GeoBuD: Une nouvelle approche pour la localisation géographique des hôtes Internet

Bamba Gueye[†] and Steve Uhlig [‡] and Artur Ziviani [§] and Serge Fdida [¶]

Actuellement, toutes les techniques de géolocalisation basées sur des mesures ne tiennent pas compte de l'influence du **buffering** (temps de traitement). Ce délai peut être introduit par les routeurs traversés par les paquets sondes. Pour combler cette lacune, nous proposons l'approche **GeoBuD**. Rechercher l'influence du délai de buffering, induit par chaque routeur, sur la localisation est un véritable challenge. En effet, pour estimer avec précision le délai de buffering au niveau des sauts intermédiaires qui composent le chemin suivi par un **traceroute**, il faut localiser avec exactitude les routeurs intermédiaires. En outre, l'estimation du délai de buffering reste difficile même avec la connaissance de la position géographique des routeurs. Ceci est dû au fait que les mesures de délai contiennent une information grossière. En tenant compte du buffering, nous montrons que l'on augmente la précision de l'estimation de localisation ainsi que la zone de confiance associée à chaque estimation de localisation.

Keywords: Géolocalisation, multilatération, estimation du délai de buffering, traceroute

1 Introduction

La localisation géographique des hôtes dans l'Internet à partir de leur adresse IP connaît un intérêt croissant. Elle permet le développement de nouvelles applications capables d'offrir des services inédits [Geo, Max]. Nous pouvons citer comme exemple la localisation des cyber-criminels, la publicité ciblée, la diffusion restreinte de contenu suivant les régulations locales ou les préférences régionales, et l'acceptation d'une transaction bancaire seulement à partir d'un endroit pré-défini. Comme les adresses IP sont allouées en général de manière arbitraire, il n'existe pas de relation évidente entre une adresse IP et la position géographique de l'équipement qui possède cette adresse. Ainsi, inférer la localisation géographique d'un hôte dans l'Internet est un véritable défi.

Les précédentes techniques de géolocalisation [PS01, GZCF06] se basent sur les mesures des *hôtes références* (hôtes dont on connaît la position géographique), pour fournir la position géographique de l'hôte cible. La technique *GeoPing* [PS01] utilise la position géographique de l'hôte référence le plus proche par rapport à l'hôte cible, en terme de délai, comme sa possible localisation. Avec cette approche, l'ensemble des endroits où on peut localiser un hôte cible est limité par le nombre d'hôtes références. Il en découle un ensemble discret de réponses. Cette approche limite la précision de l'estimation car le placement et le nombre d'hôtes références ont un impact sur l'exactitude de la localisation. L'approche "*Constraint-Based Geolocation*" (*CBG*), proposée par certains auteurs de cet article dans [GZCF06], transforme les mesures de délai entre chaque hôte référence et l'hôte cible, en distances appelées *distances géographiques surestimées*. Ces distances sont appelées ainsi car étant formées par les distances géographiques réelles et les distances induites par les délais supplémentaires qui s'ajoutent aux mesures. CBG utilise ensuite la *multilatération* pour estimer la position géographique de l'hôte cible. La multilatération permet d'estimer une position en utilisant un nombre suffisant de distances à partir de quelques points fixes.

Dès lors, la multilatération fournit un ensemble continu d'endroits où peut être localisé la cible au lieu d'un espace discret de réponses. En appliquant la multilatération, CBG déduit la zone géographique dans

[†]Université de Liège, Institut Montefiore (B28)
[‡]Delft University of Technology, Network Architectures and Services
[§]Laboratório Nacional de Computação Cientifica (LNCC)
[¶]Université Pierre et Marie Curie, Laboratoire d'Informatique de Paris 6 UMR 7606

laquelle se trouve la cible. La surface de cette zone géographique représente la zone de confiance associée à chaque estimation de localisation. Le centre de cette zone est considérée comme l'estimation de la localisation de l'hôte cible.

Ces différentes techniques de géolocalisation souffrent des distorsions que subissent les mesures de délai. Ces sources de distorsions sont dues, par exemple, à la congestion qui survient dans les réseaux et à la non-linéarité des chemins, ajoutant ainsi un délai supplémentaire dans les mesures de délais [TGSE01]. Le temps de traitement des paquets (buffering) au niveau des routeurs intermédiaires sur un chemin (source, destination) peut être aussi un facteur de distorsion. Ainsi, pour accroître la précision des techniques de géolocalisation basées sur des mesures, il est important d'estimer et de supprimer ce délai additionnel.

Dans cet article, nous examinons l'impact du délai de buffering dans les mesures de délai ainsi que les solutions envisagées pour y remédier. Notre contribution est de proposer une nouvelle méthode de géolocalisation, *GeoBuD*. Nous employons l'outil traceroute pour estimer le buffering introduit dans le délai mesuré entre chaque hôte référence et l'hôte cible. En considérant les différents *RTT* (Round Trip Time) mesurés au niveau des sauts intermédiaires découverts grâce à l'outil traceroute, nous donnons une estimation du délai buffering introduit à chaque noeud. Les résultats obtenus montrent que la prise en compte de ce buffering permet d'améliorer la précision de l'estimation fournie par CBG. Cette amélioration vient de la réduction de la surestimation de la distance géographique entre les hôtes référence et la cible. Ainsi, la zone géographique dans laquelle CBG infère l'hôte cible est rétrécie, conduisant à une meilleure estimation de la localisation.

2 Estimation du délai de buffering par l'outil traceroute

Contrairement à CBG [GZCF06] qui se base sur un calibrage entre les hôtes références pour transformer les mesures de délai en distances géographiques, GeoBuD se base sur le chemin entre les hôtes références et l'hôte cible. En effet, dans la pratique, le délai vers différentes destinations peut contenir différentes sources de distorsions. Pour en tenir compte, GeoBud modélise le délai $y_{i\tau}$ par

$$y_{i\tau} = m_i x_{i\tau} + b_{i\tau}, \tag{1}$$

où m_i représente la vitesse de propagation des données mesurées entre les hôtes références, x_{i_τ} représente la distance géographique *surestimée* entre l'hôte référence L_i et la cible τ et $b_{i\tau}$ représente le délai de buffering estimé sur le chemin entre L_i et l'hôte cible τ. Estimer le délai de buffering sur le chemin revient à l'estimer au niveau de chaque nœud entre L_i et τ. Pour ce faire, nous utilisons l'outil traceroute qui fournit le RTT vers chaque nœud intermédiaire ayant répondu par un message ICMP `TIME exceeded` sur le chemin vers une destination. Pour estimer $b_{i\tau}$ dans l'équation (1), nous estimons en fait ses différents composants b_k le long du chemin suivi par le traceroute :

$$\Delta RTT_{k+1} = RTT_{k+1} - RTT_k = m_i \times dist(k,k+1) + b_{k+1}, \tag{2}$$

où k représente le $k^{\text{ème}}$ routeur intermédiaire sur le chemin du traceroute pour lequel nous avons une mesure de délai et dont la localisation géographique est connue. Le terme RTT_k désigne le RTT minimum mesuré à un saut donné k. *dist(k,k+1)* représente la distance géographique entre le nœud k et $k+1$. La somme des $dist(k,k+1)$ pour k allant de 0 à $n-1$ donne l'estimation de la distance géographique du chemin suivi par le traceroute entre un hôte référence L_i et une cible τ. Ainsi, à partir de l'équation (2), le délai de buffering b_k à chaque saut est obtenu grâce à la formule $b_k = \Delta RTT_k - m_i \times dist(k-1,k)$.

Si on veut estimer b_k, il est évident qu'il faut d'abord estimer la distance géographique entre chaque paire de routeurs adjacents sur le traceroute.

3 Évaluation de GeoBuD

Pour estimer le délai de buffering b_k et la distance $dist(k,k+1)$ au niveau de chaque saut composant le chemin entre chaque paire hôte référence et hôte cible, nous avons utilisé deux ensembles de données :

– Premièrement, nous avons considéré 29 nœuds de PlanetLab comme hôtes références et 87 nœuds AMP [AMP98] comme hôtes cibles, tous localisés aux États Unis. Les données utilisées sont composées par les mesures de traceroute faites par les hôtes références vers les hôtes cibles durant la journée du 17 Octobre 2005.
– Le deuxième ensemble de données est formé par les traceroutes effectués à partir de 27 nœuds de PlanetLab vers 57 nœuds RIPE [RIP00] localisés en Europe Occidentale. Ces mesures ont été effectuées le 21 Novembre 2005.

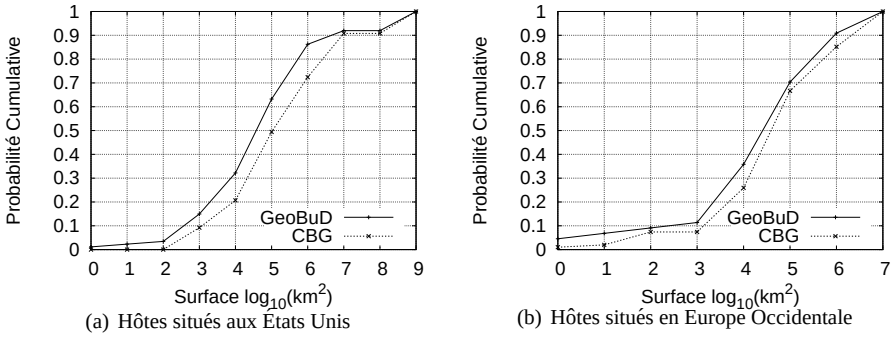
En considérant l'ensemble de données constitué par les hôtes localisés aux États Unis, sans les hôtes AMP, nous avons pu localiser 1153 routeurs intermédiaires sur un total de 1408 routeurs obtenus grâce aux mesures de traceroute. Pour les hôtes localisés en Europe Occidentale, sans les hôtes RIPE, nous avons localisé 1235 routeurs sur un total de 1328. Pour la localisation géographique de ces routeurs intermédiaires, nous avons utilisé le projet *GeoLIM* [Geo05] qui est une implémentation de l'approche CBG. Nous avons également fait un recoupement des résultats obtenus par GeoLIM avec l'outil *rockettrace*.

Après avoir localisé les routeurs intermédiaires, nous calculons la valeur des b_k sur les différents segments du traceroute composés par ces routeurs en utilisant l'équation (2). Nous avons obtenu 21% de b_k négatifs, 4043 b_k sur un nombre total de 19172 b_k calculés en localisant les hôtes AMP. Pour les hôtes RIPE, le pourcentage de b_k négatifs obtenu est de 14% pour 11908 b_k trouvés. Ainsi, un b_k est considéré si et seulement si $\Delta RTT_{k+1} > 0$. GeoBuD utilise la distance ($x_{i\tau}$) obtenue à partir de l'équation (1) pour localiser un hôte cible donné. Les distances géographiques obtenues par GeoBuD sont supposées être des bornes supérieures plus strictes sur la distance réelle que celles obtenues par la méthode CBG. En considérant ces nouvelles distances, et malgré le nombre de b_k négatifs, la zone de confiance obtenue avec la méthode GeoBuD est ainsi rétrécie.

3.1 Réduction de la zone de confiance par GeoBuD

La figure 1 compare la probabilité cumulative de la taille de la zone de confiance obtenue par GeoBuD et CBG. L'axe des abcisses représente la surface de la zone de confiance associée à chaque estimation de localisation. L'axe des ordonnées montre la probabilité que l'estimation de la localisation ait une zone de confiance inférieure à une valeur x sur l'axe des abcisses.

En tenant compte du délai de buffering nous observons une nette amélioration pour les zones de confiance inférieures à 10^7 km^2. En utilisant l'approche CBG, 72% des hôtes situés aux États Unis sont localisés avec une zone de confiance inférieure à 10^6 km^2. Pour cette même zone de confiance, nous localisons avec GeoBuD environ 86% des hôtes. En outre, avec CBG, 49% des hôtes cibles ont une zone de confiance inférieure à 10^5 km^2. Pour cette même zone de confiance, 63% des hôtes cibles sont localisés par GeoBuD. Pour les hôtes situés en Europe Occidentale, GeoBuD en localise 10% avec une zone de confiance inférieure à 10^2 km^2. Notons qu'une surface de 10^5 km^2 équivaut à peu prés à la superficie d'un pays comme le Portugal ou bien d'un état des États Unis comme l'Indiana.

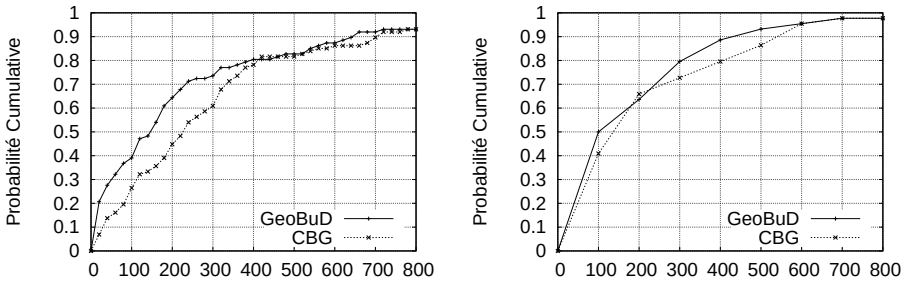


(a) Hôtes situés aux États Unis (b) Hôtes situés en Europe Occidentale

FIG. 1: Zone de confiance fournie par GeoBuD et CBG en km^2.

3.2 Erreur d'estimation de localisation

Nous nous attendons à ce que la réduction de la zone de confiance puisse avoir un impact sur l'erreur d'estimation de l'hôte cible. La figure 2 illustre la probabilité cumulative de l'erreur obtenue pour chaque estimation de localisation. L'erreur d'estimation est la différence entre la localisation géographique réelle de l'hôte cible et son estimation de localisation. Avec GeoBuD, 80% des hôtes situés aux États Unis présentent une erreur d'estimation moins importante comparé à CBG. Ainsi, l'erreur médiane est de 144 km en utilisant GeoBuD alors qu'elle est de 228 km pour CBG. Pour les hôtes situés en Europe Occidentale, elle est de 100 km pour GeoBuD et de 137 km pour CBG.



(a) Hôtes situés aux États Unis (b) Hôtes situés en Europe Occidentale

FIG. 2: Erreur d'estimation de localisation pour GeoBuD et CBG.

4 Conclusion

Cet article montre qu'en tenant compte du délai de buffering estimé au niveau des sauts intermédiaires entre un hôte référence et un hôte cible, nous pouvons améliorer la précision de l'estimation de la localisation des hôtes dans l'Internet. En se basant sur des mesures de traceroute, nous estimons le délai de buffering au niveau des sauts intermédiaires. En associant cette estimation du délai de buffering avec une technique basée sur la multilatération (CBG [GZCF06]), nous sommes capables de réduire la zone de confiance où l'hôte cible est localisé. Les résultats montrent que GeoBuD obtient aussi une meilleure précision de l'estimation de la localisation.

Nous envisageons une implémentation de l'approche GeoBuD à l'image de l'outil GeoLIM [Geo05]. Nous visons à converger vers des zones de confiance aussi petites que possible afin de fournir une meilleure estimation de localisation. Un raffinement du délai de buffering est aussi envisagé en considérant une possible existence d'un goulot d'étranglement sur le chemin.

Références

[AMP98] *NLANR Active Measurement Project*, 1998. `http://watt.nlanr.net/`.

[Geo] Geobytes, Inc. *GeoNetMap*. `http://www.geobytes.com/GeoNetMap.htm`.

[Geo05] *GeoLIM Project*, 2005. `http://planetlab-01.ipv6.lip6.fr:10000/cbg.php/`.

[GZCF06] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-based geolocation of internet hosts. *IEEE/ACM Transactions on Networking*, 14(6) :1219–1232, December 2006.

[Max] MaxMind LLC. *GeoIP*. `http://www.maxmind.com/geoip/`.

[PS01] V. N. Padmanabhan and L. Subramanian. An investigation of geographic mapping techniques for Internet hosts. In *Proc. ACM SIGCOMM*, San Diego, CA, USA, August 2001.

[RIP00] *RIPE Test Traffic Measurements*, 2000. `http://www.ripe.net/ttm/`.

[TGSE01] H. Tangmunarunkit, R. Govindan, S. Shenker, and D. Estrin. The impact of routing policy on internet paths. In *Proc. IEEE INFOCOM*, Anchorage, AK, USA, April 2001.