Vers un compromis entre mesures actives et mesures passives pour la localisation géographique des hôtes dans l'Internet

Bamba Gueye

Université Cheikh Anta Diop de Dakar
Faculté des Scicences et Techniques
Département de Mathématiques et Informatique
Dakar, Sénégal
bamba.gueye@ucad.edu.sn

Résumé L'inférence de la localisation géographique d'hôtes dans l'Internet à partir uniquement de leur identifiant a vu son importance économique grandir ces dernières années. Les techniques de géolocalisation basées sur des mesures de délai engendrent beaucoup de trafic dans le réseau lors de l' inférence de la localisation d'un hôte cible, toutefois, avec une meilleure précision comparées aux bases de données. Ainsi, nous proposons une technique hybride de géolocalisation qui combine l'utilisation des bases de données et des mesures de délai. Les résultats obtenus montrent que, si l'on choisi les hôtes références les plus proches géographiquement de la cible, pour le localiser, nous obtenons une meilleure estimation de localisation. En outre, la combinaison de la base de données avec les mesures de délai permet de réduire le nombre d'hôtes références utilisés et par ailleurs réduire le trafic injecté dans le réseau ainsi que le temps de réponse pour inférer la position de la cible.

Keywords : Géolocalisation, Mesure, Performance.

1 Introduction

La localisation géographique des hôtes dans l'Internet à partir de leur adresse IP connaît un intérêt croissant. Elle permet l'émergence de nouvelles applications variées basées sur la localisation. De nombreux services dépendent et se servent de la position des utilisateurs pour rendre des services personnalisés.

Dans l'Internet, la correspondance géographique des adresses IP est nécessaire aux services tels que : la publicité ciblée sur les pages WEB, l'utilisation de « redirect » qui permet d'orienter automatiquement un internaute vers un site national en fonction de sa localisation (ex. Google), la diffusion restreinte de contenu, identification basée sur la localisation du client pour le commerce électronique, lutte contre la cyber-criminalité. Il ne semble pas nécessaire d'insister ici sur l'enjeu que représente la lutte contre la pédocriminalité sur Internet. Un service de géolocalisation peut permettre à identifier les individus qui téléchargent et/ou proposent des contenus ou bien ceux qui ont un comportement suspect lors de discussion en ligne.

Toutefois, les identifiants (adresses IP) utilisés pour identifier les hôtes terminaux dans l'Internet sont alloués de manière arbitraire et il n'existe pas de relation entre une adresse IP et la position géographique [1–4] de l'équipement qui possède cette adresse. Ainsi les techniques qui permettent de localiser les hôtes dans l'Internet à partir de leur adresse utilisent deux types de mesure :

- les outils de localisation qualifiés de *mesures passives* utilisent des bases de données qui contiennent des blocs d'adresses IP et leur information de localisation. Les applications commerciales de géolocalisation comme [5–9] qui utilisent des bases de données peuvent être qualifiées de techniques passives. Nous avons aussi des bases de données à libre accès comme [10–12]. Ces bases de données contiennent des préfixes d'adresses IP et une information de localisation qui leur est associée. Bien qu'elles fournissent un temps de réponse assez rapide lorsqu'elles reçoivent une requête de localisation, la méthodologie et la précision de ces techniques de géolocalisation restent inconnues.
- les *mesures actives* (exemple [1, 4, 13, 14]) qui envoient des paquets sondes par des hôtes références dans le réseau pour localiser les hôtes. On appelle hôte référence tout hôte dont on connaît la position géographique.

Les techniques de géolocalisation basées sur des mesures de délai [1, 4, 13, 14] engendraient beaucoup de trafic dans le réseau lors de l' inférence de la localisation d'un hôte cible, avec une meilleure précision comparées aux bases de données, mais avec un temps temps de réponse plus important. Il faut noter que, pour certaines applications dans l'Internet qui peuvent avoir besoin d'un service de géolocalisation, ce temps de réponse doit être équivalent au temps de chargement d'une page WEB recevant une requête, soit environ une à trois secondes en moyenne. Des travaux précédents [15–17] ont montré l'incohérence et le manque de précision des applications commerciales de géolocalisation qui utilisent les bases de données. En effet, une requête de localisation adressée à la base de données fournit comme réponse la position géographique associée à ce bloc par l'entité administrative qui gère ce bloc, bien que les hôtes cibles peuvent être localisés de manière dispersée à l'intérieur de ce bloc.

Ainsi, nous proposons une technique hybride de géolocalisation qui utilise une base de données et des mesures de délai pour inférer la position des hôtes cibles dans l'Internet. Comme technique de mesures actives nous considérons la technique *Constraint-Based Geolocation (CBG)* proposée par [4] et basée sur la multilatération. En effet, la multilatération [4] permet d'estimer une position en utilisant un nombre suffisant de distances à partir de quelques points immobiles (hôtes références). Dès lors, elle fournit un ensemble continu d'endroits où on peut localiser la cible au lieu d'un espace discret de réponses comme dans [1].

L'idée principale est de localiser l'hôte cible avec la base de données et ensuite de raffiner la précision de l'estimation de localisation par des mesures actives. Ainsi, le serveur de géolocalisation implémente une heuristique qui lui permet de choisir l'ensemble des hôtes références qui feront les mesures vers l'hôte cible. Pour ce faire, le serveur envoie une requête à la base de données pour connaître la localisation géographique de cet hôte cible. Si cette localisation existe, alors

la base de données lui renvoie la latitude et la longitude du préfixe d'adresses IP auquel appartient cet hôte cible. L'heuristique consiste, pour un nombre d'hôtes références fixé, à choisir les hôtes références les plus proches à l'hôte cible en terme de distance géographique.

Ce papier est organisé comme suit. La section 2 décrit l'architecture hybride de géolocalisation. Dans la section 3 nous présentons les heuristiques considérées pour associer une technique de mesures passives et une technique de mesures actives pour inférer la position des hôtes cibles. La section 4 illustre les différents résultats obtenus en appliquant notre architecture hybride de géolocalisation. Enfin, la section 5 conclut notre travail et présente quelques perspectives de recherche.

2 Système de géolocalisation hybride

2.1 Architecture hybride

La figure (Fig. 1) illustre les différents composants de notre système de géolocalisation hybride. Ce système peut être décomposé comme suit :

- Une base de données qui contient des préfixes d'adresses IP et leur propre information de localisation. Ces informations de localisation, selon les préfixes d'adresses IP répertoriés dans la base, peuvent être à l'échelle d'un pays, d'une région, d'une ville, ou bien sous forme de coordonnées géographiques (latitude, longitude).
- Un serveur où est implémenté l'heuristique qui détermine si des mesures doivent être faites vers l'hôte cible, et si tel est le cas, avec quel ensemble d'hôtes références.
- Des mesures actives (mesures de délai) qui sont faites à partir d'un ensemble d'hôtes références prédéfini.

Le processus de localisation d'un hôte cible à partir de notre système hybride de géolocalisation est expliqué plus en détail dans la section 3.

2.2 Structure de la base de données utilisée

Dans un premier temps, quand une requête de localisation arrive au niveau du système illustré sur la figure Fig. 1, le serveur interroge la base de données pour connaître la position géographique de cet hôte cible. La base de données étudiée ici, est celle qu'utilise *GeoIP* [18]. Nous avons utilisé la version commerciale que fournit *MaxMind* [9]. La base de données de GeoIP est la plus utilisée actuellement.

En effet, la base de données contient des préfixes d'adresses IP et leur information de localisation. Elle est constituée de deux tables, *table 1* et *table 2*, comme illustré dans le tableau (Tab. 1). Dans la base de données, chaque entrée de la table 1 contient la valeur du préfixe d'adresses IP et son numéro d'identification (voir tableau (Tab. 1)). Au niveau de la table 2, chaque entrée contient l'identifiant du préfixe d'adresses IP et les informations de localisation qui lui sont associées suivant une granularité plus fine (voir tableau (Tab. 1)).

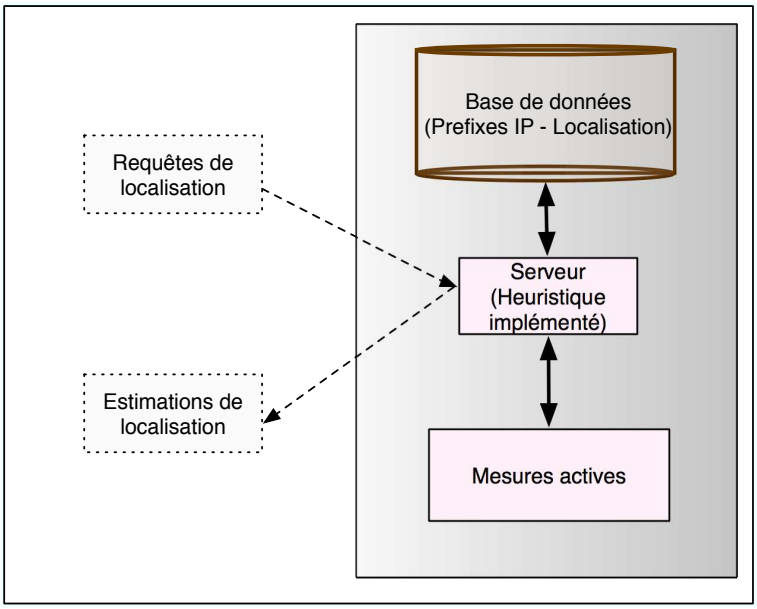


Fig. 1. Architecture d'un système de géolocalisation hybride.

Tab. 1. Bases de données des préfixes d'adresses IP.

table 1	
préfixe d'adresses IP	**loc id**

table 2						
loc id	**pays**	**région**	**ville**	**code postal**	**latitude**	**longitude**

Ensuite, une tabulation exhaustive, comme dans [8, 18–20], permet de trouver, s'il existe dans la base de données, le préfixe d'adresses IP auquel appartient l'hôte cible. En connaissant le préfixe d'adresses IP auquel appartient l'hôte cible, nous déduisons, à partir de la table 2 de la base de données, la localisation géographique de l'hôte cible. Il faut noter que, la tabulation exhaustive donne comme réponse le préfixe d'adresses auquel appartient la cible ayant le plus long préfixe dans la base de données.

Ainsi, après avoir obtenu la localisation du préfixe d'adresses IP de l'hôte cible, nous appliquons l'heuristique qui permet de définir un sous-ensemble, dans notre ensemble d'hôtes références $\mathcal{L}$, qui va procéder à la localisation de l'hôte cible. Par contre, si l'adresse IP de l'hôte cible n'appartient à aucun préfixe d'adresses IP de la base de données, *i.e* le préfixe d'adresses IP n'est pas enregistré dans la base, alors la localisation de l'hôte cible se fait avec tous les hôtes références de notre ensemble.

3 Heuristique du choix des hôtes références

Comme illustré au niveau de l'architecture de notre système hybride de géolocalisation (Fig. 1), le serveur implémente une heuristique qui lui permet de choisir l'ensemble des hôtes références qui feront les mesures vers l'hôte cible. Pour ce faire, le serveur envoie une requête à la base de données pour connaître la localisation géographique de cet hôte cible. Si cette localisation existe, alors la base de données lui renvoie la latitude et la longitude du préfixe d'adresses IP auquel appartient cet hôte cible.

L'heuristique consiste, pour un nombre d'hôtes références fixé, à choisir les hôtes références les plus proches à l'hôte cible en terme de distance géographique. Il faut noter que la position géographique (latitude, longitude) de tous les hôtes références qui constituent notre ensemble $\mathcal{L}$ est connue. Connaissant la position géographique du préfixe d'adresses de l'hôte cible, grace à la base de données, et la position géographique des hôtes références, nous calculons la distance géographique entre les hôtes références et l'hôte cible. En se basant sur [21], la distance géographique, entre chaque hôte référence L_i et l'hôte cible τ, est donnée par

$$\beta = \sqrt{\left(\sin\left(\frac{lat_i - lat_\tau}{2}\right)\right)^2 + \cos(lat_i) \times \cos(lat_\tau) \times \left(\sin\left(\frac{lon_\tau - lon_i}{2}\right)\right)^2} \quad (1)$$

$$\hat{dist}_{i\tau} = 6371 \times 2 \times \arcsin(\beta) \quad (2)$$

Dans l'équation 1, lat_i, et lon_i représentent la latitude et la longitude, exprimées en radian, de l'hôte référence L_i ; lat_τ et lon_τ représentent la latitude et la longitude, exprimées en radian, de l'hôte cible τ. La distance géographique, exprimée en km, entre l'hôte référence L_i et l'hôte cible τ, est obtenue à partir de l'équation 2. Le terme 6371, présent dans l'équation 2, représente le rayon de la terre. En effet, l'expression $2 \times \arcsin(\beta)$ fournit la distance géographique en radian. Dans la section 4, ce sont les distances géographiques exprimées en km qui sont considérées.

Pour l'hôte cible τ, nous obtenons le vecteur de distance

$$D_\tau = [\hat{dist}_{1\tau}, \hat{dist}_{2\tau}, \ldots, \hat{dist}_{K\tau}], \quad (3)$$

où K représente le nombre d'hôtes références total de notre ensemble $\mathcal{L}$ et $\hat{dist}_{i\tau}$ représente la distance géographique, en km, calculée entre l'hôte référence L_i et la cible τ pour $1 \leq i \leq K$.

Si nous fixons par exemple un nombre n d'hôtes références parmi les K que compte notre ensemble $\mathcal{L}$, les n hôtes références ayant les plus petits $\hat{dist}_{i\tau}$, $1 \leq i \leq n$, sont choisis pour inférer la position de l'hôte cible. Nous montrons dans la section 4.2, que pour un nombre réduit d'hôtes références, avec cette heuristique, nous obtenons de meilleurs résultats, de surcroît avec moins de mesures injectées dans le réseau.

4 Évaluation

4.1 Paramètres expérimentaux

Pour évaluer notre heuristique, nous avons considéré un ensemble de données constitué par des hôtes AMP [22] et des hôtes RIPE [23] comme hôte cible. Ces hôtes sont

au nombre de 127 et sont localisés essentiellement aux États Unis et en Europe. La principale raison de cette restriction est due au fait que nous avons besoin d'hôtes cibles dont on connaît la localisation géographique pour pouvoir évaluer l'erreur d'estimation. Ainsi, seuls les hôtes AMP et RIPE fournissent cette exigence.

Nous avons utilisé la base de données qu'emploie *GeoIP* [18]. GeoIP est une technologie propriétaire, un outil commercial, appartenant à l'organisation *MaxMind* [9]. Nous avons utilisé la version commerciale de GeoIP. La différence entre la version gratuite et la version commerciale de GeoIP est au niveau de la granularité de l'information de localisation fournie. Dans la base de données gratuite de GeoIP, l'information de localisation est fournie seulement à l'échelle d'un pays. Cette base de données est formée par des préfixes d'adresses IP et chaque préfixe d'adresses IP possède une information de localisation suivant plusieurs granularités comme le pays, la région, la ville ou latitude/longitude (Tab. 1). La base de données contient 1873596 préfixes d'adresses IP et le masque de ces préfixes d'adresse varie entre 8 et 32.

Pour inférer la localisation de ces hôtes cibles, nous avons considéré un ensemble de 74 nœuds PlanetLab [24] comme hôtes références. Les points illustrés sur la figure (Fig. 2) représentent la distribution géographique de ces hôtes références.

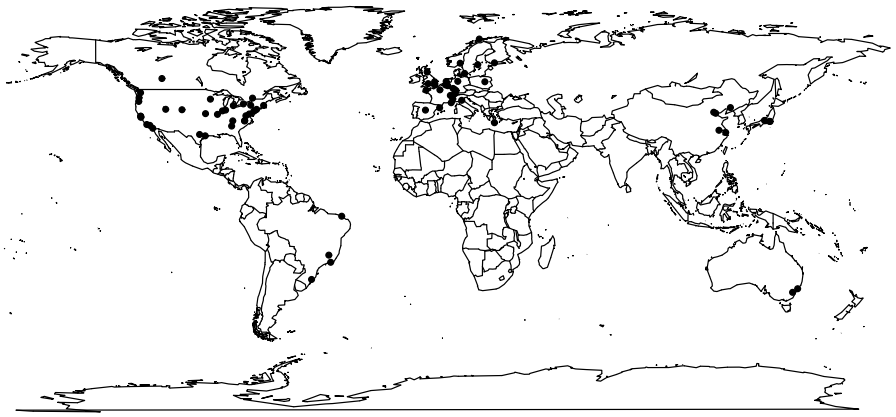


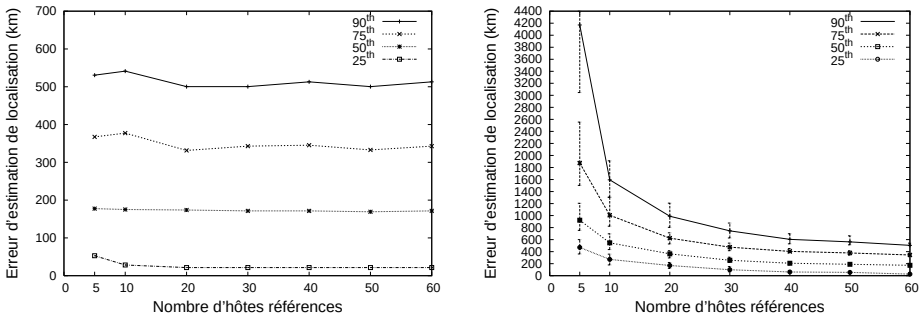
Fig. 2. Distribution géographique des hôtes références.

Nous avons fait les mesures de délai entre les hôtes références et les hôtes cibles. Ainsi, chaque hôte référence exécute des pings vers les hôtes cibles pour mesurer le délai entre eux. Chaque mesure de ping, entre un hôte référence L_i et un hôte cible τ, est composé de 10 paquets sondes envoyés par intervalle de 1 seconde. Nous espaçons les paquets sondes pour que nos mesures ne soient pas considérées comme des attaques de déni de service. Chaque paquet envoyé a une taille de 1024 Ko. Seul le RTT minimum est considéré pour chaque mesure de ping. Il est plus vraisemblable, que le RTT minimum reflète le mieux le délai de propagation et qu'il soit le moins assujetti aux congestions et autres sources de distorsions. Ainsi, nous avons considéré le RTT minimum entre chaque hôte référence et chaque hôte cible de notre ensemble de données.

Pour inférer la localisation géographique des hôtes cibles nous utilisons l'approche CBG décrite dans le papier [4].

4.2 Résultats

La figure (Fig. 3) montre l'impact du choix des hôtes références et de leur nombre sur les performances de CBG. Le choix des hôtes références, pour localiser les hôtes cibles, se fait soit de manière aléatoire ou, suivant l'heuristique étudiée dans la section 3.



(a) Choix des hôtes références par heuristique

(b) Choix aléatoire des hôtes références

Fig. 3. Erreur d'estimation de localisation en fonction du nombre d'hôtes références.

La figure (Fig. 3(a)) montre différents centiles de l'erreur d'estimation de localisation obtenue en fonction du nombre d'hôtes références considérés en utilisant la technique CBG. Le choix des hôtes références, pour un nombre k d'hôtes références fixé, se fait suivant l'heuristique présentée dans la section 3. L'axe des abscisses représente le nombre d'hôtes références choisi dans notre ensemble composé de 74 hôtes références, pour inférer la localisation d'un hôte cible. Le nombre d'hôtes références varie entre 5 et 60. L'axe des ordonnées représente l'erreur d'estimation de localisation obtenue pour un nombre k d'hôtes références utilisés. Par exemple, la courbe qui montre les 90ème de centiles, représente l'erreur d'estimation de localisation, où la courbe de la fonction de probabilité cumulative de l'erreur moyenne d'estimation de localisation rencontre le point dont la probabilité est 90%. Nous remarquons qu'à partir de 20 hôtes références considérés (Fig. 3(a)), l'erreur d'estimation de localisation reste stable. Toutefois, pour la courbe des 90 et 75 centiles nous notons une légère augmentation de l'erreur d'estimation de localisation si l'on augmente le nombre d'hôtes références considérés. Ceci est certainement dû à la présence de "bruit" (distorsions) au niveau de nos mesures de délai induit par les hôtes références ajoutés, et qui sont un peu éloignés de la cible. Plus les hôtes références sont proches géographiquement de la cible meilleur est l'estimation de localisation (Fig. 3(a)). En ne considérant que les 20 hôtes références les plus proches des hôtes cibles, 50% des hôtes cibles sont localisés avec une erreur inférieure à 175 km (voir Fig. 3(a)).

La figure (Fig. 3(b)) montre différents centiles de l'erreur d'estimation de localisation en fonction du nombre d'hôtes références considérés. Le choix des hôtes références,

pour un nombre k d'hôtes références fixé, s'est fait de manière aléatoire. Nous avons considéré 30 échantillons de mesures pour chaque k hôtes références choisis. Le nombre d'hôtes références considéré varie entre 5 et 60. Toutefois, comme le nombre de possibilité de placement des hôtes références devient de plus en plus important lorsque k augmente, nous n'avons pas considéré toutes les façons de choisir k hôtes références dans chaque ensemble de données.

Les barres sur la figure (Fig. 3(b)) représentent les intervalles de confiance, pour l'ensemble des échantillons de mesures considérées, pour un nombre d'hôtes références fixé. Nous remarquons qu'à partir de 30 hôtes références, l'erreur d'estimation de localisation est pratiquement stable. Toutefois, avec un choix aléatoire des hôtes références, l'erreur d'estimation de localisation reste assez importante Ainsi, en considérant 30 hôtes références choisis aléatoirement, 50% des hôtes cibles sont localisés avec une erreur d'estimation inférieure à 400 km. En considérant notre heuristique, pour ce même nombre d'hôtes cibles et d'hôtes références, l'erreur d'estimation de localisation est inférieure à 175 km (Fig. 3(a)).

Ainsi, en considérant un nombre restreint d'hôtes références pour localiser les hôtes cibles, nous réduisons le temps de traitement nécessaire à CBG pour traiter une requête de géolocalisation. Par conséquent le temps de réponse est fortement diminué.

5 Conclusion

Ce papier a proposé et a évalué une technique utilisant à la fois des mesures actives et passives pour inférer la position géographique des hôtes Internet. Nous avons mis en place un système hybride de géolocalisation qui permet d'utiliser une base de données (mesure passive) que l'on associe à des mesures de délai (mesure active). Ainsi, grace à la base de données, il est possible d'inférer la position de l'hôte cible avec les hôtes références qui lui sont le plus proche géographiquement. En effet, l'heuristique que nous avons développée permet de choisir les hôtes références les plus proches géographiquement de la cible.

Les résultats obtenus montrent que, si l'on choisi les hôtes références les plus proches géographiquement de la cible, pour le localiser, nous obtenons une meilleure estimation de localisation comparé au choix aléatoire des hôtes références. En ne considérant que les 20 hôtes références les plus proches des hôtes cibles (choix par heuristique), 50% des hôtes cibles sont localisés avec une erreur inférieure à 175 km. En outre, 20 hôtes références suffisent pour stabiliser l'erreur d'estimation de localisation. La combinaison de la base de données avec les mesures de délai permet de réduire le nombre d'hôtes références utilisés et par ailleurs réduire le trafic injecté dans le réseau ainsi que le temps de réponse pour inférer la position de la cible.

Cependant, la mise-à-jour des bases de données reste difficile à faire. Malgré l'utilisation de la base de données pour choisir les hôtes références les plus proches à l'hôte cible, nous notons une erreur d'estimation au niveau de la localisation des hôtes cibles. Nous envisageons de déployer cette technique hybride sur PlanetLab mais aussi d'évaluer la précision de la base de données.

Références

1. V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *Proc. ACM SIGCOMM*, August 2001.

2. M. Freedman, M. Vutukurum, N. Feamster, and H. Balakrishnan, "Geographic locality of IP prefixes," in *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, October 2005.
3. B. Wong, I. Stoyanov, and E. G. Sirer, "Geolocalization on the Internet through constraint satisfaction," in *Proc. USENIX Workshop on Real, Large Distributed Systems (WORLDS)*, Seattle, WA, November 2006.
4. 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, December 2006.
5. GeoBytes Inc., "GeoNetMap - geobytes' IP address to geographic location database," http ://www.geobytes.com/GeoNetMap.htm.
6. Qwerks Inc., "WhereIsIP - IP whois tool," http ://www.jufsoft.com/whereisip.
7. Hexasoft Development Sdn. Bhd, "IP address geolocation to identify website visitor's geographical location," http ://www.ip2location.com.
8. Quova Inc., "GeoPoint - IP geolocation experts," http ://www.quova.com.
9. MaxMind, "Geolocation and online fraud prevention from MaxMind," http ://www.maxmind.com/.
10. "Host IP," http ://www.hostip.info.
11. IPInfoDB, "Free IP address geolocation tools," http ://ipinfodb.com/.
12. Software 77, "Free IP to country database," http ://software77.net/geo-ip/.
13. E. Katz-Bassett, J. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, October 2006.
14. B. Gueye, S. Uhlig, A. Ziviani, and S. Fdida, "Leveraging buffering delay estimation for geolocation of Internet host," in *Proc. IFIP Networking Conference*, Coimbra, Portugal, May 2006.
15. S. Siwpersad, B. Gueye, and S. Uhlig, "Assessing the geographic resolution of exhaustive tabulation for geolocating Internet hosts," in *Proc. of PAM*, Cleveland, Ohio, USA, April 2008.
16. B. Gueye, S. Uhlig, and S. Fdida, "Investigating the imprecision of IP block-based geolocation," in *Proc. Passive and Active Measurement Conference (PAM)*, Louvain-la-neuve, Belgium, April 2007.
17. C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang, "Mining the web and the Internet for accurate IP address," in *Proc. IEEE INFOCOM*, April 2009.
18. "MaxMind LLC," http ://www.maxmind.com/geoip.
19. "Akamai Inc.," http ://www.akamai.com.
20. "GeoURL," http ://www.geourl.org.
21. *Ed Williams*
, http ://http ://williams.best.vwh.net/avform.html.
22. *NLANR Active Measurement Project, 1998*, http ://watt.nlanr.net/.
23. *RIPE Test Traffic Measurements, 2000*, http ://www.ripe.net/ttm/.
24. PlanetLab, "An open platform for developing, deploying, and accessing planetary-scale services," 2002, http ://www.planet-lab.org.