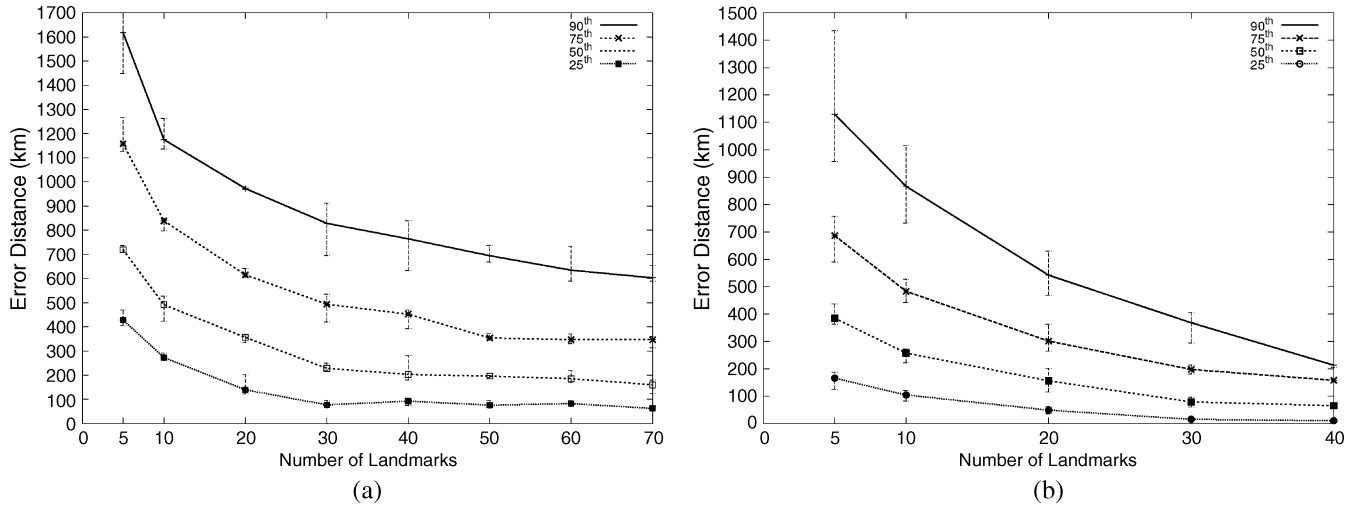


Fig. 10. Error distance as a function of the number of landmarks. (a) U.S. dataset. (b) Western Europe dataset.

In Fig. 8, we compare further the CBG results in error distance for the U.S. and W.E. datasets. The mean error distance in the U.S. dataset is 182 km, whereas for the W.E. dataset the mean error distance is 78 km. Most hosts in both landmark datasets have a quite good location estimation. The median error distance and the 80th percentile for the U.S. dataset are 95 km and 277 km, respectively. In the W.E. dataset, the median error distance is 22 km and the 80th percentile is 134 km. We identify and discuss reasons of inaccurate estimations in further detail in Section IV-F.

D. Confidence Region of a Location Estimation

The total area of the intersection region $\mathcal{R}$ is somewhat related to the confidence that CBG assigns to the resulting location estimate. Intuitively, this area quantifies the geographic extent or spread of each location estimate in km^2. The smaller the area of region $\mathcal{R}$, the more confident CBG is in this location estimate. Therefore, in contrast to previous measurement-based geolocation techniques, CBG assigns a confidence region in km^2 to each location estimate. We believe this is important because this confidence region may be used by location-aware applications to evaluate to which extent they can rely on the given location estimate. Furthermore, we envisage location-aware applications with different requirements on accuracy. By using the confidence region, these location-aware applications may decide if the provided location estimate has sufficient resolution with respect to their particular needs.

Fig. 9 presents the CDF of confidence regions in km^2 for location estimates in both the U.S. and W.E. landmark datasets. Results show that, for the U.S. dataset, CBG assigns a confidence region with a total area less than 10^5 km^2 for around 80% of location estimates. This area is slightly larger than Portugal or the U.S. state of Indiana. For the W.E. dataset, 80% of location estimates have a confidence region of up to 10^4 km^2, thus enabling regional host location. A confidence region of less than 10^3 km^2, which is equivalent to a large metropolitan area, is achieved by 25% of target hosts for the U.S. dataset and by 65% of target hosts for the W.E. dataset.

E. Impact of the Number of Landmarks

In this section, we evaluate the impact of the number of adopted landmarks in the performance of CBG. For each dataset, we compute the mean error distance as the average of all error distances corresponding to several random sets of k landmarks chosen out of the total number of available landmarks (42 for the W.E. dataset and 95 for the U.S. dataset). Because the number of possible placement combinations become very large as we increase k, we do not consider all the possible choices of k landmarks out of each dataset.

Fig. 10 shows different percentile levels of the error distance of CBG location estimates as a function of the number of adopted landmarks. For example, the 90th percentile curve represents the error distance at which the CDF plot of mean error distance meets the 0.90 probability mark. Error bars indicate the 99% confidence interval. These results suggest that a certain number of landmarks, typically about 30, is needed to level off the mean error distance for both datasets.

F. On the Reasons of Inaccurate Estimations

Two aspects contribute to add basic robustness to the location inference performed by CBG against factors that may weaken the relationship between network delay and geographic distance. First, delay is measured from multiple geographically distributed landmarks rather than from three locations as would be sufficient for a triangulation with "perfectly" accurate measurements like in GPS. Second, the minimum RTT, among several RTT samples, is considered rather than an individual delay sample to avoid considering queueing delay. Besides these two sources of distortion, the conversion from delay measurements to geographic distance constraints may be also distorted by other sources as well and these are discussed in the following.

1) Circuitous Routing: Route circuitousness indicates the degree to which the network path deviates from the great-circle path between two nodes. Subramanian *et al.* [6] examine how circuitous Internet paths are and show that the level of network connectivity and the interconnection policies between autonomous systems directly impact the circuitousness of a path.

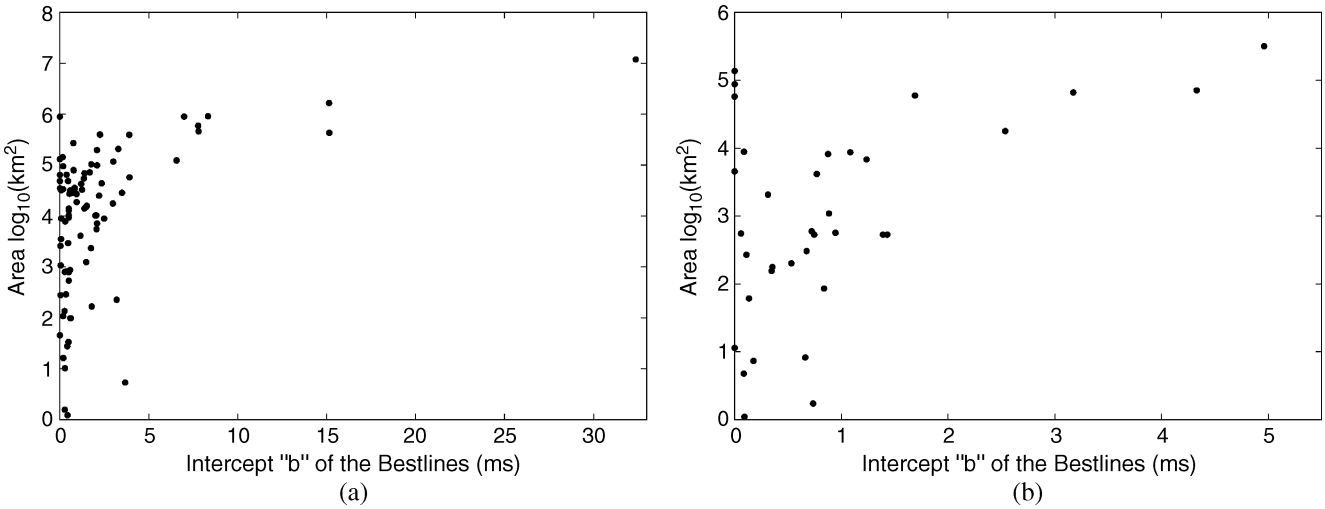


Fig. 11. Confidence region as a function of the intercept b (localized delay). (a) U.S. dataset. (b) Western Europe dataset.

Further, at the network level, Internet paths are not necessarily optimal since end-to-end paths can be significantly longer than needed. This phenomenon has been recently analyzed under different names, such as path inflation [26] or routing stretch [27], and also contributes to path circuitousness.

CBG deals with these deviations from the idealized great-circle paths between hosts. This is done as each landmark self-calibrates its vision to the relationship between network delay and geographic distance when computing the bestline. The bestline at each landmark reflects the known path that is the closest to the great-circle path (represented by the baseline). Therefore, the bestline incorporates the deviations from the great-circle path as they are seen with respect to all other landmarks.

2) Localized Delay: Localized delay refers to the situation in which there is a constant amount of delay that appears to be added to all delay measurements to a given host. Localized delays may emerge from low-speed access links, local congestion, or both. In CBG, localized delay is represented by the intercept b_i of the computed bestlines. In other words, the target sees landmarks as having a nonzero minimum delay even for landmarks that are collocated with the target. The presence of excessive localized delays is misleading because the geographic distance constraints tend to be largely overestimated, leading to large confidence regions.

Fig. 11 compares the intercept b_i found in the bestline on each landmark L_i and the resulting confidence region when this landmark is used as a target host. It should be noted that Fig. 11(a) and (b) are not in the same scale. The U.S. dataset presents some landmarks with very large intercepts in their bestlines as compared to the European landmarks, leading to large confidence regions for some U.S. target hosts. However, regardless of the dataset, all landmarks that have large intercepts b_i also have a large confidence region when being used as target hosts. This clearly indicates that excessively large localized delays lead to large confidence regions. Nevertheless, the contrary is not necessarily true. From Fig. 11, small intercepts do not directly result in small confidence regions. A large confidence region may be the result of an overestimation of the distance constraints by the remaining landmarks due to how they currently observe the network conditions, and not necessarily related to local conditions of the target host. If shared paths hide the target host behind a single point, all landmarks overestimate the distance constraints, even if the target host presents no localized delay as is further discussed in next section.

3) Shared Paths: Measurements from different landmarks that share some paths toward the target host provide redundant information. If all measurements travel past a single point and share the remaining paths toward the target host, the location estimate is limited to a region around that single point. This potentially leads to inaccurate estimates, i.e., large confidence regions. We observe some inaccurate location estimates due to shared paths in our experiments, as some cases shown in Fig. 11 that have large confidence regions although the host presents small or no localized delay.

An interesting example of shared paths is the case of the RIPE hosts located in Lisbon and Porto, both cities in Portugal. When the Porto landmark is used as a target host, this leads to an inaccurate location estimation with a confidence region of about 57 000 km^2, which is about 2/3 of the size of Portugal. Fig. 12 shows the bestline that reflects how the Lisbon and Porto landmarks best observe the relationship between network delay and geographic distance within the network. It should be noted that the Porto landmark determines the bestline of the Lisbon landmark in Fig. 12(a), and *vice versa* in Fig. 12(b). We observe that without the Lisbon landmark in Fig. 12(b) the bestline of the Porto landmark would be shifted toward the remaining landmarks. The resulting figure would be virtually the same as of the bestline of the Lisbon landmark in Fig. 12(a), except that an intercept b_i of about 5 ms would be present in the "new" bestline of the Porto landmark. The measured delay between the Porto landmark and the Lisbon landmark is indeed about 5 ms. In other words, the network perception that all landmarks have from the Porto host is the same that they have from the Lisbon host with an additional delay of 5 ms. Clearly, from the viewpoint of the remaining landmarks, the Porto landmark is to some extent hidden behind the Lisbon landmark. We suggest that this is an indication that all traffic from Porto toward the remaining landmarks, and vice versa, travels through the Lisbon

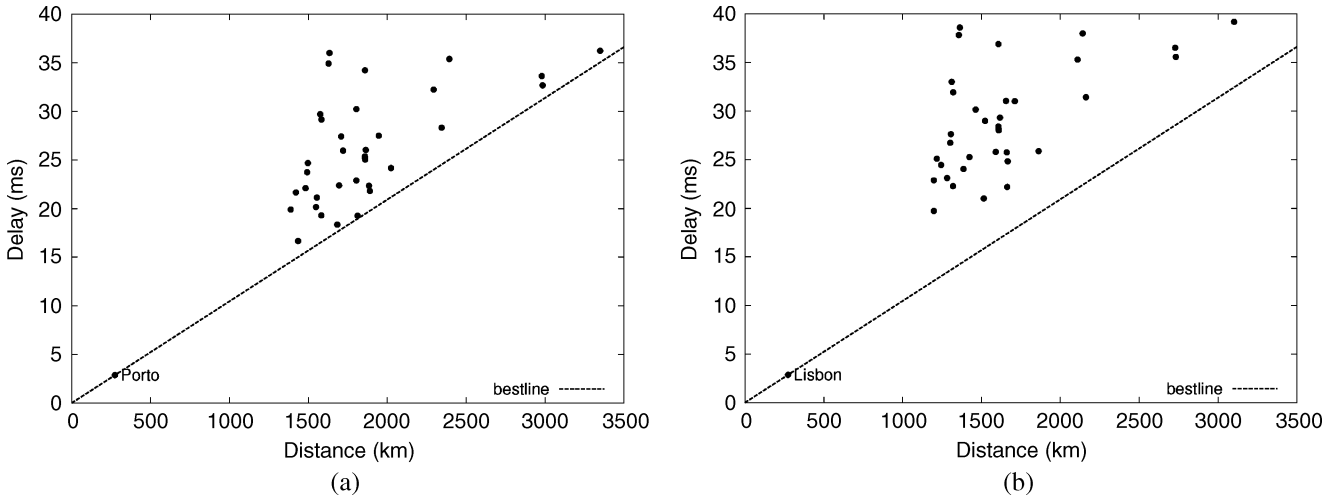


Fig. 12. Example of inaccurate location estimation caused by shared paths. (a) Bestline of the Lisbon landmark. (b) Bestline of the Porto landmark.

urban area. As a consequence, when the Porto landmark is used as the target host, the confidence region is inferred as a relatively large circle around Lisbon, i.e., an inaccurate location estimate.

In the U.S. dataset, we observe a similar typical case of shared paths that leads to inaccurate location estimations. The AMP hosts `amp − wsu` and `amp − montana`, respectively located in Pullman (Washington—WA) and in Bozeman (Montana—MT), seem to be hidden by the `amp − uwashington` host in Seattle (WA). All the remaining landmarks in the U.S. dataset see the `amp − wsu` and `amp − montana` hosts with a constant extra delay of 10 ms and 15 ms, respectively, added to their visions of `amp − uwashington`. This leads to inaccurate confidence regions. Measurements from all other landmarks share paths to `amp − wsu` and `amp − montana` after traveling through the Seattle area as indicate the respective `traceroute` traces available at AMP [25]. It is reasonable to suppose that the traffic to these hosts passes through somewhere in the Seattle area. We believe that these results on shared paths obtained using CBG are an indication that similar methods may be used for topology inference, but this still needs further investigation.

V. Experimental Results Using PlanetLab

We also present experimental results for a CBG deployment on PlanetLab [11]. These results have been taken in early May 2005. We adopt 57 landmarks, i.e., PlanetLab nodes, distributed in the following way: 24 in the U.S., 24 in Europe, five in Asia, three in South America, and one in Oceania. These landmarks are used to geolocate using CBG methodology 42 target hosts in the U.S. and 43 target hosts in Europe. Among these target hosts, there are three behind modems, three connected through wireless links, and six through ADSL, while the remaining have Internet access using broadband links. The response time for all target hosts is within a 2–3 minutes range.

Fig. 13 shows the CDF of the observed error distance using CBG in our PlanetLab experiment. The mean error distance for the target hosts located in the U.S. is 209 km, whereas for the target hosts located in Europe the mean error distance is 106 km. The median error distance and the 80th percentile for the U.S. hosts are 130 km and 411 km, respectively. For the target hosts located in Europe, the median error distance is 42 km and the 80th percentile is 218 km. Fig. 14 presents the CDF of the confidence regions in km^2 for the location estimates of the target hosts located in both the U.S. and Europe. Although confidence regions in the PlanetLab experiments are in general larger than those found in the dataset evaluation, there is a larger number of highly confident estimates, i.e., with a confidence region of less than 10^2 km.

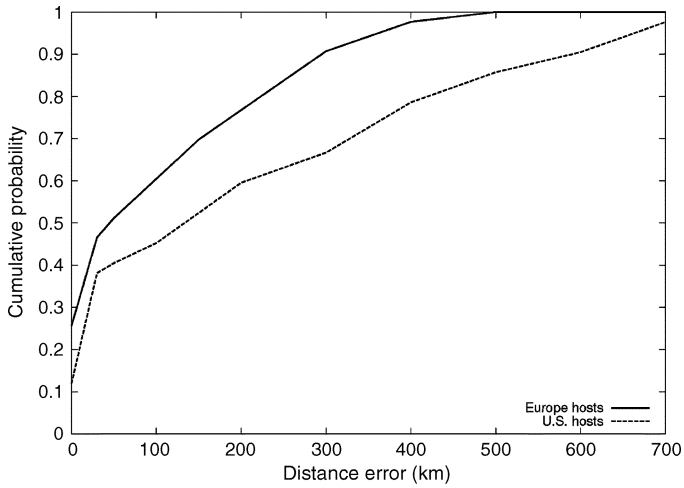


Fig. 13. Error distance for CBG using PlanetLab.

The network diversity issues associated with PlanetLab experiments are well-known [5], [28] and as such, our results must be evaluated in that light. However, there is no widely available alternative system to PlanetLab for these sorts of experiments at the current time.

VI. Discussion

In this section we address topics related to Internet geolocation technology in general. We emphasize that the raised issues do not necessarily affect CBG more than they do with any other geolocation technique.

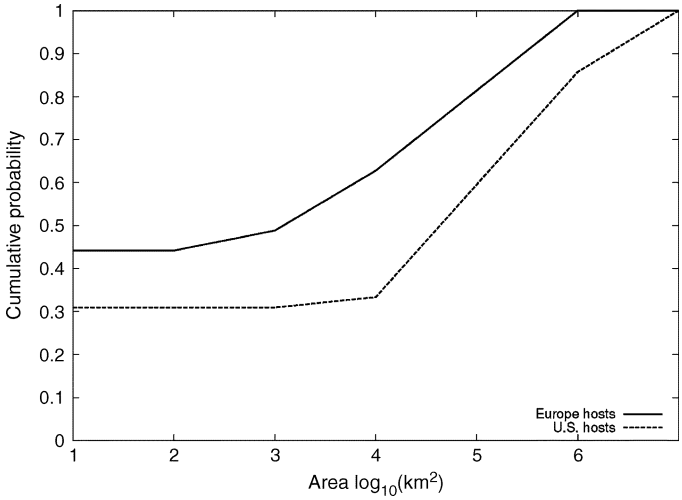


Fig. 14. Confidence regions provided by CBG in km^2 using PlanetLab.

The development and use of geolocation technology can give rise to privacy and security concerns. The Geographic Location/Privacy (geopriv) IETF working group [29] focuses on establishing policies to control the exchange of geolocation information with privacy in mind, whereas the development of geolocation technology is out of its scope of work. Thus, our research is actually complementary to their work. We believe that any geolocation technology, including CBG, has to consider privacy and security issues in the use of the provided location information. Further, the proposed approach at the geopriv community is to provide less location information, i.e., with reduced resolution, to unprivileged users. The confidence region assigned by CBG to each location estimate may be directly used to this purpose.

Proxies and firewalls impose a fundamental limitation on measurement-based geolocation techniques that depend on the client IP address. Since the IP address seen by the external network may actually correspond to the address of a proxy, the geolocation techniques infer the geographic location of the proxy, which may be inaccurate in the case the client and the proxy are not in relatively close proximity. As a practical countermeasure to this, commercial geolocation services that rely on exhaustive tabulation (Section II-B) keep an extensive database of known proxy servers from large ISPs in order to refrain from inferring a geolocation in these cases. Denying a location answer is a first step, but not exactly a solution to the problem. This is an area for further research.

Measurement-based geolocation techniques assume that the target host is able to answer measurements (a `ping` request for instance). Nevertheless, even if the target host does not directly echo `ping` requests, a measurement-based geolocation may still be possible. A possible countermeasure that we have considered is to use `traceroute` and look for secondary targets to be measured that are relatively close in hop count to the originally intended target host. By limiting the distance in hop count and inferring the location of these secondary targets, a location estimate may be feasible at a lower accuracy.

VII. Conclusion

In this paper, we have proposed the Constraint-Based Geolocation (CBG), a measurement-based method to estimate the geographic location of Internet hosts. CBG establishes a dynamic relationship between network delay and geographic distance. This is done in a distributed and self-calibrating fashion among the adopted landmarks using the bestline method. CBG points out that accurate transformation from delay measurements to geographic distances *constraints* is indeed feasible and that in practice these constraints are often tight enough to allow an accurate location estimation using multilateration.

Our experimental results show that CBG outperforms previous geolocation techniques. The median error distance obtained in our experiments for the U.S. dataset is below 100 km while for the Western Europe dataset this value is below 25 km. These results contrast with median error distances of about 150 km for the U.S. dataset and 100 km for the Western Europe dataset when GeoPing-like methods are used. Moreover, in contrast to previous approaches, CBG assigns a confidence region to each location estimate. This is important to allow a location-aware application to assess whether the location estimate is sufficiently accurate for its needs. Our findings indicate that an accurate location estimate, i.e., with a relatively small confidence region, is provided for most cases in both datasets, thus enabling location information at a regional level granularity. Similar results have been found in a PlanetLab deployment of CBG. It might be possible, once the confidence region has been determined, to use other methods if necessary to geolocate more precisely the target host using regional landmarks. This is left for future work.

Our results are based on measurements taken in well-connected, geographically contiguous networks. To some extent our work takes advantage of the fact that network connectivity has improved dramatically in the last decade, and that the relationship between network delay and geographic distance is strong in these regions [2], [30]. Thus, one must be cautious before extrapolating our results to arbitrary network regions. Generalized geolocation to or from typical end-systems and investigation on methods to address other sources of distortion in the relationship of delay and distance that result in inaccurate estimations are part of our future work.

References

[1] M. J. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic locality of IP prefixes," in *Proc. ACM Internet Measurement Conf. (IMC 2005)*, Berkeley, CA, Oct. 2005, pp. 153–158.

[2] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. B. Duarte, "Improving the accuracy of measurement-based geographic location of Internet hosts," *Comput. Netw.*, vol. 47, no. 4, pp. 503–523, Mar. 2005.

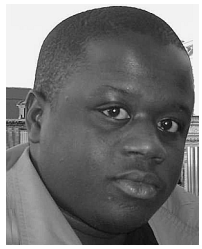
[3] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *Proc. ACM SIGCOMM*, San Diego, CA, Aug. 2001, pp. 173–185.

[4] P. Enge and P. Misra, "Special issue on global positioning system," *Proc. IEEE*, vol. 87, no. 1, pp. 3–15, Jan. 1999.

[5] S. Banerjee, T. G. Griffin, and M. Pias, "The interdomain connectivity of PlanetLab nodes," in *Proc. Passive and Active Measurement Workshop (PAM 2004)*, Antibes Juan-les-Pins, France, Apr. 2004.

[6] L. Subramanian, V. N. Padmanabhan, and R. Katz, "Geographic properties of Internet routing," in *Proc. USENIX 2002*, Monterey, CA, Jun. 2002, pp. 243–259.

[7] T. S. E. Ng and H. Zhang, "Predicting Internet network distance with coordinates-based approaches," in *Proc. IEEE INFOCOM*, New York, Jun. 2002, pp. 170–179.
[8] L. Tang and M. Crovella, "Virtual landmarks for the Internet," in *ACM Internet Measurement Conf. 2003*, Miami, FL, Oct. 2003, pp. 143–152.
[9] F. Dabek, R. Cox, F. Kaashoek, and R. Morris, "Vivaldi: A decentralized network coordinate system," in *Proc. ACM SIGCOMM 2004*, Portland, OR, Aug. 2004, pp. 15–26.
[10] R. Percacci and A. Vespignani, "Scale-free behavior of the Internet global performance," *Eur. Phys. J. B—Condensed Matter*, vol. 32, no. 4, pp. 411–414, Apr. 2003.
[11] PlanetLab: An Open Platform for Developing, Deploying, and Accessing Planetary-Scale Services. 2002 [Online]. Available: http://www.planet-lab.org
[12] C. Davis, P. Vixie, T. Goodwin, and I. Dickinson, "A means for expressing location information in the domain name system," Internet RFC 1876, Jan. 1996.
[13] IP Address to Latitude/Longitude. Univ. Illinois, Urbana-Champaign [Online]. Available: http://cello.cs.uiuc.edu/cgi-bin/slamm/ip2ll/
[14] D. Moore, R. Periakaruppan, J. Donohoe, and K. Claffy, "Where in the world is netgeo.caida.org?," presented at the INET 2000 Conf., Yokohama, Japan, Jul. 2000.
[15] GeoURL. [Online]. Available: http://www.geourl.org/
[16] Net World Map. [Online]. Available: http://www.networldmap.com/
[17] GeoNetMap. Geobytes, Inc. [Online]. Available: http://www.geobytes.com/GeoNetMap.htm
[18] GeoPoint. Quova Inc. [Online]. Available: http://www.quova.com/
[19] GTrace. CAIDA [Online]. Available: http://www.caida.org/tools/visualization/gtrace/
[20] Sarangworld Traceroute Project. 2003 [Online]. Available: http://www.sarangworld.com/TRACEROUTE/
[21] P. Bahl and V. N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *Proc. IEEE INFOCOM 2000*, Tel Aviv, Israel, Mar. 2000, pp. 775–784.
[22] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. B. Duarte, "Toward a measurement-based geographic location service," in *Proc. Passive and Active Measurement Workshop (PAM 2004)*, Antibes Juan-les-Pins, France, Apr. 2004, pp. 43–52.
[23] C. J. Bovy, H. T. Mertodimedjo, G. Hooghiemstra, H. Uijterwaal, and P. van Mieghem, "Analysis of end-to-end delay measurements in Internet," in *Proc. Passive and Active Measurement Workshop (PAM 2002)*, Fort Collins, CO, Mar. 2002.
[24] RIPE Test Traffic Measurements. 2000 [Online]. Available: http://www.ripe.net/ttm/
[25] NLANR Active Measurement Project. 1998 [Online]. Available: http://watt.nlanr.net/
[26] N. Spring, R. Mahajan, and T. Anderson, "Quantifying the causes of path inflation," in *Proc. ACM SIGCOMM 2003*, Karlsruhe, Germany, Aug. 2003, pp. 113–124.
[27] D. Krioukov, K. Fall, and X. Yang, "Compact routing on Internet-like graphs," in *Proc. IEEE INFOCOM 2004*, Hong Kong, Mar. 2004, pp. 209–219.
[28] H. Zheng, E. K. Lua, M. Pias, and T. G. Griffin, "Internet routing policies and round-trip-times," in *Proc. Passive and Active Measurement Workshop (PAM 2005)*, Boston, MA, Mar. 2005, pp. 236–250.
[29] Geographic location/privacy (geopriv). IETF working group, 2003 [Online]. Available: http://www.ietf.org/html.charters/geopriv-charter.html
[30] S.-H. Yook, H. Jeong, and A.-L. Barabási, "Modeling the Internet's large-scale topology," *Proc. National Academy of Sciences (PNAS)*, vol. 99, pp. 13382–13386, Oct. 2002.



Bamba Gueye received the B.Sc. degree in computer science from the University Cheikh Anta Diop, Dakar, Senegal, and the M.Sc. degree in networking from the University of Paris 6, France, in 2003. Currently, he is working toward the Ph.D. degree in computer networking, also at the University of Paris 6.

He is developing the GeoLIM project that aims at providing measurement-based geolocation of Internet hosts. His research interests are in Internet measurements, focusing on measurement-based geolocation and bandwidth estimation.

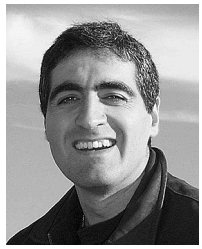
Artur Ziviani (S'99–M'04) received the B.Sc. degree in electronics engineering in 1998 and the M.Sc. degree in electrical engineering (emphasis in computer networking) in 1999, both from the Federal University of Rio de Janeiro (UFRJ), Brazil. In 2003, he received the Ph.D. degree in computer science from the University of Paris 6, France, where he was also a Lecturer during 2003–2004.

Since 2004, he has been with the National Laboratory for Scientific Computing (LNCC), Brazil. His research interests include QoS, wireless computing, Internet measurements, and the application of networking technologies in telemedicine.

Dr. Ziviani has been a member of the ACM since 2004.

Mark Crovella (M'94) is Professor of computer science at Boston University, Boston, MA. During 2003-2004, he was a Visiting Associate Professor at the Laboratoire d'Informatique de Paris VI (LIP6). His research interests are in performance evaluation, focusing on parallel and networked computer systems. In the networking arena, he has worked on characterizing the Internet and the World Wide Web; on analysis of Internet measurements, including traffic and topology measurements; and on the implications of measured Internet properties for the design of protocols and systems. He is coauthor of *Internet Measurement: Infrastructure, Traffic and Applications* (Wiley, 2006).

Dr. Crovella has been a member of the ACM since 1994.



Serge Fdida (M'88–SM'98) has been a Full Professor at the Université Pierre et Marie Curie (Paris 6) since 1991. He received the Doctorat de 3eme Cycle in 1984, and the Habilitation a Diriger des Recherches specializing in modeling of computer networks in 1989 from the University of Paris 6.

From 1989 to 1995, he was a Full Professor to the Université Rene Descartes (Paris). His research interests are in the area of high-speed networking, pervasive communication, resource management, and performance analysis. He is heading the Network and Performance group of the LIP6 Laboratory (CNRS-University of Paris 6). He was a Visiting Scientist at IBM Research during the 1990–1991 academic year. He has been the editor of the proceedings of several networking conferences and is the author of a book on performance evaluation and a book on networking. He is involved in many research projects in high-performance networking in France and Europe. He is also the Co-Director of EURONETLAB, a joint laboratory established in 2001, between University Paris 6, CNRS, THALES and 6WIND.

Prof. Fdida has been a member of the ACM since 1988.