Geolocation of Internet hosts using smartphones and crowdsourcing

Gloria Ciavarrini, Francesco Disperati, Luciano Lenzini, Valerio Luconi, Alessio Vecchio
Dip. di Ingegneria dell'Informazione, Università di Pisa, Pisa, Italy
gloria.ciavarrini@for.unipi.it, valerio.luconi@iet.unipi.it, l.lenzini@iet.unipi.it, alessio.vecchio@unipi.it

Abstract—**Knowing the position of an Internet host enables location-aware applications and services, such as restriction of content based on user's position or customized advertising. Active IP geolocation techniques estimate the position of an Internet host using measurements of the end-to-end delay between the target and a number of landmarks (hosts whose positions are known in advance). We present an IP geolocation method that operates in a crowdsourcing perspective and uses mobile devices as landmarks, since their position can be easily computed using the GPS unit. A specific calibration has been included to take into account the particular operating environment which, differently from the past, includes the presence of wireless links.**

I. INTRODUCTION

Knowing the position of an Internet host can be useful in several applications and services. Examples include the distribution of content based on user's location, authorization of transactions only when user's position corresponds to trusted areas, or the analysis of the Internet along a geographical dimension. As known, there is no direct relationship between an IP address and the position of the host with such address [1]. Some databases provide a mapping between hosts and their believed coordinates. Unfortunately, geographic data stored in these databases is not particularly accurate [2], since the position of a host is determined using administrative information. As a consequence, when organizations are particularly large, the error between the estimated and the real position can be in the order of thousands of kilometers [3], [4].

Active IP geolocation techniques determine the position of an Internet host via network measurements. In general, the basic idea is to find the distance between the host to be localized and a number of landmarks (a set of hosts whose position is known in advance), then trilateration (or another geometrical technique) is applied. The distance between the target and a given landmark is estimated using end-to-end delay measurements. Thus, landmarks actively send probes towards the target host to measure the delay; delays are converted into distances according to a delay-distance model; finally results are collected on a central server where the position is calculated.

We devised an active IP geolocation method that uses smartphones as landmarks. The proposed method is based on crowdsourcing principles: a number of users participate voluntarily to the system and provide their devices as measuring elements [5], [6].

With respect to existing literature, this work contributes as follows: i) for the first time the use of mobile devices, whose position is known thanks to their GPS receiver, is considered to determine the location of an Internet host; ii) the localization procedure is calibrated according to the particular operating conditions of mobile devices that, differently from previous studies, are connected to the Internet via wireless links (cellular and Wi-Fi); iii) landmarks do not belong to research/academic networks and are enrolled via crowdsourcing.

The paper is organized as follows: in Section II literature concerning geolocation of Internet hosts is summarized, focusing on active techniques; Section III describes the localization method and collection of data; in Section IV, the calibration procedure for our particular environment is described; Section V discusses some landmark selection techniques (which are relevant for our crowdsourcing-based approach); Section VI reports results and Section VII draws the conclusions.

II. RELATED WORK

The problem of finding the geographical position of a host with a given IP address has been studied quite extensively in the last decade, although a complete and reliable solution still has to be found. Several methods are based on registries or other static sources of information. The approach discussed in [7] relies on the information extracted from the DNS. Nevertheless, this data is provided by administrators and not automatically produced, thus it is not frequently updated and it is characterized by significant errors (as shown in [8]).

The WHOIS protocol is a de-facto standard for querying domain name information. Several systems use WHOIS databases for IP geolocation purposes, but records are often unsynchronized or inconsistent. To overcome some of the previous problems, the traceroute utility has been used. Traceroute returns the addresses of router interfaces along a network path between two hosts. In fact, when addresses can be converted in names via DNS, these frequently contain city, country or airport codes, and GeoTrack uses this information to infer the location of the target [2]. However the format of DNS names is not standardized and this makes the technique not always applicable.

GeoCluster, developed by the same authors of GeoTrack, organizes IP addresses in geographical clusters using information extracted from BGP routing tables [2]. Subsequently, if some IP addresses in a cluster can be localized using external sources, then the position of the other addresses in the same cluster can be computed.

Other techniques do not rely on information contained in databases and try to determine the position of a host using

active measurements. Delay-distance approaches use the end-to-end delay for estimating the distance between targets and landmarks. Unfortunately, the relationship between network performance and geographic distance is affected by errors introduced by the circuitousness of Internet routes and the presence of multiple ISPs in a path [9]. In [10] the role of topology on localization is investigated. For overcoming the circuitousness and irregularity of Internet paths, and hence improving the performance, landmarks should be uniformly distributed on the surface of Earth.

In [4] the authors propose a technique that does not require landmark-specific calibration. The technique relies on a probabilistic approach, where the distribution of distances for a given delay is independent from the position of the landmark. This makes the system more resilient to measurement errors and possible network anomalies.

Ziviani et al. [11] explored how a demographic landmark distribution provides more accurate location estimation. They also highlighted that the density of connectivity influences the correlation between geographic distance and network delay.

In [12], the authors propose a multi-phase approach to IP geolocation. In a first phase, the coarse region where the target is located is estimated using the absolute delays with respect to a number of hosts with known position (similarly to other works). Then, the position of the target is determined more precisely by measuring the relative delay against a set of passive landmarks located in its surroundings. Passive landmarks are dynamically found by mining resources available on the Web (e.g. businesses, universities or offices with a given ZIP code).

All the previously mentioned works rely on measurements that have been collected in rather protected environments, mostly characterized by high speed wired links. In such scenarios the measurement process is relatively stable and reliable. This paper discusses a novel approach where the measuring endpoints are smartphones enrolled according to crowdsourcing principles. Since measurements are collected via wireless links and using devices belonging to normal users, specific procedures for calibrating the delay-distance model and filtering devices have been developed.

III. SMARTPHONE-BASED LOCALIZATION

We propose an active IP geolocation technique that uses smartphones as landmarks. The use of smartphones for this purpose is motivated by their always increasing number, which makes them an attractive platform for geographically distributed applications. Moreover, smartphones participating in measurements are enrolled according to crowdsourcing principles; as a consequence, they are under control of their respective owners and the set of devices dynamically changes according to to their will.

These particular features (mobility of landmarks and crowdsourcing) make our approach rather different from the ones proposed in the past, where landmarks were always fixed hosts belonging to a rather homogeneous set (usually they were placed within academic/research networks).

$$\mathcal{S} \leftarrow \emptyset$$
for all $i \in \{1, \ldots, n\}$ **do**
    if $\mathcal{S} = \emptyset$ **then**
        $\mathcal{S} \leftarrow C_{L_i}$
    else
        if $\mathcal{S} \cap C_{L_i} \neq \emptyset$ **then**
            $\mathcal{S} \leftarrow \mathcal{S} \cap C_{L_i}$
        end if
    end if
end for

Fig. 1: Calculation of $\mathcal{S}$.

A. Method

Active geolocation of Internet hosts is based on collecting round trip time (RTT) measurements from a set of landmarks to the target IP address. Then, after having converted RTTs into distances, a geometric technique is used to estimate the position of the target.

Literature abounds of geometrical techniques for finding the position of a target. We used a variation of the intersection technique described in [13], as we found it to be rather resilient to measurement errors. This technique is based on the assumption that smaller RTTs correspond to measurements affected, in general, by smaller errors. In our scenario RTT values show larger variability with respect to previous studies, because of two main factors: i) wireless networks are characterized by higher variability, in terms of delay, than wired networks (as a consequence of collisions and/or interferences); ii) smartphones' HW and OS are less suitable for collecting accurate timestamps (e.g., because measurement are collected at user-level in the presence of multitasking, energy saving policies, etc).

More formally, let $\mathcal{L}^k = \{L_1^k, L_2^k, ..., L_n^k\}$ be the set of landmarks participating in the measurement towards the k-th target, where L_i^k is the i-th landmark. Let $\mathcal{M}_i^k = \{m_{i,1}^k, m_{i,2}^k, \ldots, m_{i,R_i^k}^k\}$ be the set of RTTs between L_i^k and the target T^k, where R_i^k is the number of RTTs collected by such landmark. To make notation easier to read, from now on we will not use the k superscript when describing a given measurement unless it is strictly necessary. Each landmark selects the minimum RTT $\hat{m}_i$ towards the target host ($\hat{m}_i = min(\mathcal{M}_i)$). Then it computes the distance from the target T as:

$$r_i = \frac{\hat{m}_i}{2} \cdot CF \cdot c \qquad (1)$$

where c is the speed of light in vacuum ($c \simeq 3 \cdot 10^6$ km/s), and CF is the Conversion Factor, a coefficient that models the distance-delay ratio (discussed in Section IV).

For each L_i the circular area C_{L_i} of radius r_i and centered in L_i is calculated, as the target T is supposed to be located in such region. Let $\mathcal{C} = \{C_{L_1}, C_{L_2}, \ldots, C_{L_n}\}$ be the set of all circles ordered in ascending order of r_i. Thus C_{L_1} will be the circle with the smallest radius and C_{L_n} the circle with the largest radius. Let then $\mathcal{S}$ be the convex region where the target T should be located, according to all measurements. The algorithm shown in Figure 1 is used to compute $\mathcal{S}$. In
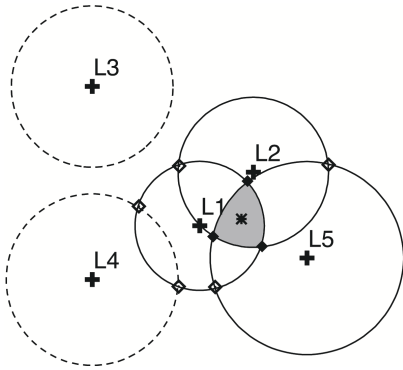
Fig. 2: Example of localization using intersection.

Fig. 3: Geographic position of landmarks (blue) and targets (red).

practice, the intersection between circular areas is calculated starting from the smallest ones and discarding those with empty intersection with the previous ones. An example of the localization technique is shown in Figure 2. $L1, ..., L5$ is the set of landmarks. Circles correspond to the distances calculated according to Equation 1. $L1$ is the landmark that registers the smallest RTT towards the target. The intersection between the circle centered in $L2$ and the previous one is then calculated. Circles centered in $L3$ and $L4$ do not overlap with this area and thus do not contribute further to calculate the position of the target. The circle centered in $L5$ partially overlaps the area defined by the intersection between the circles generated by $L1$ and $L2$ and defines the area depicted in gray. The estimated position of T is finally calculated as the barycenter of $\mathcal{S}$.

B. Collection of measurements

Smartphones operate as landmarks, since they are generally equipped with a GPS receiver and thus their position is known. The set of smartphones involved in measurements is not constant: as these devices have been enrolled according to crowdsourcing principles, device owners installed and uninstalled the measurement software according to their necessities; moreover, the set of smartphones participating in measurements also depends on their status at the time of measurements (turned on/off, connected/disconnected). In the end, we registered a rather large amount of variability in the number of smartphones participating in measurements (from 2 to 157). Enrolled devices are participants of the Portolan project [14], a smartphone-based network measurement platform. To be part of the Portolan project, users just have to install an app on their devices. For the experiments described in this paper, only Android-based smartphones have been considered (Portolan is also available for standard PCs running Linux, Windows or Mac OS X).

The set of targets is composed of 401 hosts participating in the PlanetLab network [15]. We used PlanetLab nodes as targets because their real position is known. The real position of targets has been used to compute the error of the proposed method (i.e. the difference between the real position and the estimated position).

For performing a measurement each landmark sent a number of probes. The average number of collected RTTs is approximately equal to 30 ($R \simeq 30$). The exact number
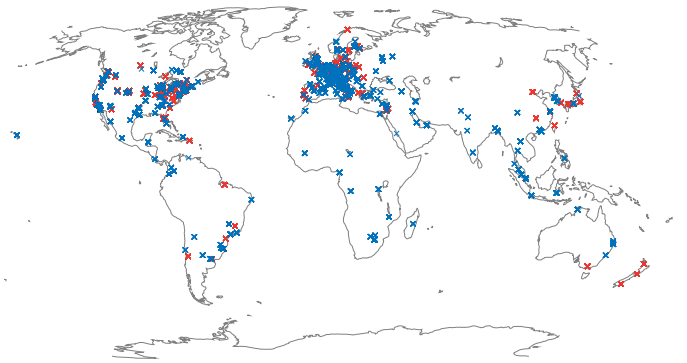
depended on a variety of factors, related to both wireless access and crowdsourcing approach. Events that may reduce the number of collected RTTs include interrupted connections because of smartphone movement, running out of energy, or reaching the limits of users' data plan. Measurements were triggered by commands issued by the Portolan server. Actual measurements towards a given target took place at slightly different times for the involved landmarks, even though the commands are sent by the server almost simultaneously. Also in this case, the reasons are mostly related to the wireless access and crowdsourcing approach: if the command is sent by the server when a smartphone is not connected to the Internet, the command is delivered as soon as the device reconnects. Strong synchronization among landmarks is not needed for IP geolocation purposes. Nevertheless, a large interval between measurements increases the probability of finding slightly different network conditions (e.g. because of congestion) and thus, in turn, to increase the error affecting the estimated distances. Probes were based on UDP, instead of ICMP as commonly done in similar systems. The reason for this choice is due to the unavailability of ICMP sockets in Android without superuser privileges (the Portolan app runs as a standard app at the user level) [16].

All measurements performed by smartphones have been collected on a central server, where they have been saved onto persistent storage. This data has been used for both calibrating and evaluating the proposed method. Both calibration and evaluation have been performed off-line, to ensure repeatability of experiments and the evaluation of different strategies on the same dataset.

The geographic position of all landmarks (smartphones) and targets is depicted in Figure 3. Both landmarks and targets are more dense in Europe and North America than in other continents. This reflects the distribution of both Portolan and PlanetLab nodes, which are not uniformly spread on the globe.

IV. CALIBRATION FOR A WIRELESS SCENARIO

The outcome of Equation 1 strongly depends on the CF coefficient. In this section we describe the procedure followed to calibrate CF according to the collected data.

In the Internet, the end-to-end delay is composed of a deterministic part and a stochastic part. The deterministic
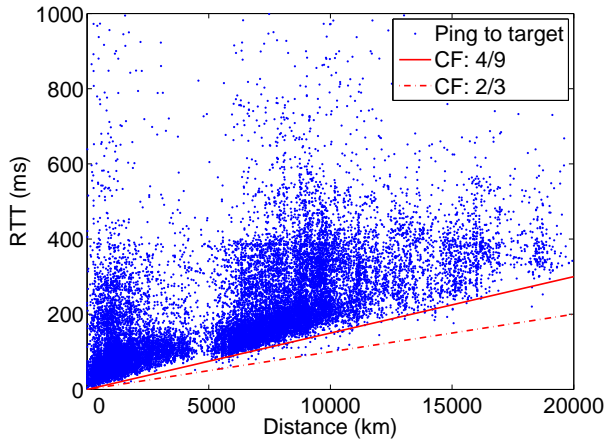
Fig. 4: Scatterplot of minimum RTT vs great circle distance between landmark and target.

delay is due to the propagation delay, the transmission delay, and the minimum processing delay achievable in each router along the path. The stochastic delay is variable and it is due to intermediate queues and extra processing time in routers along the path. The conversion from delay measurement to geographic distance is also distorted by other unpredictable factors such as circuitous routing and the possible different paths followed to reach the target. In our scenario, landmarks (i.e. smartphones) are connected to the Internet via wireless connections which, as mentioned, are characterized by greater delay and variability with respect to wired links.

The theoretical minimum RTT that can be obtained is equal to the required time for a signal to travel in optic fiber along the great circle distance[1] between the sender and the receiver, without additional time for processing and queuing in intermediate routers. Let d the great circle distance and c the light speed in vacuum. The speed in optical fiber is $v_{opt} \simeq 2/3 \cdot c$ [17], which thus represent the maximum speed for a real system. Nevertheless, such value cannot be easily obtained in practice, and measurements in our dataset allow us to state that in our scenario the packet speed is at most $v_{max} = 4/9 \cdot c$. This upper bound is consistent with [10].

Figure 4 shows the experienced RTT vs the real distance in our dataset (the real distance can be calculated because all targets belong to PlanetLab, and thus their coordinates are known, whereas coordinates of landmarks are acquired via GPS). Each point represents the minimum RTT obtained during measurements. The dotted red line is the theoretical minimum RTT obtained as $RTT_{min} = d/v_{opt}$. The continuous red line is obtained using $CF = 4/9$. Figure 4 shows that more than 99% of ~ 21700 sample points are above the continuous red line ($CF = 4/9$) and that only ~ 90 sample points have $CF = 2/3$ as lower bound.

The choice of CF has a strong influence on the localization process: if CF is too large it produces overestimated circular areas, whereas a small CF value leads to underestimate them. In the first case, the intersection among circular areas will be

[1] The great circle distance is the shortest distance between two points on the surface of a sphere, measured along the surface of the sphere.

TABLE I: Access technologies, percentages of use, and CF values.

Type	%	CF
2.5G	2.18%	0.1553
3G	27.71%	0.1930
4G	7.75%	0.2519
Wi-Fi	60.14%	0.2414
Not Available	2.22%	-

large and the estimated position will be affected by a relatively large inaccuracy, as shown in Figure 5a. On the other hand, in the second case, areas may not cover the target position or may generate an empty intersection, as shown in Figure 5b. Figure 5c depicts the ideal conditions.

In previous works, like CBG [13], an offline calibration was performed. In particular, the line of best fit is calculated independently for each landmark. In a scenario characterized by mobile devices this approach cannot be applied: a smartphone changes its position and access technology in an unpredictable way. Moreover, the set of landmarks participating in a measurement can be completely different from the set of landmarks available during the calibration phase (because of the crowdsourcing approach).

We performed an offline statistical analysis of our dataset in order to estimate the optimal conversion factor CF. In details, we used linear least squares regression, and we obtained a value equal to 0.2234 (let us call CF_{ALL} this value).

A. Calibration based on access technology

When a measurement is performed, each landmark sends to the remote database additional information, including its geographical coordinates, network operator name and the type of access technology (Wi-Fi, 3G, 4G, etc.).

We investigated if information concerning the technology used for accessing the Internet by landmarks can be used to improve the performance of the localization method. The basic idea is to perform a customized calibration of CF for the different access technologies, instead of using a single value for all of them, in order to increase the accuracy of distance estimation.

Table I shows the fraction of measurements for the different access technologies. As can be noticed, the majority of measurements have been collected using Wi-Fi, whereas 4G and 3G are the other most used technologies. The values of CF obtained according to a per-technology calibration are also reported (let us call CF_{PT} this values). As expected the conversion factor grows when faster wireless connections are used. Figure 6 shows the RTT vs distance scatterplots, separated per access technology. For all of them, the coefficient has been calculated using linear least squares regression. In terms of general trend the scatterplots do not differ significantly, with the exception of the one concerning 2.5G that is characterized by a reduced number of measurements.

V. SELECTION OF LANDMARKS

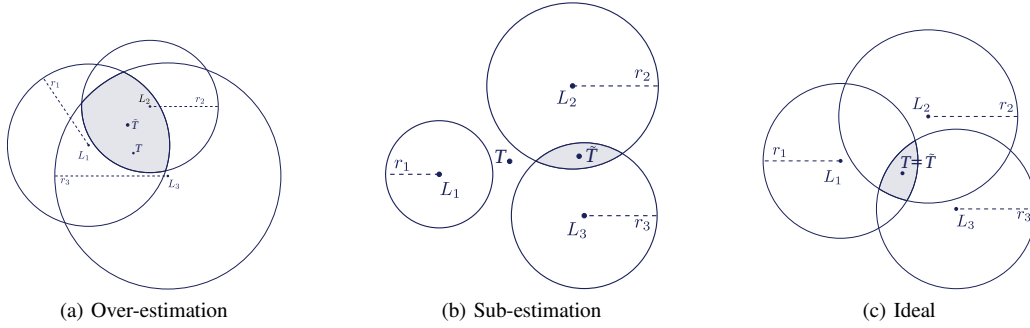From preliminary experiments we observed that, in many cases, only a fraction of landmarks contributes to localize the

(a) Over-estimation   (b) Sub-estimation   (c) Ideal

Fig. 5: Varying CF: effects on intersection-based localization. T is the real position, $\tilde{T}$ is the estimated position.

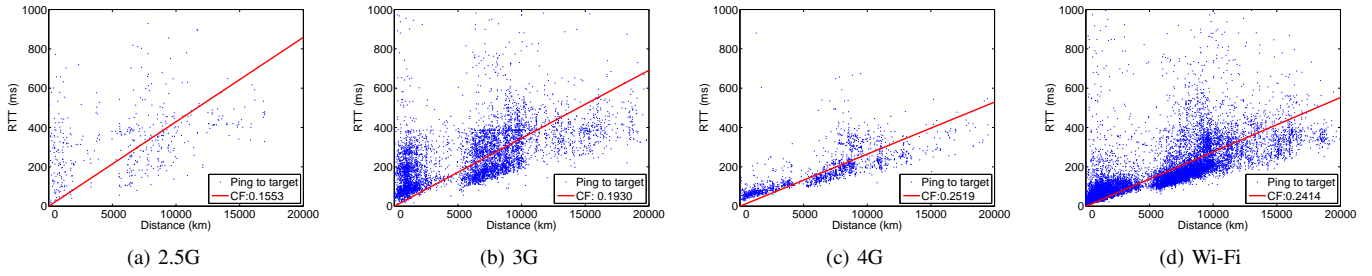(a) 2.5G   (b) 3G   (c) 4G   (d) Wi-Fi

Fig. 6: Scatterplot of RTT vs distance separated per access technology.

(a) Landmarks in North America have no influence on the result (landmarks in blue, estimated position in red).

(b) Wrong result caused by a low-latency transatlantic link (landmarks in blue, estimated position in red, real position in green).
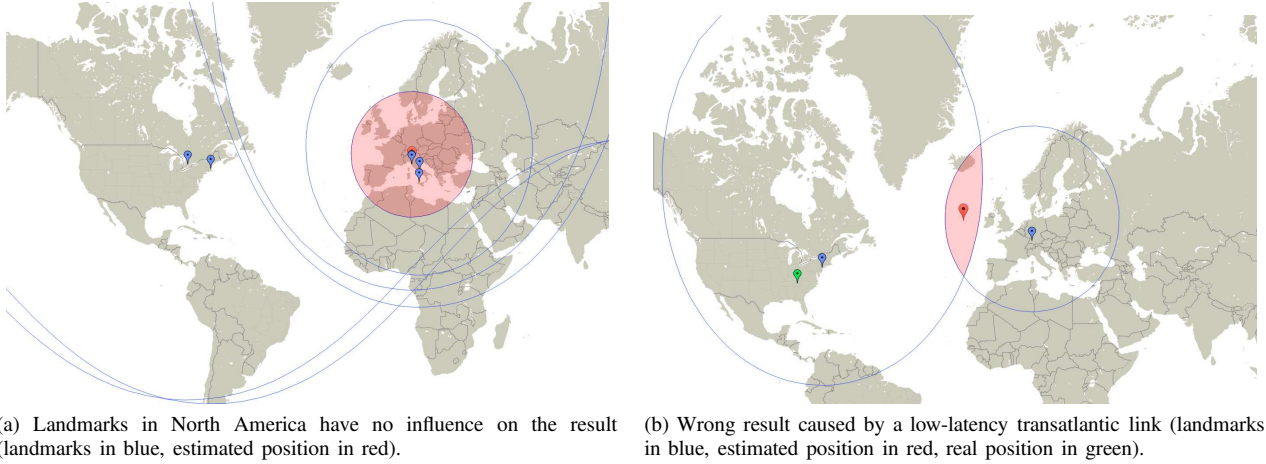
Fig. 7: Examples of localization when using landmarks placed in different continents.

target. In fact, RTTs registered by some landmarks are so large that the circular area associated to such landmarks can cover an hemisphere. This situation occurs when landmarks are very far from the target, e.g. when they are in Europe and the target is located in South America. For instance, in Figure 7a the region obtained at the end of the process (in red) is completely contained within the regions produced using landmarks located in North America [2].

Moreover, intercontinental backbone links can be the source of additional errors because the CF value is calibrated for paths that are generally slower. Figure 7b shows this problem:

[2] Regions in Figure 7a do not appear as circular because of the Mercator projection, which is based on a cylindrical representation. Deformations are particularly evident for large circles, which look like "sinusoids".

the target, located in North America, is erroneously positioned in the middle of Atlantic Ocean as results from the intersection of the two circular regions. In this specific case, the region produced by the European landmark is largely sub-estimated due to a particularly fast and direct link (with respect to the average paths from which CF has been calculated).

If the RTT scatterplots are observed in detail, it is easy to notice that measurements at approximately 5000 km are less dense. This somehow empty region approximately corresponds to intercontinental distances. To confirm this hypothesis, the source and destination continent for all RTTs has been determined: Figure 8a makes clear that almost all paths above 5000 km belong to different continents.

The separation degree between these two classes can be
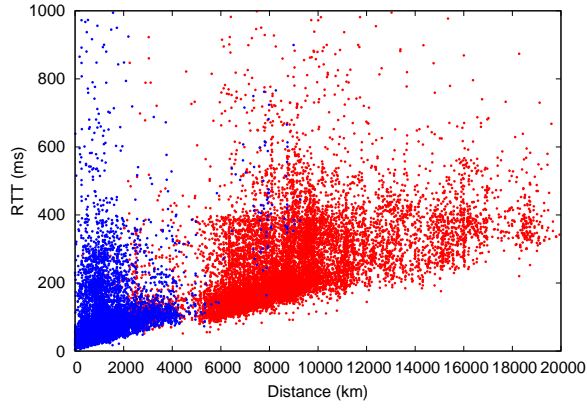
(a) Scatterplot of RTT vs distance for intra- and inter-continental paths (in red and blue, respectively).
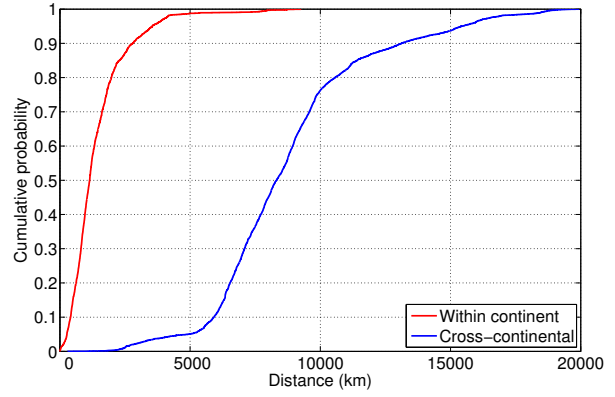
(b) Classification of paths as intra- and inter-continental (CDF).

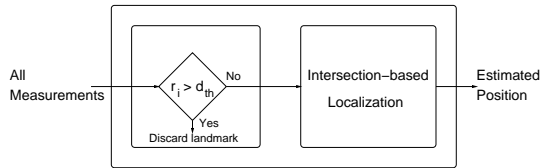Fig. 8: Intra- and inter-continental paths.

Fig. 9: Global localization procedure.

observed in Figure 8b: 99% of the distances having landmark and target in the same continent are below 5900 km, whereas only 12% of the intercontinental distances are below the same value.

In general, landmarks with higher RTT values are less accurate with respect to landmarks with smaller values. This is reasonable, as the larger the distance, the higher the probability to incur in congested nodes or more indirect paths [18]. For these reasons, it is preferable to not include landmarks with these characteristics in the localization process. Thus, the global localization procedure operates as shown in Figure 9: all measurements are given as input to the localization system; a filter removes all measurements that correspond to landmarks that are distant from the target more than a given threshold (d_{th}); remaining measurements are provided as input to the intersection localization method.

By limiting the maximum estimated distance, the localization accuracy of the system increases (when the distance between a landmark and the target is higher than a given threshold the measurement is discarded). Nevertheless, if too many measurements are discarded the system may become unable to provide an estimated position for the target.

VI. RESULTS

Figure 10 reports the localization results when using the literature-based value of CF (4/9) and the CF calibrated for our scenario. The localization procedure performs better when using conversion factors specifically calibrated for a wireless environment than the literature-based value (calculated in wired environment). In particular, the median localization error is equal to ~ 820 km when using CF_{ALL} and ~ 953 km when
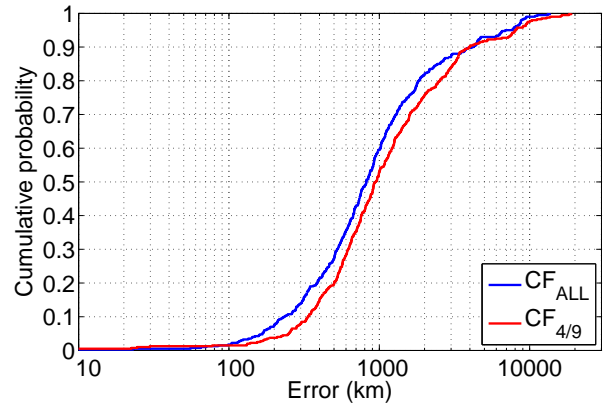

Fig. 10: CDF of localization error when using different CF values.

using $4/9$. These results have been obtained without selection of landmarks ($d_{th} = +\infty$ and thus all landmarks have been used, independently from their distance from the target).

The performance of the localization method when using CF_{ALL} and CF_{PT} is depicted in Figure 11a. Since the performance of the localization procedure at the same time depends on the maximum landmark-target distance, Figure 11a reports the obtained results for a set of possible values (d_{th} is varied between 1000 km and 7000 km). As expected, when the threshold used to discard landmarks is smaller, the localization accuracy gets better. This improvement has a negative side-effect: the fraction of localized targets becomes smaller as well. This is due to the fact that the localization procedure is unsuccessful when all targets are discarded during the selection phase. Such phenomenon is mostly a consequence of the crowdsourcing approach: the number of smartphones involved in measurements is not under control and in some cases is rather small. Figure 11b shows the distribution of estimation error when CF_{ALL} and CF_{PT} are used. Both curves have been calculated with $d_{th} = 2500$ km. In both cases, the median localization error is ~ 580 km. Selection of landmarks with such threshold improves localization, in terms
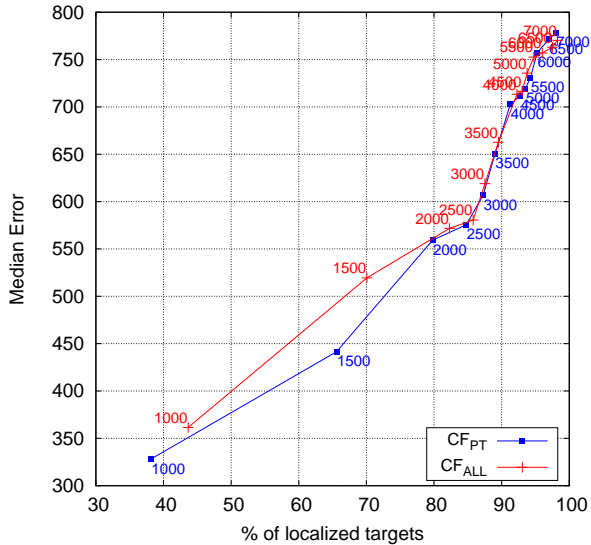
(a) Combined performance of localization (error and percentage of success) depending on landmark-target distance
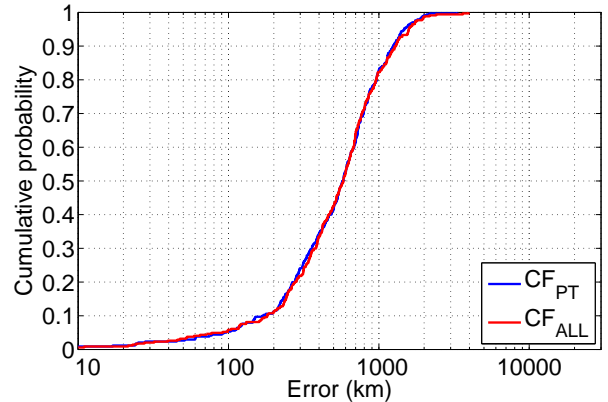
(b) Localization error when varying the maximum landmark-target distance (logarithmic scale)

Fig. 11: Selection of landmarks depending on their distance from the target.

of median error, of approximately 250 km, with respect to absence of filtering. No significant differences emerge when using CF_{ALL} or CF_{PT}, with exception of the cases where d_{th} is below 2000 km (i.e. when aggressive selection is used). This may be due to the large fluctuations that affect RTTs, thus the adoption of a better delay-distance model is almost nullified by the large noise.

Error significantly increases when the distance between targets and landmarks becomes higher, as the delay component due to queues and possible congestion is more significant. Figure 12a depicts this finding. At the same time, the localization error is inversely proportional to the number of landmarks that take part in localizing a target: the higher the number of used landmarks, the better the localization accuracy (as shown in Figure 12b). Both these problems can be reduced by increasing significantly the number of devices that participate in the localization process. Unfortunately this is not easy to achieve, as participation in a crowdsourcing-based platform depends on the utility perceived by users. Moreover, participation in a crowdsourcing platform can be either voluntary, as it happens in our case, or based on incentives (see [19], [20] for a discussion about incentived crowdsourcing and its use in smartphone-based scenario, respectively). However, also in the latter case, finding the appropriate reward mechanisms is not trivial (possible incentives include money, access to services, and altruism) [21], [22].

We verified the above considerations by restricting the analysis to Europe, i.e. considering only the targets located in Europe. This increases the density of landmarks in the surrounding of targets because a relatively large fraction of Portolan users are located in this continent. At the same time, also the average distance between targets and landmarks gets reduced with respect to the global case. In particular, 157 of the 401 targets are located in Europe. In such conditions, the median localization error becomes 383 km. It is interesting

to notice that this last result has been obtained using CF_{PT}, whereas the use of CF_{ALL} provides worse results (429 km). We may conclude that when measurements are less noisy, e.g. when the average landmark-target distance is smaller, the use of per-technology calibration brings some improvement.

VII. Conclusion

We presented an IP geolocation method that is based on a crowdsourcing approach and that uses smartphones as landmarks. Wireless access makes the localization process particularly troublesome. Independently from the specific access technology (cellular, Wi-Fi) the end-to-end delay is larger and less stable than the one obtained using wired access. As a consequence, despite using a distance-delay model specifically calibrated for a wireless scenario, estimated distances are less accurate than the ones obtained in a wired scenario. Another factor is the impact on the end-to-end delay of inter-AS (autonomous system) communication. In our scenario, smartphones addresses belong to ISPs and cellular operators ("commercial" ASes), whereas targets are part of academic and research networks. In several previous studies both landmarks and targets were placed in "similar" or "close" ASes (in some cases both categories were hosted in the PlanetLab network), thus reducing the number of ASes to be traversed and in the end reducing also the jitter that affects the delay. Nevertheless, the influence of these factors on the end-to-end delay was not quantified in existing literature and is left for future work.
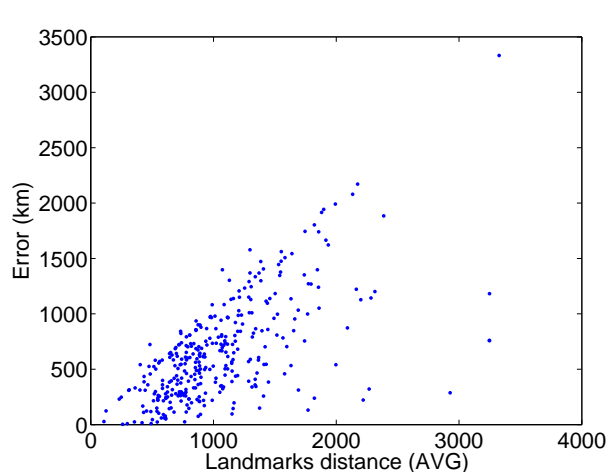
(a) Localization error vs average landmark-target distance.
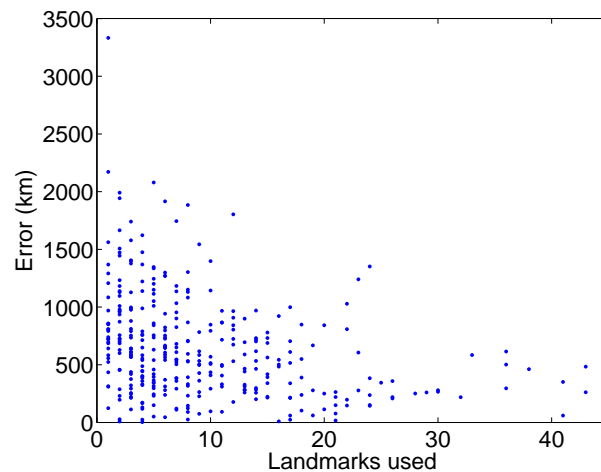
(b) Localization error vs number of used landmarks

Fig. 12: Localization error depending on the landmark-target distance and number of used landmarks.

REFERENCES

[1] M. J. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic Locality of IP Prefixes," in *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '05. Berkeley, CA, USA: USENIX Association, 2005, pp. 13–13. [Online]. Available: http://dl.acm.org/citation.cfm?id=1251086.1251099

[2] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," *SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 4, pp. 173–185, Aug. 2001. [Online]. Available: http://doi.acm.org/10.1145/964723.383073

[3] S. Laki, P. Mátray, P. Hága, I. Csabai, and G. Vattay, "A model based approach for improving router geolocation," *Comput. Netw.*, vol. 54, no. 9, pp. 1490–1501, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2009.12.004

[4] S. Laki, P. Matray, P. Haga, T. Sebok, I. Csabai, and G. Vattay, "Spotter: A model based active geolocation service," in *Proceedings of IEEE INFOCOM*, April 2011, pp. 3173–3181.

[5] J. Howe, *Crowdsourcing: How the power of the crowd is driving the future of business.* Random House, 2008.

[6] D. R. Choffnes, F. E. Bustamante, and Z. Ge, "Crowdsourcing service-level network event monitoring," *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 387–398, Aug. 2010. [Online]. Available: http://doi.acm.org/10.1145/1851275.1851228

[7] C. Davis, P. Vixie, T. Goodwin, and I. Dickinson, "A means for expressing location information in the Domain Name System. Network Working Group RFC 1876," 1996.

[8] M. Zhang, Y. Ruan, V. S. Pai, and J. Rexford, "How DNS misnaming distorts Internet topology mapping," in *Proceedings of the USENIX Annual Technical Conference*, 2006, pp. 369–374.

[9] L. Subramanian, V. N. Padmanabhan, and R. H. Katz, "Geographic properties of Internet routing," in *Proceedings of the General Track of the Annual Conference on USENIX Annual Technical Conference.* Berkeley, CA, USA: USENIX Association, 2002, pp. 243–259. [Online]. Available: http://dl.acm.org/citation.cfm?id=647057.713857

[10] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, ser. IMC '06. New York, NY, USA: ACM, 2006, pp. 71–84. [Online]. Available: http://doi.acm.org/10.1145/1177080.1177090

[11] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. Duarte, "Improving the accuracy of measurement-based geographic location of Internet hosts," *Computer Networks*, vol. 47, no. 4, pp. 503 – 523, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138912860400249X

[12] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards street-level client-independent IP geolocation," in *Proceedings of the 8th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2011, pp. 365–379.

[13] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of Internet hosts," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1219–1232, 2006.

[14] A. Faggiani, E. Gregori, L. Lenzini, V. Luconi, and A. Vecchio, "Smartphone-based crowdsourcing for network monitoring: Opportunities, challenges, and a case study," *Communications Magazine, IEEE*, vol. 52, no. 1, pp. 106–113, January 2014.

[15] "PlanetLab," https://www.planet-lab.org/.

[16] A. Faggiani, E. Gregori, L. Lenzini, S. Mainardi, and A. Vecchio, "On the feasibility of measuring the Internet through smartphone-based crowdsourcing," in *Proceedings of the 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2012, pp. 318–323.

[17] R. Percacci and A. Vespignani, "Scale-free behavior of the Internet global performance," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 32, no. 4, pp. 411–414, 2003. [Online]. Available: http://dx.doi.org/10.1140/epjb/e2003-00123-6

[18] B. Wong, I. Stoyanov, and 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.* Berkeley, CA, USA: USENIX Association, 2007, pp. 23–23. [Online]. Available: http://dl.acm.org/citation.cfm?id=1973430.1973453

[19] J. J. Horton and L. B. Chilton, "The labor economics of paid crowdsourcing," in *Proceedings of the 11th ACM conference on Electronic commerce.* ACM, 2010, pp. 209–218.

[20] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: ACM, 2012, pp. 173–184. [Online]. Available: http://doi.acm.org/10.1145/2348543.2348567

[21] A. Faggiani, E. Gregori, L. Lenzini, V. Luconi, and A. Vecchio, "Lessons learned from the design, implementation, and management of a smartphone-based crowdsourcing system," in *Proceedings of First International Workshop on Sensing and Big Data Mining.* New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2536714.2536717

[22] A. Mao, E. Kamar, Y. Chen, E. Horvitz, M. E. Schwamb, C. J. Lintott, and A. M. Smith, "Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing," in *First AAAI Conference on Human Computation and Crowdsourcing*, Mar. 2013. [Online]. Available: http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7497