

Web Neighborhood Watch

Gene Connolly, Anatoly Sachenko, George Markowsky

In the nine years that separated the terrorist attacks on the World Trade Center, the United States pioneered a revolution in information technology by popularizing and expanding the Internet and the World Wide Web. Now, with homeland security of foremost importance following the most recent attack, the domain of the Internet must be considered when defining the boundaries of the 'homeland' being protected. One of the characteristics of the Internet is that it offers geographic anonymity to its users. This has led to an increase in terrorist content and communication on the web, with few means of determining if it is located within the homeland, and where. The first step in securing the Internet homeland is devising a method of mapping Internet content to geographical locations.

The architecture of the Internet does not lend well to obtaining specific geographical location of any of the computers that comprise it. Should someone wish to use the Internet from a remote, geographically undisclosed location, they can achieve this goal. Most web content and IP addresses however, can be traced to the geographical region in which they reside. Although this information is of little use in the pursuit of specific offenders, it may be used effectively to monitor trends of potentially dangerous content in geographical regions. Fully realized, this system would act as a 'neighborhood watch' for the Internet.

The problem of geographically mapping IP addresses has garnered a fair amount of research, as there are many applications that would benefit from such information. The Cooperative Association for Internet Data Analysis (CAIDA) has several projects involved with geographic mapping of IP addresses [5][6], such as NetGeo, a database and collection of Perl scripts used for network visualization and analysis. Sarang Gupta [7] and Ankit Fadia [2][3] each separately offer extensive documentation on the Internet regarding using the traceroute utility to determine the geographical path data travels en route to its target computer.

Traceroute

These projects, as well as many others available, use the traceroute utility as their means of obtaining geographical information. Traceroute is a method that determines all of the routers a data packet travels through over a network to a target computer. The resulting path may be examined for network analysis or optimization. One of the pieces of information the traceroute populates is the hostname of each of the IP devices the traceroute encounters. Although traceroute was never intended to be used for geographical purposes, often these hostnames contain clues regarding the geographical position of the IP device. 'BOS' or 'NYC' may be found within the hostname of a router, included as part of a naming scheme used by the often expansive, or worldwide network that it is on. Similar to airline codes, 'BOS' and 'NYC' may easily be interpreted to represent Boston and New York City, respectively. Using this method, determining the geographical route traveled by an Internet Data packet is simply the matter of matching each hostname with the appropriate network naming scheme, extracting the location code, and translating it to the appropriate location.

This traceroute approach, however, is not all together an effective method. Many smaller networks do not include geographical codes within the naming scheme of their IP devices. Often, as a data packet approaches its target computer, descending from a backbone server to a local ISP, geographical codes become more scarce on networks that are not geographically expansive, and the target computer remains unable to locate. The best geographical information that may be obtained regarding the target computer in this scenario is a city that is more local to the target than any other city representing the backbone network the data traveled on.

Projects such as CAIDA's NetGeo, and the investigation pursued by Sarang Gupta use this traceroute approach as a means to determine the geographical route the data took to arrive at the destination. Traceroute accomplishes this task well, as its geographical information is 'route-based' and not 'destination-based'. Geographical location, for the purposes of developing a 'Web Neighborhood Watch', is interested in obtaining 'destination-based' information.

Considering this objective, the effectiveness of the traceroute approach becomes a question of the relevance that the final city located by the traceroute has upon the actual location of the target computer. If the knowledge base of network naming schemes is robust, then the final city located should always have geographical bearing upon the target computer. Networks are optimized to have data packets travel from the fastest, high-capacity nodes, so backbone servers will transport data as geographically close as it can to the target computer. With every hop the data makes following the final located city, however, is a degree of geographical anonymity that the target computer has achieved, effectively lowering the relevance that the location of the final city has upon it. Furthermore, there is no information suggesting the scope of the relevance that the final located city has upon the target.

The traceroute utility may be the most effective approach to geographically mapping IP addresses, but executing a single trace upon a target computer is not going to provide the geographical data needed classify the target to a region or neighborhood of the country. In order for it to produce effective results it would have to (1) provide more 'destination-based' information as to confirm the relevance of the location of the final located city, and (2) offer a geographical basis to determine a scope, or a set of boundaries that the target computer exists within.

Distributed Traceroute Approach

Interestingly, both of these conditions can be satisfied if traceroute analysis is examined from multiple tracroutes executed from geographically and network-diverse locations. Each trace that travels upon another network to achieve the same final located city as the other traces support the relevance to the target computer. Moreover, diverse network paths to the computer may offer more specific, or supporting data because they have major network hubs located in different cities. Each additional final located city discovered further refine the geographical region in which the target computer is located.

By approaching from different networks, the additional data is collected to achieve the 'destination-based' results. A geographical basis is achieved because multiple geographically

distributed traceroutes will approach the target from different directions, encircling the target computer in a geographical region.

This distributed approach, building upon the single traceroute approach, is an effective method of obtaining more geographical information regarding the destination computer. It does, however, hinge upon the ability to execute traces from multiple, geographically diverse locations. Thomas Kernan, author of traceroute.org, has compiled a list of traceroute servers available publicly [4]. Many educational institutions and networking corporations offer a web-based interface to access the traceroute utility from their servers. Utilizing these available resources provide the basis for the distributed approach.

Heretofore, traceroute has been the only solution to the problem of locating IP addresses proposed. Although it may be the only approach that is effective as a standalone solution, there are several resources that may be used to enhance its accuracy. The WHOIS database, for example, is a resource for obtaining information regarding every network. Each entry in the database contains, among other things, an address for administrative and technical contact. The address may be evaluated for its geographic relevance to the computers on that network. Unfortunately, as networks become more expansive, the administrative address becomes less relevant. Computers that make up worldwide networks, could be located anywhere worldwide.

WHOIS Database

Using WHOIS as an accessory to the distributed traceroute approach, however, complements the shortcomings of traceroute very well. Small networks whose administrative address in their WHOIS entry reflect upon the region of the target computer often are the networks that do not contain geographical codes in the naming scheme of their computers. Conversely, Worldwide networks that contain geographical codes in their computer-naming scheme do not contain relevant data in their WHOIS database entry.

By examining the WHOIS information of each hop that exists between the final located city of each trace and the target computer, much information can be obtained regarding the relationship between the two locations. Local network travel suggests a stronger bond between the final located city and the target computer, whereas worldwide network travel suggests a weaker bond.

Needless to say, incorporating WHOIS data with the distributed traceroute approach will only provide the more information regarding the networks surrounding the target computer. Developing a better understanding of the expansiveness of the networks that approach the target computer will enhance the certainty of the geographical region it is associated with.

Neither the traceroute utility nor the WHOIS database were ever intended to be used as tools for the geographical location of IP addresses. To that end, they cannot be relied upon to always produce correct results. By using the distributed traceroute approach, however, these tools are given the geographical basis to solve this geographical problem, and produce results, that although non-specific, conform to the limitations of the problem.

Future Work

Locating IP addresses, satisfied by the distributed traceroute approach, is only the first step in developing a 'Web Neighborhood Watch'. Using this approach to determine geographical trends of terrorist content will be the true test of the effectiveness of the system. There exist many doubts as to whether there is significance to the geographical position of web content. By and large, most web hosting companies offer large amounts of web space to a many people, all from a single remote location. If people can maintain content on their servers from anywhere in the world, does the content being stored on those servers reflect upon the geographical location in which they are located?

Many of these large web hosting companies offer terms of service to their clients declaring that (1) the web space is not to be used for illegal or potentially dangerous web content, and (2) the hosting company will not be held responsible for the content of its servers. Depending on how these terms of service are enforced, an analysis of the geographical trends will reveal two possible scenarios. Either (1) large amounts of potentially dangerous content is being stored on the remote servers of web hosting companies, with no geographical bearing, or (2) the majority of potentially dangerous web content is located on many different small web servers, whose location reflects upon the terrorist activity in the surrounding area.

-
- [1] Connolly, Gene M., George Markowsky, and Anatoly Sachenko. Distributed Traceroute Approach to Geographically Locating IP Devices. 21 Apr. 2003 <http://homeland.cs.umaine.edu/web_watch.htm>.
 - [2] Fadia, Ankit . Getting geographical Information using an IP Address. Astalavista.net. 30 Jan. 2003 <<http://www.astalavista.com/library/basics/organisation/ip-geography.shtml>>.
 - [3] Fadia, Ankit. Tracing The Traceroute. 24 Jan. 2002. 30 Jan. 2003 <<http://www.ankitfadia.com/traceroutew.html>>.
 - [4] Kernen, Thomas. traceroute.org. 30 Jan. 2003 <<http://www.traceroute.org>>.
 - [5] Moore, David, Jim Donohoe, and Ram Periakaruppan. Where in the World is netgeo.caida.org? Cooperative Association for Internet Data Analysis. 30 Jan. 2003 <http://www.caida.org/outreach/papers/2000/inet_netgeo/inet_netgeo.html>.
 - [6] Nemeth, Evi, and Ram Periakaruppan. "GTrace - A Graphical Traceroute Tool." (n.d.). 30 Jan. 2003 <<http://www.caida.org/tools/visualization/gtrace/paper/GTrace.pdf>>.
 - [7] SarangWorld Traceroute Project. Comp. Sarang Gupta. 9 Mar. 2003. 30 Jan. 2003 <<http://www.sarangworld.com/TRACEROUTE/>>.