CBG: Geolocalização na Internet usando medições de atraso

Artur Ziviani[1], Bamba Gueye[2], Mark Crovella[3], Serge Fdida[2]

[1]Laboratório Nacional de Computação Científica (LNCC)
Av. Getúlio Vargas, 333 – 25651-075 – Petrópolis, RJ

[2]Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie (Paris 6)
8, rue du Capitaine Scott – 75015 – Paris, França

[3]Department of Computer Science
Boston University
111 Cummington St – 02215 – Boston, MA, EUA

ziviani@lncc.br, gueye@rp.lip6.fr, crovella@cs.bu.edu, fdida@rp.lip6.fr

Abstract. *Geolocation of Internet hosts enables a diverse and interesting 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, thus establishing a continuous space of answers instead of a discrete one. 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 the previous geolocation techniques. Moreover, in contrast to previous approaches, our method is able to assign a confidence region to each given location estimate.*

Resumo. *A geolocalização de nós na Internet permite o surgimento de uma variada e interessante nova classe de aplicações conscientes de localização. Abordagens anteriores baseadas em medições usam nós de referência, com posição geográfica bem conhecida, para fornecer uma estimativa de localização para um nó-alvo. Isto leva a um espaço discreto de respostas, limitando o número de possíveis estimativas de localização ao número de nós de referência adotados. Contrastando com isto, este artigo propõe CBG (Constraint-Based Geolocation) para a geolocalização de nós na Internet. CBG infere a localização geográfica de nós na Internet usando multilateração com restrições de distância, estabelecendo assim um espaço contínuo de respostas ao invés de um espaço discreto. CBG transforma de forma acurada medições de atraso em restrições de distância geográfica, as utilizando então em multilateração para inferir a geolocalização do nó-alvo. Nossos resultados experimentais mostram que CBG apresenta um melhor desempenho do que técnicas anteriores de geolocalização. Além disto, em contraste com as demais propostas anteriores, nosso método é capaz de associar uma região de confiabilidade para cada estimativa de localização.*

1. Introdução

O desenvolvimento de uma maneira eficiente de inferir a localização geográfica de nós na Internet abre perspectivas para novas aplicações conscientes da localização dos usuários [Zook, 2001, Lakhina et al., 2003]. Exemplos destas aplicações incluem a publicidade direcionada em páginas web, a seleção automática do idioma para a apresentação de um determinado conteúdo, a distribuição restrita de conteúdo seguindo normas regionais e a autorização de transações somente quando realizadas a partir de locais pré-estabelecidos. No entanto, a inferência da localização geográfica de nós na Internet a partir de seus endereços IP constitui um problema desafiador, pois não há uma relação direta entre o endereço IP de um nó e a sua localização geográfica. Trabalhos anteriores [Padmanabhan e Subramanian, 2001, Ziviani et al., 2004] usam a posição de nós de referência, que possuem localização geográfica bem conhecida, como possíveis estimativas de localização para o nó-alvo. Isto leva a um espaço discreto de respostas, ou seja, o número de respostas equivale ao número de nós de referência, o que pode levar a resultados inacurados porque o nó de referência mais próximo pode estar afastado do alvo.

Para superar a limitação de um espaço discreto de respostas, nós propomos a abordagem CBG (*Constraint-Based Geolocation*)[1], que infere a localização geográfica de nós na Internet usando multilateração. Multilateração refere-se ao processo de estimar uma posição usando um número suficiente de distâncias a alguns pontos fixos. Como resultado, a multilateração estabelece um espaço contínuo de respostas no lugar de um espaço discreto. Nós utilizamos um conjunto de nós de referência para estimar a localização de outros nós na Internet. A idéia fundamental é que dadas as distâncias geográficas até um determinado nó-alvo a partir dos nós de referência, uma estimativa de localização do nó-alvo seria viável usando multilateração, assim como faz o sistema GPS (*Global Positioning System*) [Enge e Misra, 1999].

Um elemento-chave de CBG é a sua habilidade em transformar de forma acurada medições de atraso em restrições de distância. O ponto de partida consiste na constatação de que a informação digitalizada trafega por cabos de fibra ótica a quase exatamente 2/3 da velocidade da luz no vácuo [Bovy et al., 2002, Percacci e Vespignani, 2003]. Isto significa que qualquer medição de atraso fornece imediatamente um *limite superior* na distância entre os pontos finais. Este limite superior é a medição de atraso multiplicada pela velocidade da luz na fibra. Do ponto de vista de um par de pontos finais, nós argumentamos que há algum atraso mínimo teórico para a transmissão de pacotes que é ditado pela distância geográfica entre eles. Portanto, o atraso real medido entre estes pontos envolve somente uma distorção *aditiva*.

Contudo, se CBG usasse as medições de atraso para inferir diretamente as restrições de distância, a proposta não seria muito acurada. Para resultados acurados, é importante estimar e remover o tanto quanto for possível da distorção aditiva. CBG realiza esta tarefa auto-calibrando as medições de atraso tomadas de cada ponto de medida. Isto é feito de uma forma distribuída como explicado com mais detalhes na Seção 3. Após a auto-calibração, CBG é capaz de transformar de forma mais acurada um conjunto de medições de atraso até um alvo em restrições de distância. CBG então usa multilateração com estas restrições de distância para estabelecer uma região geográfica que contenha o nó-alvo. Determinada esta região, uma estimativa razoável da localização do nó-alvo é o centróide desta região, o que é usado por CBG como estimativa pontual da posição do alvo. Deve-se ressaltar que, em contraste com outras propostas, CBG associa uma região de confiabilidade para cada estimativa de localização. Isto permite a uma aplicação consciente de localização avaliar se a qualidade da estimativa fornecida é suficiente às suas necessidades.

Nós avaliamos nossa proposta CBG usando bases de dados com nós que estão geograficamente distribuídos pelos EUA e pela Europa Ocidental. Nossos resultados experimentais são

[1] Este artigo é uma versão estendida do artigo curto [Gueye et al., 2004].

promissores e mostram que CBG supera em desempenho outras técnicas de geolocalização. A mediana do erro em distância está abaixo de 25 km para os dados da Europa e abaixo de 100 km para os dados dos EUA. Para a maioria dos nós-alvo analisados, as regiões de confiabilidade obtidas permitem uma resolução em nível regional, ou seja, aproximadamente o tamanho de estados brasileiros relativamente pequenos em extensão territorial, como Rio de Janeiro ou Santa Catarina.

Este artigo está organizado da seguinte forma. A Seção 2 revisa os trabalhos relacionados à geolocalização e destaca as contribuições da proposta CBG em relação a estes trabalhos. Na Seção 3, nós introduzimos a proposta CBG. Em seguida, nós apresentamos na Seção 4 nossos resultados experimentais. Na Seção 5, nós discutimos alguns aspectos pertinentes às tecnologias de geolocalização. Finalmente, nós concluímos na Seção 6.

2. Geolocalização de nós na Internet

2.1. Motivação

A ampla disponibilidade de informação de localização permite o desenvolvimento de aplicações conscientes de localização que podem ser úteis tanto aos usuários corporativos quanto aos particulares. Por exemplo, estas aplicações podem incluir:

- Publicidade direcionada em páginas web – usuários podem ter diferentes preferências regionais. Ser capaz de adaptar regionalmente produtos, serviços, estratégias de propaganda e conteúdo, provê um diferencial de atratividade;
- Distribuição restrita de conteúdo – seguindo alguma regulamentação regional, um serviço de geolocalização fornece subsídios para a definição de quais usuários estão autorizados, ou não, a receber um determinado conteúdo;
- Verificação de segurança baseada em localização – se localizações autorizadas são conhecidas, uma transação de comércio eletrônico que for requisitada de algum outro lugar pode gerar avisos sobre um comportamento atípico ou não autorizado de um cliente.

Uma grande variedade de aplicações conscientes de localização pode ser vislumbrada com base em um serviço de mapeamento de endereços IP em localizações geográficas, podendo-se beneficiar assim tanto usuários finais quanto o gerenciamento de redes. Além disto, diferentes aplicações podem requerer diferentes níveis de acurácia na informação de localização. Nosso objetivo é portanto fornecer uma metodologia que seja capaz de geolocalizar nós na Internet de forma acurada e simultaneamente associar uma região de confiabilidade à estimativa de localização.

2.2. Trabalhos relacionados

Uma extensão do serviço DNS para o fornecimento também de um serviço de geolocalização é proposta na RFC 1876 [Davis et al., 1996]. No entanto, a utilização do serviço DNS para geolocalização foi limitada, pois isto requer mudanças nos registros DNS e os administradores tem pouca motivação para cadastrar novos registros de localização. Ferramentas, tais como NetGeo [Moore et al., 2000] consultam bases de dados Whois de modo a obter informação de localização para inferir a localização geográfica de um nó, mas estas bases são pouco atualizadas e podem ser potencialmente imprecisas.

Em [Padmanabhan e Subramanian, 2001] são investigadas três diferentes técnicas para inferir a localização geográfica de um nó na Internet. A primeira técnica infere a localização de um nó com base na extração de indicações geográficas do nome DNS do próprio nó ou de um nó próximo. A segunda técnica divide o espaço de endereçamento IP em agrupamentos de forma que todos os nós com endereços no mesmo agrupamento estejam possivelmente co-localizados.

Conhecendo a localização de alguns nós do agrupamento e supondo que estes são coerentes entre si, a técnica infere a localização de todo o agrupamento. A terceira técnica, chamada GeoPing, é a mais próxima da nossa, visto que ela se baseia em uma possível correlação entre o atraso na rede e a distância geográfica entre os nós. Dado um conjunto de nós de referência com localização geográfica bem conhecida, a estimativa de localização para o nó-alvo é a localização do nó de referência cujo padrão de atraso seja o mais similar ao observado para o nó-alvo.

Na técnica GeoPing, o número de possíveis estimativas de localização é limitado ao número de nós de referência adotado, caracterizando um espaço discreto de respostas. Logo, para melhorar a acurácia de técnicas como GeoPing, é necessário adicionar nós de referência ao sistema [Ziviani et al., 2005]. Na Seção 4.2, nós comparamos CBG com métodos do tipo GeoPing e do tipo DNS, e mostramos que CBG os supera.

2.3. Contribuições

Nesta subseção, nós destacamos as contribuições do CBG em relação aos trabalhos relacionados em geolocalização de nós na Internet:

- CBG estabelece uma relação dinâmica entre endereços IP e a sua localização geográfica. Esta relação dinâmica resulta de uma abordagem baseada em medições onde nós de referência cooperam de uma maneira distribuída e auto-calibrável, permitindo assim ao CBG adaptar-se às condições da rede variantes no tempo. Isto contrasta com os trabalhos relacionados baseados em uma relação estática;
- uma contribuição importante de CBG é demonstrar que medições de atraso podem ser transformadas de forma acurada em restrições de distância geográfica para serem usadas com multilateração. Isto potencialmente leva a estimativas mais acuradas da localização geográfica de nós na Internet;
- CBG oferece um espaço contínuo de respostas ao invés de um espaço discreto como fazem as outras abordagens baseadas em medições;
- CBG associa uma região de confiabilidade a cada estimativa de localização, permitindo às aplicações conscientes de localização avaliar se a estimativa de localização fornecida possui resolução suficiente em relação às necessidades de cada aplicação.

3. A proposta CBG

3.1. Multilateração com restrições de distância

A posição física de um determinado ponto pode ser estimada usando um número de medições de distâncias ou ângulos em relação a alguns pontos fixos, cujas posições sejam conhecidas. Quando lida-se com distâncias, este processo é chamado multilateração. De forma similar, quando lida-se com ângulos, o processo é chamado de multiangulação. Formalmente, triangulação refere-se ao processo de estimativa de uma posição baseado em ângulos usando três pontos de referência. Entretanto, freqüentemente este mesmo termo costuma ser adotado para qualquer estimativa de posição, seja ela baseada em ângulos ou em distância. Apesar da popularidade do termo triangulação, ao longo deste artigo adotamos o termo mais preciso e adequado à nossa técnica: multilateração.

O principal problema que surge da utilização de multilateração é a medição acurada das distâncias entre o nó-alvo a ser localizado e os nós de referência. Por exemplo, o sistema GPS usa multilateração com diversos satélites para estimar a posição de um determinado receptor GPS. No caso do GPS, a distância entre um receptor GPS e um satélite é medida pelo tempo que um sinal leva do satélite ao receptor. A medição precisa de tempo e de intervalos de tempo é o

ponto-chave na acurácia do sistema GPS. Em contraste com o GPS, é um problema desafiador transformar medições de atraso na rede em distâncias geográficas de forma acurada. Esta é a razão mais plausível em nossa opinião para justificar que a adoção direta de multilateração tenha permanecido inexplorada para a geolocalização de nós na Internet. A partir deste ponto, nós explicamos os fundamentos de CBG que o habilitam a utilizar multilateração com restrições de distância geográfica.

Para a localização de nós na Internet usando multilateração, nós lidamos com o problema de estimar a distância geográfica de um nó-alvo a ser localizado aos nós de referência, dadas as medições de atraso destes nós de referência ao nó-alvo. O princípio fundamental da metodologia CBG é que, não importa a razão, o atraso é somente distorcido de forma aditiva com relação ao tempo para a luz em fibras óticas passar sobre uma mesma distância. Portanto, nós nos interessamos em nos beneficiar desta invariante e desenvolver um método para estimar restrições de distância geográfica a partir de medições de atraso distorcidas aditivamente. A maneira pela qual CBG utiliza esta idéia para inferir restrições entre os nós de referência e o nó-alvo a partir de medições de atraso é detalhada na Seção 3.2. Também é mostrado que, como uma conseqüência da distorção de atraso aditiva, as restrições de distância geográfica resultantes são em geral super-estimadas com relação às distâncias geográficas reais.

A Figura 1 ilustra o processo de multilateração usado por CBG com um conjunto de nós de referência $\mathcal{L} = \{L_1, L_2, L_3\}$ na presença de alguma distorção aditiva de distância devido a medições imprecisas. Cada nó de referência L_i deve inferir a sua restrição geográfica de distância ao nó-alvo τ, cuja localização é desconhecida. No entanto, a restrição de distância geográfica estimada é na realidade determinada por $\hat{g}_{i\tau} = g_{i\tau} + \gamma_{i\tau}$, ou seja, a distância geográfica real $g_{i\tau}$ acrescida de uma distorção aditiva de distância geográfica representada por $\gamma_{i\tau}$. Esta distorção de distância puramente aditiva $\gamma_{i\tau}$ resulta da presença eventual de alguma distorção aditiva de atraso. Como uma conseqüência da existência da distorção aditiva de distância, a estimativa de localização do nó-alvo τ encontra-se provavelmente em alguma parte no interior da área acinzentada (ver Figura 1) que corresponde à interseção das restrições de distância geográfica super-estimadas dos nós de referência ao nó-alvo.

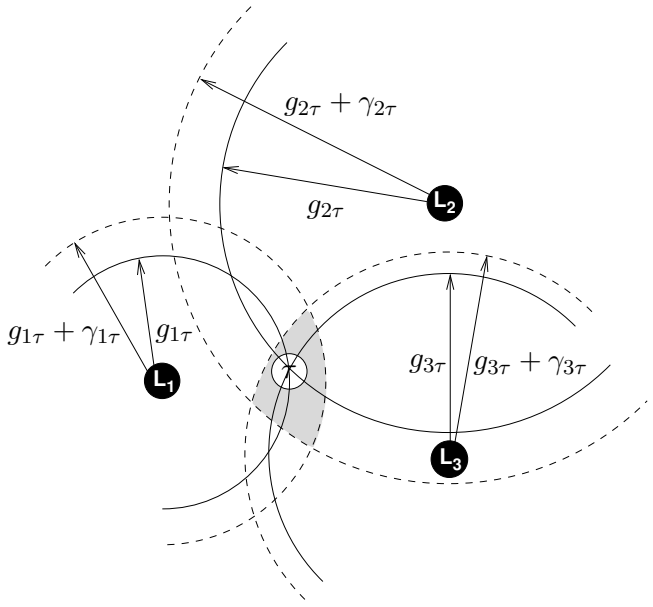


Figura 1: Multilateração com restrições de distância geográfica.

3.2. Transformando medições de atraso em restrições de distância

Antes de introduzirmos como CBG converte medições de atraso em restrições de distância, devemos primeiro observar um gráfico de espalhamento (*scatter plot*) relacionando distância geográfica e atraso na rede. Esta amostra, mostrada na Figura 2, faz parte dos dados experimentais descritos na Seção 4. O eixo dos x apresenta o atraso de rede entre um determinado nó de referência L_i e os demais nós de referência. O significado de "reta_de_base" e "melhor_reta" na Figura 2 são explicados ao longo desta seção. Trabalhos recentes [van Langen et al., 2004, Ziviani et al., 2004] investigam o coeficiente de correlação observado neste tipo de gráfico de espalhamento, obtendo um ajuste de mínimos quadrados para caracterizar a relação entre distância geográfica e atraso na rede. Em contrapartida, nós consideramos as *razões* pelas quais os pontos encontram-se espalhados em gráficos deste tipo e argumentamos que o mais importante não é o ajuste de mínimos quadrados, mas o maior limite inferior.

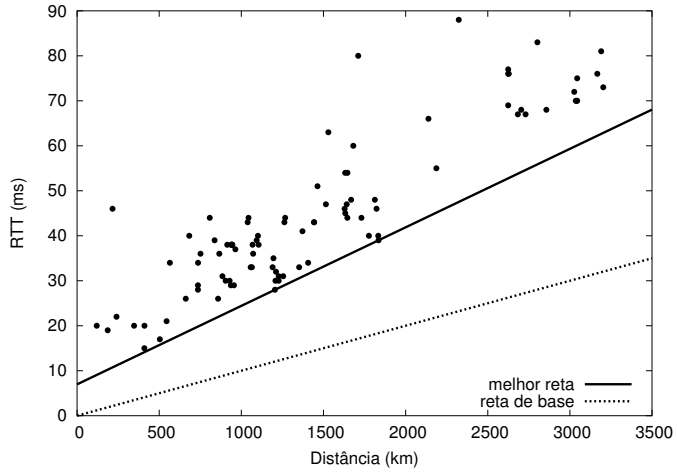


Figura 2: Exemplo de gráfico de espalhamento entre distância e atraso.

Baseados nestas considerações, nós propomos uma abordagem inédita para estabelecer uma relação dinâmica entre o atraso de rede e a distância geográfica. De maneira a ilustrar esta abordagem, suponha a existência de caminhos diretos entre um nó de referência L_i e cada um dos demais nós de referência. Mais do que isto, considere também que, quando sendo transmitidos nestes caminhos diretos, os dados estão apenas sujeitos ao atraso de propagação do meio de comunicação. Neste caso ideal, nós deveríamos ter uma reta representando esta relação que é dada por $y = mx + b$, onde $b = 0$ pois não há atrasos locais e m está somente relacionado com a velocidade na qual os bits trafegam no meio de comunicação. Como já mencionado, informação digital trafega ao longo de cabos de fibra ótica a quase exatamente 2/3 da velocidade da luz no vácuo [Percacci e Vespignani, 2003]. Isto fornece uma regra bastante conveniente de 1 ms de RTT para cada 100 km de cabo. Esta relação pode ser usada para obter um limite inferior físico absoluto no RTT (ou no atraso unidirecional) entre sítios cuja localização geográfica seja conhecida. Este limite inferior é representado pela "reta_de_base" na Figura 2. Neste caso idealizado, nós poderíamos usar esta regra conveniente para extrair de uma maneira simples e direta distâncias geográficas acuradas entre sítios na rede a partir dos atrasos medidos. No entanto, na prática, estes caminhos diretos raramente existem [Subramanian et al., 2002]. Portanto, nós devemos lidar com caminhos que desviam do modelo idealizado por diversas razões, incluindo atrasos por congestionamento e pela falta de caminhos diretos entre os nós.

Como colocado na Seção 3.1, a principal idéia da proposta CBG consiste no fato de que a combinação de diferentes fontes de distorção de atraso em relação ao caso idealizado de caminho

direto produz um fator de incremento puramente geométrico do atraso. Nós então modelamos a relação entre o atraso de rede e a distância geográfica usando medições de atraso da seguinte maneira. Nós definimos a "melhor_reta" para um dado nó de referência L_i como a reta $y = m_i x + b_i$ que é a mais próxima, mas encontra-se abaixo, de todos os pontos (x, y) e que possui uma interseção positiva com o eixo y, pois não faz sentido considerar atrasos negativos. Deve-se ressaltar que cada nó de referência calcula a sua própria "melhor_reta" em relação aos demais nós de referência. Portanto, a "melhor_reta" pode ser vista como a reta que captura a relação menos distorcida entre a distância geográfica e o atraso de rede do ponto de vista de cada nó de referência.

O cálculo da "melhor_reta" é formulado como um problema de programação linear. Para um determinado nó de referência L_i, há o atraso d_{ij} e a distância g_{ij} até cada nó de referência L_j, onde $i \neq j$. Nós precisamos encontrar, para cada nó de referência L_i, o coeficiente angular m_i e o coeficiente linear b_i que determinam a "melhor_reta" dada a equação $y = m_i x + b_i$. A condição de que a "melhor_reta" para cada nó de referência L_i deva estar abaixo de todos os pontos (x, y) define a região onde uma solução deve encontrar-se:

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

onde o coeficiente angular é $m_i = (d_{ij} - b_i)/g_{ij}$. A função objetivo para minimizar a distância entre a reta com coeficiente linear não-negativo e todas as medições de atraso é dada por

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

onde m é o coeficiente angular da "reta_de_base". A Equação (2) é usada para encontrar a solução m_i e b_i da Equação (1) que determina a "melhor_reta" para cada nó de referência L_i. Cada nó de referência L_i então utiliza a sua própria "melhor_reta" para converter a medida de atraso até o nó-alvo em uma distância geográfica. Logo, a restrição de distância geográfica estimada $\hat{g}_{i\tau}$ entre um nó de referência L_i e o nó-alvo τ é obtida do atraso $d_{i\tau}$ usando a "melhor_reta" do nó de referência L_i como segue

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

Se os atrasos entre os nós de referência forem periodicamente colhidos, isto leva a um algoritmo auto-calibrável que determina como cada nó de referência observa em um determinado momento a relação dinâmica entre o atraso de rede e a distância geográfica dentro da rede.

3.3. Usando restrições de distância distribuídas para geolocalizar nós na Internet

CBG utiliza uma abordagem geométrica usando multilateração para estimar a localização de um dado nó-alvo τ. Cada nó de referência L_i infere a sua restrição de distância geográfica ao nó-alvo τ, que na realidade é a distância aditivamente distorcida $\hat{g}_{i\tau} = g_{i\tau} + \gamma_{i\tau}$ usando a Equação (3). Portanto, cada nó de referência L_i estima que o nó-alvo τ esteja em algum lugar no interior de um círculo $\mathcal{C}_{i\tau}$, centrado no nó de referência L_i e com um raio igual à restrição de distância geográfica estimada $\hat{g}_{i\tau}$. Dados K nós de referência, o nó-alvo τ possui uma coleção de curvas fechadas $\mathbf{C}_\tau = \{\mathcal{C}_{1\tau}, \mathcal{C}_{2\tau}, \ldots, \mathcal{C}_{K\tau}\}$ que pode ser vista como um diagrama de Venn de ordem-K. Das 2^K possíveis regiões definidas por este diagrama de Venn de ordem-K para o nó-alvo τ, nós estamos interessados na região de interseção $\mathcal{R}$ de todas as curvas fechadas $\mathcal{C}_{i\tau} \in \mathbf{C}_\tau$, dada por

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

Note que $\mathcal{R}$ é uma região convexa, visto que as regiões $\mathcal{C}_{i\tau}$ são convexas, e a interseção de conjuntos convexos é também convexo.

4. Resultados experimentais

4.1. Dados utilizados

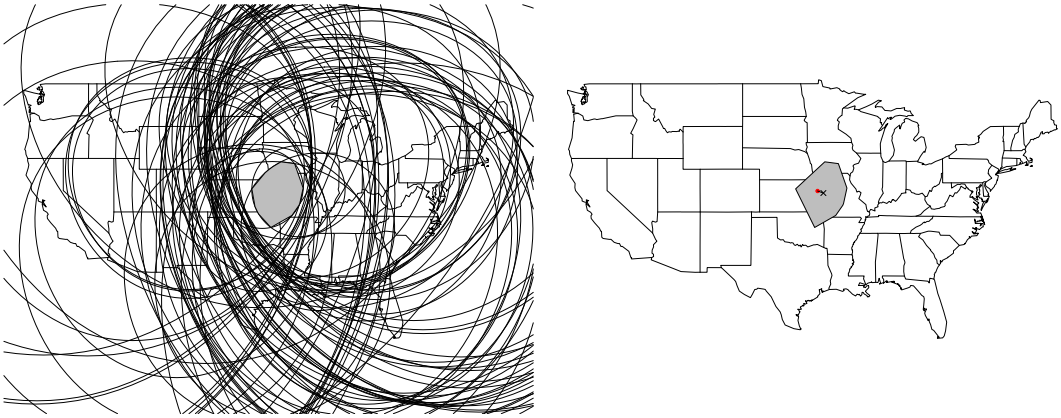
- RIPE – o conjunto de dados coletados no projeto *Test Traffic Measurements* (TTM) [RIPE, 2000] que nós consideramos é composto pelo percentil 2,5 do atraso unidirecional observado de cada nó da rede RIPE a cada um dos outros nós desta rede durante o período de 10 semanas de dezembro de 2002 até fevereiro de 2003. Cada nó da rede RIPE gera aproximadamente 300 kB de tráfego por dia até cada um dos outros nós da rede RIPE com uma média de dois pacotes enviados por minuto. A maioria dos nós da rede RIPE encontra-se localizada na Europa e todos são equipados com placas GPS, assim permitindo que a sua localização geográfica exata seja conhecida. Nós então usamos os 42 nós da rede RIPE na Europa Ocidental para compor o conjunto de dados da Europa.
- NLANR AMP – o conjunto de dados coletados no *Active Measurement Project* (AMP) [AMP, 1998] que nós consideramos é composto pelo percentil 2,5 do RTT entre todos os nós participantes localizados na porção continental dos EUA, totalizando 95 nós. Estes dados foram coletados em 30 de janeiro de 2003 e são simétricos. O atraso é amostrado, em média, uma vez por minuto. Isto leva a uma carga de medições média de aproximadamente 144 kB por dia enviados por cada nó da rede AMP até cada um dos outros nós da rede AMP. A localização exata de cada um dos nós participantes também está disponível. Estes 95 nós da rede AMP compõem nosso conjunto de dados dos EUA.

Em nossos experimentos, os nós em cada conjunto de dados fazem um por vez o papel de nó-alvo a ser localizado. Os nós restantes no mesmo conjunto de dados são então considerados como nós de referência para realizar a estimativa de localização do nó-alvo. A "melhor_reta" de cada nó de referência é calculada usando o conjunto de nós de referência de cada cenário, portanto excluindo o nó-alvo. Nós repetimos este procedimento para avaliar as estimativas de localização resultantes para cada nó em ambos os conjuntos de dados, Europa e EUA.

4.2. Geolocalizando nós na Internet usando CBG

A partir das restrições de distância geográfica, CBG determina para cada nó-alvo τ um conjunto de curvas fechadas $\mathbf{C}_\tau = \{\mathcal{C}_{1\tau}, \mathcal{C}_{2\tau}, \ldots, \mathcal{C}_{K\tau}\}$ (veja Seção 3.3), onde $K = 42$ para o conjunto de dados da Europa e $K = 95$ para o conjunto dos EUA. Cada curva em $\mathbf{C}_\tau$ é centrada no seu respectivo nó de referência L_i e tem um raio equivalente à restrição de distância estimada $\hat{g}_{i\tau}$.

Para ilustrar a metodologia proposta por CBG, a Figura 3(a) mostra um exemplo de um conjunto de curvas fechadas extraído de nosso estudo experimental. A área da região de interseção $\mathcal{R}$, ou seja, a área acinzentada na Figura 3(a), indica a região de confiabilidade que CBG associa esta estimativa de localização. Deve-se ressaltar que a maioria das regiões de confiabilidade observadas possuem uma área relativamente pequena, não visíveis em ilustrações similares com todas as curvas fechadas presentes (a Seção 4.4 apresenta os resultados sobre o tamanho das regiões de confiabilidade). Este exemplo possui uma região de confiabilidade maior do que



(a) Determinação da região de confiabilidade (b) Estimativa de localização do nó-alvo

Figura 3: Exemplo do procedimento de geolocalização usando CBG.

o tamanho típico, mas foi selecionado exatamente por possuir uma região suficientemente visível para ilustrar a metodologia de CBG.

A determinação da estimativa pontual de localização, apresentada na Figura 3(b), é discutida a seguir. A região $\mathcal{R}$ é a estimativa de localização fornecida por CBG. Uma aproximação intuitiva da localização pontual do nó-alvo é o centróide desta região. Portanto, CBG usa o centróide da região $\mathcal{R}$ como estimativa pontual da posição do nó-alvo.

Nós adotamos a seguinte heurística para aproximar a região de interseção $\mathcal{R}$, ou seja, a estimativa de localização associada pelo CBG com o nó-alvo τ, por um polígono. O polígono resultante é usado para aproximar a área da região $\mathcal{R}$ e fornecer uma estimativa da localização pontual do nó-alvo. Para formar o polígono, nós consideramos como vértices os pontos de cruzamento dos círculos $\mathcal{C}_{i\tau}$ que pertencem a todos os círculos. Como a região $\mathcal{R}$ é convexa, o polígono é uma subestimava da área de $\mathcal{R}$. Nós então aproximamos a região $\mathcal{R}$ pelo polígono formado dos segmentos de reta entre os N vértices $v_n = (x_n, y_n)$, $0 \leq n \leq N-1$. O último vértice $v_N = (x_N, y_N)$ é suposto como sendo o mesmo que o primeiro, ou seja, o polígono é fechado. Estes vértices do polígono associado ao nó-alvo τ são os pontos de interseção que pertencem a todos os círculos $\mathcal{C}_{i\tau}$. A área de um polígono, que não se auto-intercepta, com vértices $v_0 = (x_0, y_0), \ldots, v_{N-1} = (x_{N-1}, y_{N-1})$ é dada por

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

onde $|\mathbf{M}|$ denota a determinante da matriz $\mathbf{M}$. O centróide c do polígono, ou seja, a estimativa pontual de localização do nó-alvo τ, está posicionado em (c_x, c_y), coordenadas estas obtidas por

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

e

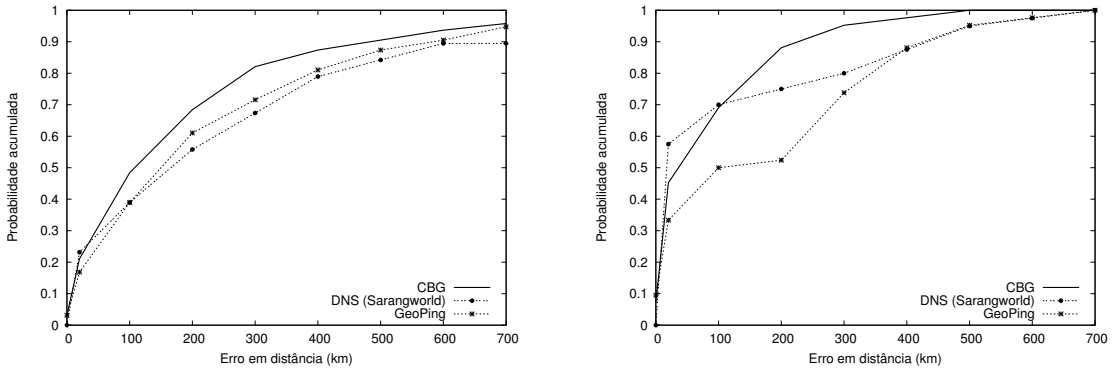
$$c_y = \frac{1}{6A} \sum_{n=0}^{N-1} (y_n + y_{n+1}) \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix}. \tag{7}$$

A estimativa pontual de localização do nó-alvo e a estimativa da região de confiabilidade são, respectivamente, o centróide (c_x, c_y) e a área A do polígono aproximado. A Figura 3(b) indica uma amostra de um polígono obtido por esta heurística. A área acinzentada apresentada na Figura 3(b) é a aproximação por um polígono da região de interseção mostrada na Figura 3(a). No interior deste polígono, o círculo indica a localização real do nó-alvo, enquanto a cruz indica a estimativa pontual de localização fornecida pelo CBG que corresponde ao centróide do polígono.

4.3. Erro em distância da estimativa de localização

Depois de inferir a estimativa pontual para cada nó-alvo considerado, nós calculamos o erro em distância, que corresponde à diferença entre a posição estimada e a localização real do nó-alvo τ. Nós comparamos o desempenho de CBG com os resultados obtidos por um sistema de geolocalização baseado em medições com um espaço discreto de respostas do tipo GeoPing [Padmanabhan e Subramanian, 2001, Ziviani et al., 2004], ou seja, onde a localização dos nós de referência é usada como estimativa de localização. CBG também é comparado a um método baseado em DNS (*SarangWorld Traceroute project* [Sarangworld, 2003]) que realiza um `traceroute` em direção ao nó-alvo e infere a geolocalização dos nós intermediários com base nos seus nomes DNS. A geolocalização inferida do nó mais próximo ao nó-alvo é usada como a estimativa de localização.

A Figura 4 mostra a distribuição de probabilidade acumulada do erro em distância observado usando os métodos CBG, DNS e GeoPing. CBG supera em desempenho tanto o método DNS quanto o método GeoPing com um espaço discreto de respostas. A diferença de desempenho entre as duas abordagens é mais acentuada no conjunto de dados referente à Europa. Isto é justificado pela presença menos numerosa de nós de referência no conjunto da Europa do que no conjunto dos EUA. Na abordagem com espaço discreto de respostas, como as respostas estão limitadas às localizações dos nós de referência, a quantidade e a localização dos nós de referência constituem pontos-chave no seu desempenho [Ziviani et al., 2005]. Na Seção 4.5, nós investigamos o impacto do número de nós de referência adotados no desempenho de CBG.



(a) EUA (b) Europa

Figura 4: Erro em distância para os métodos CBG, DNS e GeoPing.

Nos resultados de CBG, o erro em distância médio no conjunto dos EUA é 182 km, enquanto o resultado equivalente para o conjunto da Europa é de 78 km. A maioria dos nós em ambos os conjuntos de nós de referência teve uma boa estimativa de localização. A mediana do erro em distância e o percentil 80 para o conjunto dos EUA é de 95 km e 277 km, respectivamente. No conjunto da Europa, a mediana do erro em distância é de 22 km e o percentil 80 é de 134 km.

4.4. Região de confiabilidade de uma estimativa de localização

A área total da região de interseção $\mathcal{R}$ encontra-se de certa forma relacionada com a confiabilidade que CBG associa às estimativas de localização resultantes de sua metodologia. Intuitivamente, esta área quantifica a extensão geográfica de cada estimativa de localização em km^2. Quanto menor for a área da região $\mathcal{R}$, mais confiante CBG está na sua estimativa de localização. Portanto, em contraste com as técnicas anteriores de geolocalização baseadas em medições, CBG associa uma região de confiabilidade em km^2 para cada estimativa de localização fornecida. Nós acreditamos que isto seja de grande importância porque esta região de confiabilidade pode ser utilizada por aplicações conscientes de localização para avaliar o quanto elas podem confiar na estimativa de localização fornecida. Além disto, nós vislumbramos aplicações com diferentes requisitos em termos de acurácia na estimativa de localização. Usando a região de confiabilidade, estas aplicações podem decidir se a estimativa de localização fornecida possui resolução suficiente em relação às suas necessidades.

A Figura 5 apresenta a probabilidade acumulada da extensão das regiões de confiabilidade em km^2 para as estimativas de localização de ambos os conjuntos de nós de referência, EUA e Europa. Os resultados mostram que, para o conjunto dos EUA, CBG associa uma região de confiabilidade com área total menor que 10^5 km^2 para aproximadamente 80% das estimativas de localização. Esta área é um pouco maior do que a superfície do estado de Santa Catarina. Para o conjunto da Europa, 80% das estimativas de localização possuem uma região de confiabilidade de até 10^4 km^2 (menos da metade da superfície do menor estado brasileiro, Sergipe), permitindo assim uma localização ao nível regional. Uma região de confiabilidade de menos do que 10^3 km^2, o que equivale à extensão da região metropolitana de uma grande cidade, é alcançada por 25% dos nós-alvos nos EUA e por 65% dos nós-alvos na Europa.

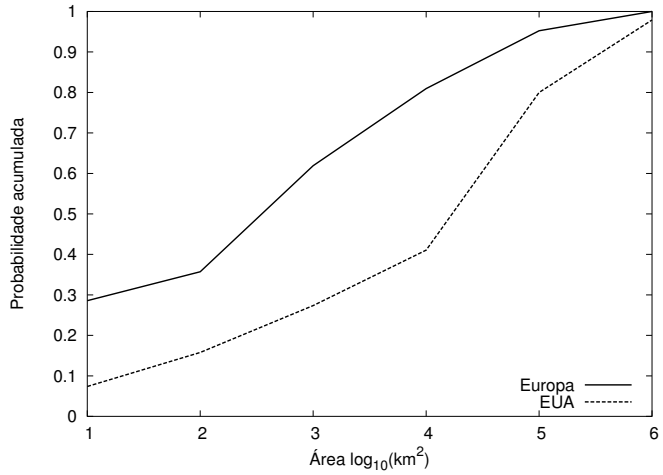
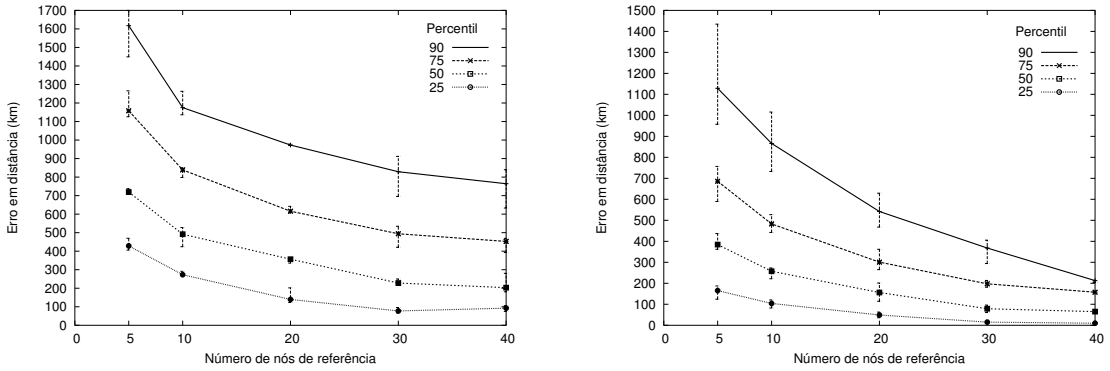


Figura 5: Regiões de confiabilidade fornecidas por CBG.

4.5. Impacto do número de nós de referência

Nesta subseção, nós avaliamos o impacto do número de nós de referência adotados no desempenho de CBG. Para cada conjunto de dados, nós calculamos o erro médio em distância como a média de todos os erros em distância correspondentes a vários conjuntos aleatórios de k nós de referência escolhidos em meio ao número total de nós de referência disponíveis (42 para o conjunto da Europa e 95 para o conjunto dos EUA). Como o número de possíveis combinações torna-se bastante grande conforme nós aumentamos o valor de k, nós não levamos em conta todas as possíveis escolhas de k nós de referência de cada conjunto de dados.

A Figura 6 mostra diferentes níveis de percentil para o erro em distância das estimativas de localização fornecidas pelo CBG como uma função do número de nós de referência adotados. Por exemplo, a curva correspondente ao percentil 90 representa o erro em distância no qual o gráfico de probabilidade acumulada do erro médio em distância alcança a probabilidade 0,90. Estes resultados sugerem que um certo número de nós de referência, tipicamente em torno de 30, é suficiente para estabilizar o erro médio em distância para ambos os conjuntos de dados.



(a) EUA (b) Europa

Figura 6: Erro em distância em função do número de nós de referência adotados.

5. Discussão

Nesta seção, nós discutimos tópicos relacionados a tecnologias de geolocalização de nós na Internet de forma geral. Nós enfatizamos, contudo, que as questões aqui levantadas não necessariamente afetam CBG mais do que afetariam qualquer outra técnica de geolocalização.

O desenvolvimento e o uso de tecnologias de geolocalização geram preocupações relativas a temas como privacidade e segurança. Um grupo de trabalho do IETF, chamado *Geographic Location/Privacy* (geopriv) [Geopriv, 2003], está atualmente dedicado a estabelecer políticas para controlar a troca de informações de localização tendo a privacidade como meta. O desenvolvimento de tecnologias de geolocalização é definido como fora do escopo deste grupo de trabalho. Nosso trabalho de pesquisa é complementar ao esforço do grupo de trabalho *geopriv*, pois nós investigamos a inferência da geolocalização de nós na Internet. Nós acreditamos que qualquer tecnologia de geolocalização, incluindo CBG, deva considerar as questões de privacidade e segurança no uso da informação de localização fornecida. Além disto, a abordagem proposta na comunidade do grupo *geopriv* consiste em fornecer informação de localização com resolução reduzida a usuários não-privilegiados. A região de confiabilidade atribuída pelo CBG a cada estimativa de localização pode ser diretamente utilizada com este propósito.

Procuradores (*proxies*) impõem um limite fundamental em técnicas de geolocalização baseadas em medições. Como o endereço IP visto pela rede externa pode corresponder na realidade ao endereço do procurador, as técnicas de geolocalização baseadas em medições inferem a localização do procurador, o que pode ser inacurado no caso do cliente e do procurador não estarem relativamente próximos. Um cliente e um procurador podem estar em relativa proximidade como no caso de um procurador em um campus de uma universidade ou em um ISP (*Internet Service Provider*) local. Neste caso, a estimativa de localização não tende a ser excessivamente inacurada. Em alguns casos, entretanto, o cliente e o procurador podem estar distantes, como em alguns

grandes ISPs que concentram um agrupamento de procuradores para seus clientes em uma única localização. Como uma contra-medida prática a isto, alguns serviços de geolocalização mantêm uma base de dados de procuradores de grandes ISPs para abster-se de estimar a localização nestes casos. A recusa em fornecer uma estimativa de localização pode ser um primeiro passo, mas não exatamente uma solução ao problema. Esta área é objeto de investigação futura.

Tecnologias de geolocalização baseadas em medições supõem que o nó-alvo é capaz de responder às medições, por exemplo um pedido de `ping`. Nós também supomos neste artigo que o nó-alvo responde às medições exatamente como os nós de referência o fazem na proposição de CBG. Isto foi suposto para favorecer a simplicidade na apresentação de CBG. No entanto, mesmo se o nó-alvo não responder diretamente a pedidos `ping`, uma geolocalização baseada em medições pode ainda ser viável. Uma possível contra-medida que nós consideramos é a utilização de `traceroute` para então buscar alvos secundários a serem medidos que estejam em relativa proximidade em número de saltos do nó-alvo original. Ao limitar a distância em número de saltos e inferir a localização destes alvos secundários, uma estimativa de localização pode ser viável com uma menor acurácia.

6. Conclusão

Neste artigo, nós propusemos CBG (*Constraint-Based Geolocation*), um método baseado em medições para estimar a localização geográfica de nós na Internet. Baseado em medições de atraso, CBG utiliza multilateração para inferir a geolocalização de um determinado nó-alvo. A transformação acurada de medições de atraso em distância geográfica é desafiadora devido a muitas características de uso e de implementação da Internet atual. Entre estas características estão o atraso em filas e a ausência de caminhos diretos entre os nós. CBG apresenta uma contribuição ao apontar que uma transformação acurada de medições de atraso em *restrições* de distâncias geográficas é entretanto viável. Além disto, CBG demonstra que na prática estas restrições são suficientemente justas para permitir uma estimativa de localização acurada usando multilateração. CBG também estabelece uma relação dinâmica entre o atraso de rede e a distância geográfica. Isto se realiza de uma maneira distribuída e auto-calibrável entre os nós de referência adotados usando o método da "melhor_reta".

Nossos resultados experimentais mostraram que CBG supera em desempenho técnicas anteriores de geolocalização. A mediana do erro em distância obtida em nossos experimentos para o conjunto dos EUA está abaixo de 100 km, enquanto que para o conjunto da Europa este valor está abaixo de 25 km. Estes resultados contrastam com medianas do erro em distância de aproximadamente 150 km para o conjunto dos EUA e de 100 km para o conjunto da Europa quando métodos similares ao GeoPing são utilizados. Além disto, em contraste com as abordagens anteriores, CBG atribui uma região de confiabilidade a cada estimativa de localização. Isto é importante, pois permite a aplicações conscientes de localização avaliar se a estimativa de localização é suficientemente acurada para as suas necessidades. Nossos resultados indicam que uma estimativa de localização acurada, ou seja, com uma região de confiabilidade relativamente pequena, é alcançada na maioria dos casos em ambos os conjuntos de dados utilizados, assim fornecendo uma informação de localização em nível regional. Por nível regional, nós entendemos uma região de tamanho equivalente a de um pequeno estado brasileiro.

Nossos resultados são baseados em medições tomadas em redes geograficamente contíguas e com alto grau de conectividade. De certa maneira, nosso trabalho se beneficia do fato que a conectividade na rede aumentou significativamente na última década e que a relação entre o atraso e a distância é mais forte nestas regiões [Yook et al., 2002, Ziviani et al., 2004]. A localização de sistemas finais típicos faz parte do nosso trabalho futuro.

Referências

AMP (1998). *NLANR Active Measurement Project*. http://watt.nlanr.net/.

Bovy, C. J., Mertodimedjo, H. T., Hooghiemstra, G., Uijterwaal, H., e van Mieghem, P. (2002). Analysis of end-to-end delay measurements in Internet. In *Proc. of the Passive and Active Measurement Workshop - PAM'2002*, Fort Collins, CO, EUA.

Davis, C., Vixie, P., Goodwin, T., e Dickinson, I. (1996). A means for expressing location information in the domain name system. *Internet RFC 1876*.

Enge, P. e Misra, P. (1999). Special issue on global positioning system. *Proceedings of the IEEE*, 87(1):3–15.

Geopriv (2003). Geographic location/privacy (geopriv) IETF working group. http://www.ietf.org/html.charters/geopriv-charter.html.

Gueye, B., Ziviani, A., Crovella, M., e Fdida, S. (2004). Constraint-based geolocation of Internet hosts. In *Proc. of ACM/SIGCOMM Internet Measurement Conference – IMC 2004*, Taormina, Itália.

Lakhina, A., Byers, J. W., Crovella, M., e Matta, I. (2003). On the geographic location of Internet resources. *IEEE Journal on Selected Areas in Communications*, 21(6):934–948.

Moore, D., Periakaruppan, R., Donohoe, J., e Claffy, K. (2000). Where in the world is netgeo.caida.org? In *Proc. of the INET'2000*, Yokohama, Japão.

Padmanabhan, V. N. e Subramanian, L. (2001). An investigation of geographic mapping techniques for Internet hosts. In *Proc. of the ACM SIGCOMM'2001*, San Diego, CA, EUA.

Percacci, R. e Vespignani, A. (2003). Scale-free behavior of the Internet global performance. *The European Physical Journal B - Condensed Matter*, 32(4):411–414.

RIPE (2000). *RIPE Test Traffic Measurements*. http://www.ripe.net/ttm/.

Sarangworld (2003). *Sarangworld Traceroute Project*. http://www.sarangworld.com/TRACEROUTE/.

Subramanian, L., Padmanabhan, V. N., e Katz, R. (2002). Geographic properties of Internet routing. In *Proc. of USENIX 2002*, Monterey, CA, EUA.

van Langen, S., Zhou, X., e van Mieghem, P. (2004). On the estimation of Internet distances using landmarks. In *Proc. of the International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking – NEW2AN'04*, São Petersburgo, Rússia.

Yook, S.-H., Jeong, H., e Barabási, A.-L. (2002). Modeling the Internet's large-scale topology. *Proc. of the National Academy of Sciences (PNAS)*, 99:13382–13386.

Ziviani, A., Fdida, S., de Rezende, J. F., e Duarte, O. C. M. B. (2004). Toward a measurement-based geographic location service. In *Proc. of the Passive and Active Measurement Workshop - PAM'2004*, Lecture Notes in Computer Science (LNCS) 3015, pages 43–52, Antibes Juan-les-Pins, França.

Ziviani, A., Fdida, S., de Rezende, J. F., e Duarte, O. C. M. B. (2005). Improving the accuracy of measurement-based geographic location of Internet hosts. *Computer Networks, Elsevier Science*, 47(4):503–523.

Zook, M. (2001). Connected is a matter of geography. *ACM NetWorker*, 5(3):13–17.