Geographically Speaking

Ing-wher Chen
Wen W. Chiang
Syed Adnan
Lucas Silacci

CSE 222 Winter 2001
University of California, San Diego

1 Abstract:

It is becoming increasingly important for Internet designers to have a better understanding of dynamically changing Internet. One important area in Internet behavior studies is Wide Area Network IP traffic patterns, which include traffic distribution, traffic flow, and routing behavior, among other things. In this paper, we have examined the WAN IP traffic pattern using traces collected by the Cooperative Association for Internet Data Analysis (CAIDA). Our analysis of the traces is centered on the following questions: (1) Geographically, what countries/states/cities are the biggest sources and destinations of IP traffic? (2) Where is the traffic from/to a particular geographic source/destination flowing? (i.e. who is talking to whom?) (3) How far does IP-traffic travel in relation to the actual distances between source and destination?

After analyzing the traces, we have observed that for any given geographical region (city/state/country) there is a small number of regions that are the dominant sources/destinations of the given region. Furthermore, after looking at all the destinations and sources of a few well-chosen regions, we have observed that state and city level traffic tends to travel outside its own state and city. However, at the country level, most of the traffic tends to stay within the country.

2 Analysis Tools and Data Set

The traces for this study were collected from the NASA AMES Internet exchange (AIX) in Mountain View, CA as part of an NSF/NASA collaborative effort with NLANR/MOAT. The trace data contains approximately 3.6 millions IP flow traces, which is collected over an hour of monitoring time.
We wrote several PERL scripts, patterned after CoralReef [3] applications, to parse the raw data to a set of entries, where each entry has the format of source IP address, destination IP address, source network number, destination network number, source AS number, and destination AS number. Then we took unique network numbers and used them as input to NetGeo [4] to obtain information on each network, including geographic location. Based on the NetGeo output, we built a hash table with the network numbers as the keys and geographical data such as longitude, latitude, country, state and city as the values.

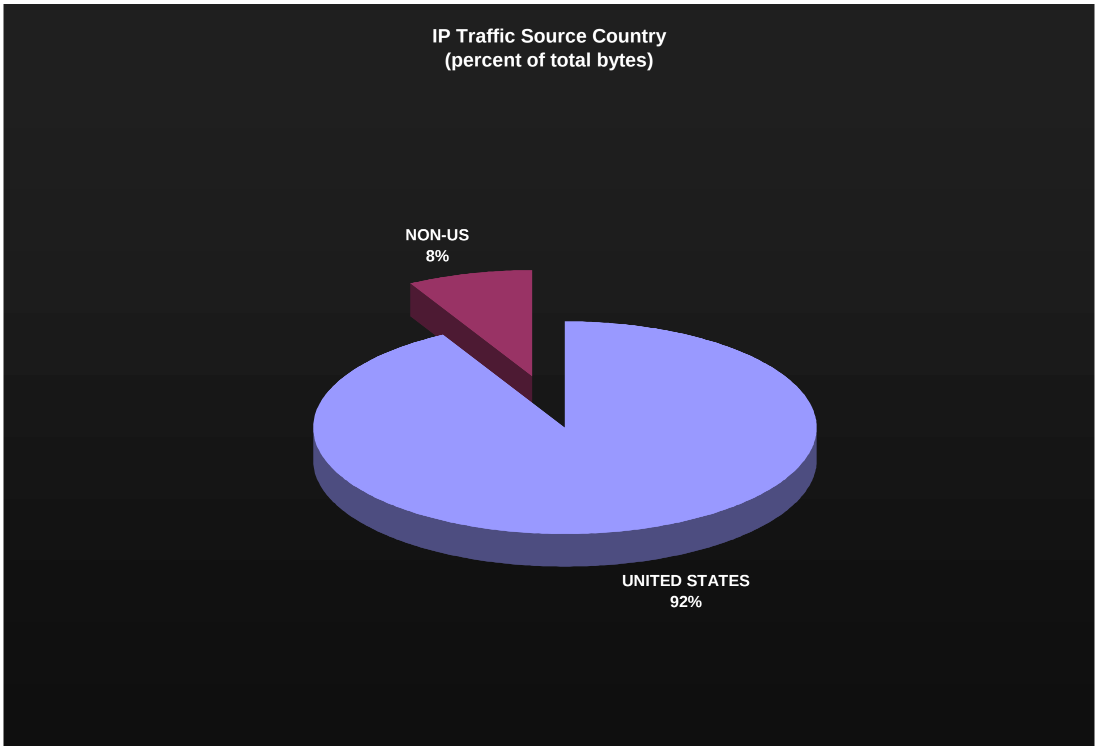
First off, we looked at the traffic distributions across the entire trace at the levels of countries, states, and cities. By looking at the results of that study, we were then able to choose a few countries, states, and cities with a large amount of the overall byte flows as either a source or a destination. This narrows our focus to a single region and allows us to analyze the behavior of the traffic for that one region. Lastly, we look at the physical distances between the source and destination for each flow and compared it with a computed lower bound of the actual distance traveled by packets in that flow. In all of this it is interesting to note that we only observed inter-AS traffic. Since the trace was taken at an Internet Exchange Point, we have no visibility into traffic which does not cross an AS boundary.

3 Traffic Distribution

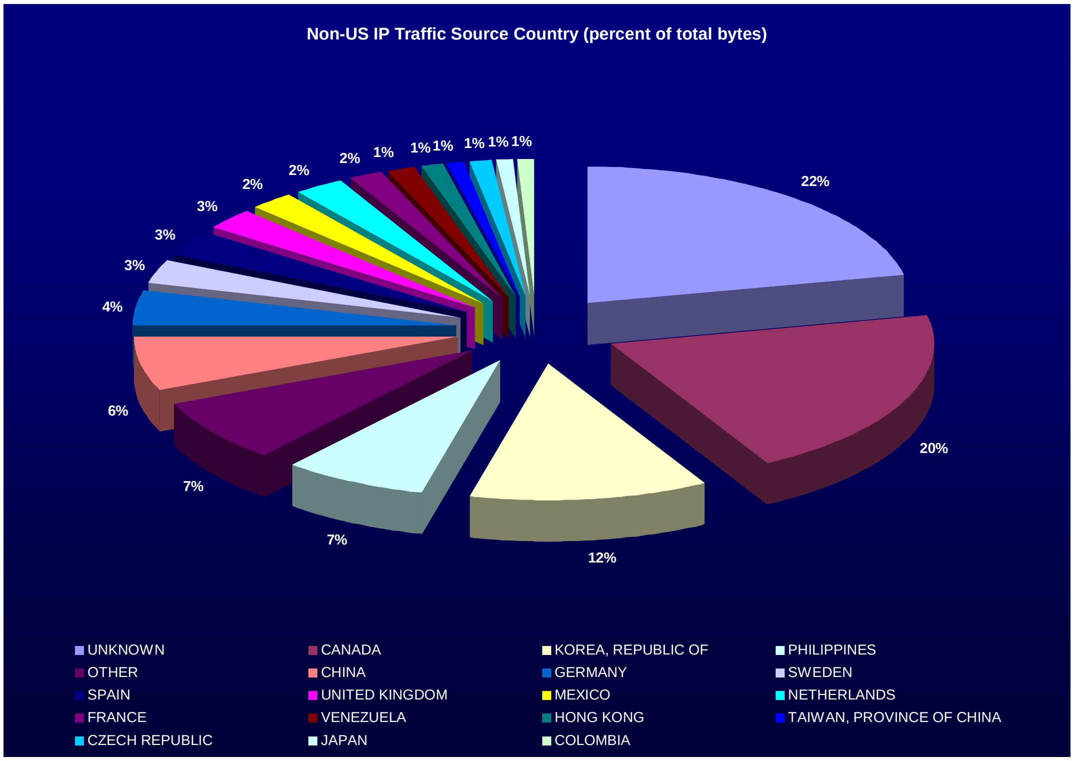
We start the analysis of the traces we collected by looking at the overall geographic distribution of the traffic. This analysis gives us the overall picture of where bytes are flowing during the time the trace was taken. Specifically we look at the distribution of distinct geographical locations for the source and destination of all the IP packets in the trace. This will allow us to zero in on the geographic regions that are generating or consuming the most data during the trace period.

3.1 Source Country

Looking at the byte-traffic in the trace, we find that hosts in the United States account for 92% of all the byte-traffic (by source). This is not surprising, considering that the trace was gathered in the San Francisco Bay Area.

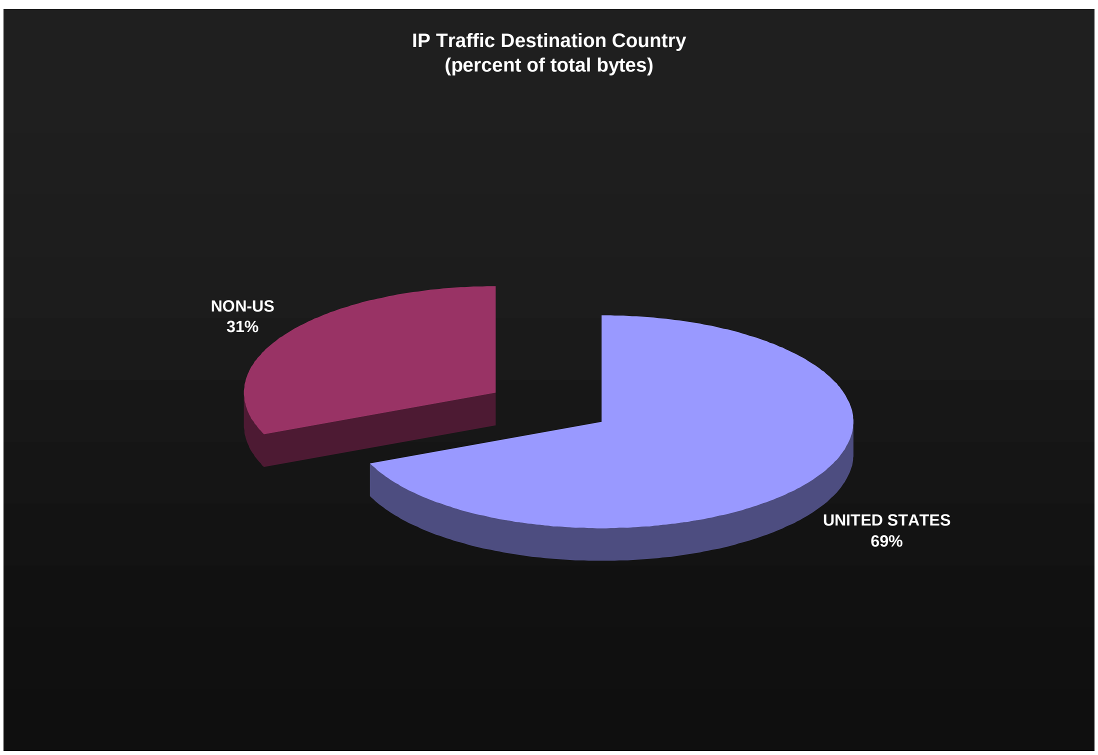


The chart below represents the breakdown of 8% of the traffic that was not generated by hosts in the United States. The largest byte-traffic generating country after the United States is Japan, at less than 2% (22% of 8%), followed by Canada, China, Korea, and the Philippines. We see here is that hosts in Asia dominate the other non-US hosts by a fair amount.

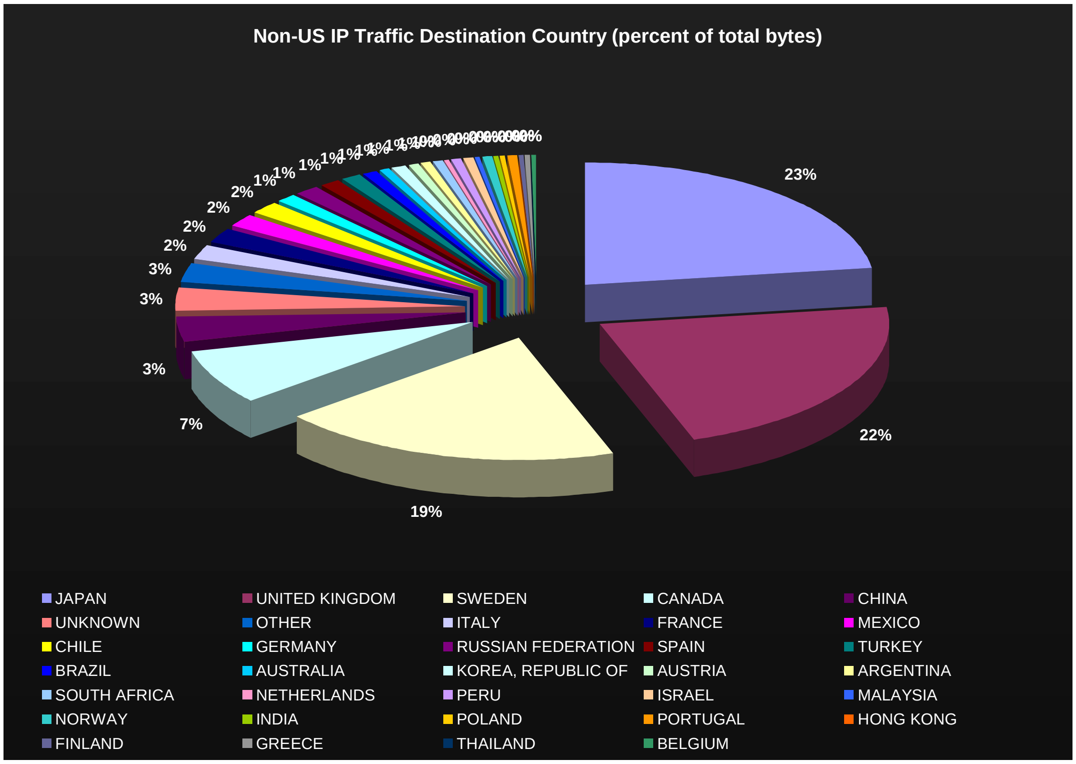


3.2 Destination Country

In comparison with the previous graphs, hosts in the United States are the destinations of only 69% of the byte-traffic. This means that there are significantly more hosts pulling data from the United States than there are hosts in the United States pulling data from the outside.



Again, Japan is the second biggest country (now as a destination) with (31% x 23%) ~7% of the bytes headed in that direction. The geographical locations of non-US hosts by destination are much more scattered around the world than the locations in the source graph.

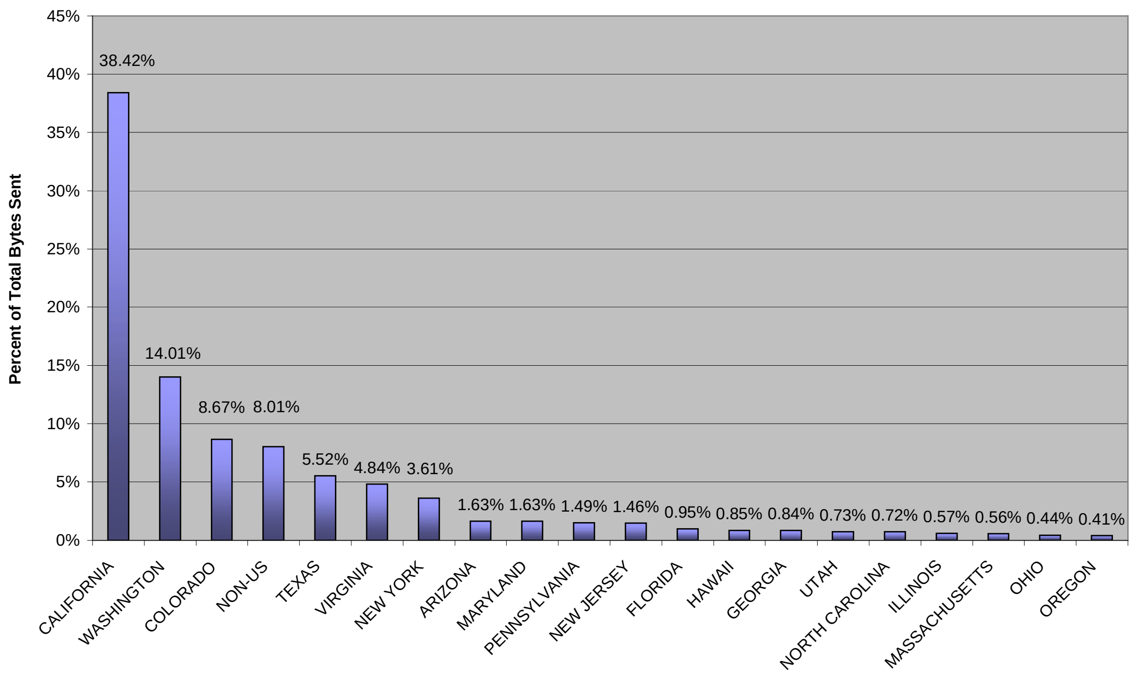


So the United States is by far (an order of magnitude) the biggest source and biggest destination on the Internet in our trace. The next sections investigate this further by looking at the United States traffic in more detail.

3.3 Top 20 Source States in the United States

As can be seen in the graph below, 38% of the traffic in the trace is coming from California when broken down by state within the United States (all of the non-US traffic is shown for comparison). Again, this is not surprising since the trace was taken in the San Francisco Bay Area. Washington and Colorado are relatively close to California geographically so having them at positions two and three makes sense as well.

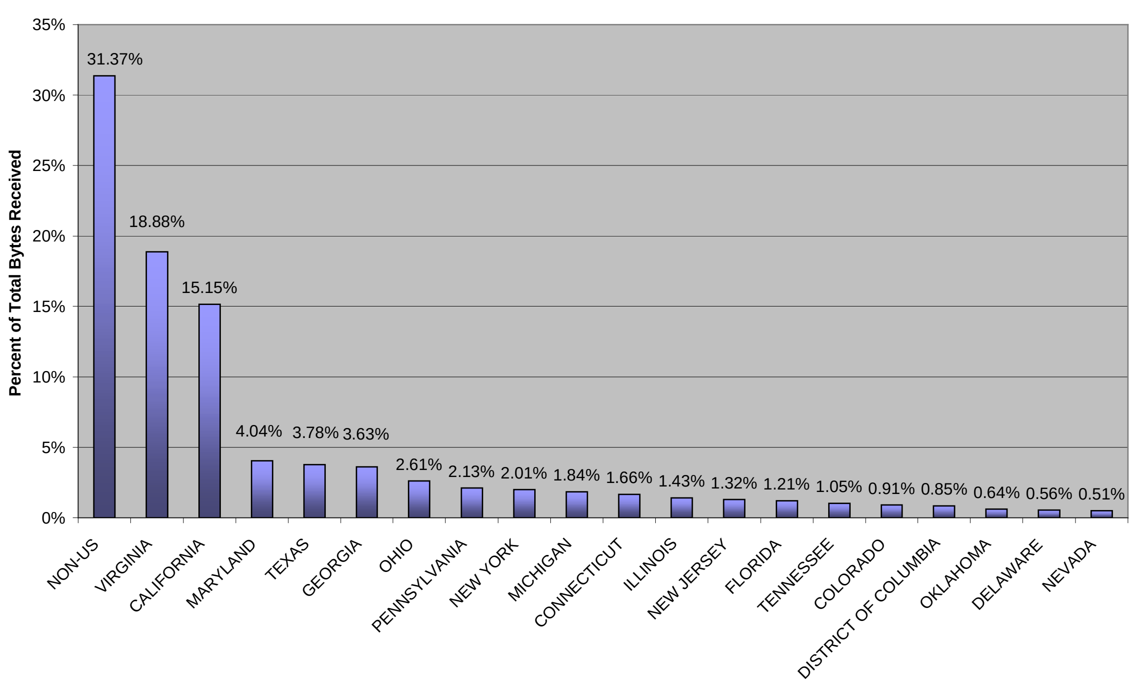
Top 20 Sources by State



3.4 Top 20 Destination States in the United States

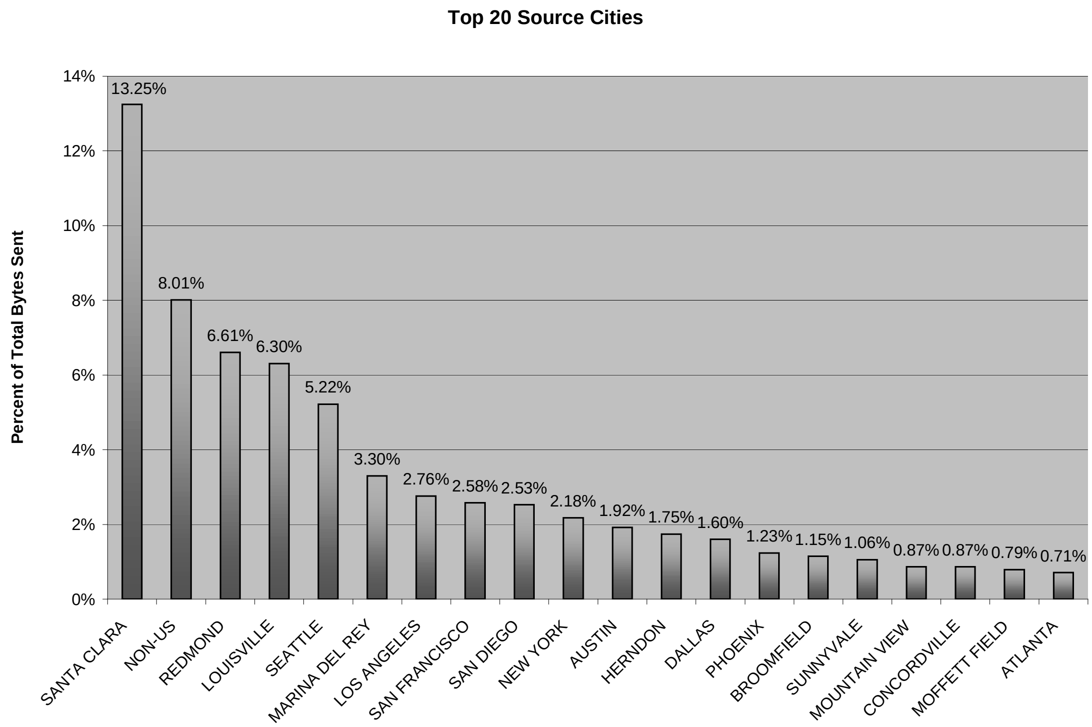
The interesting thing to note about the graph below is that although the US was by far the biggest destination of packets (69%), non-US hosts still received more bytes than any single US state. This could be an indication of an imbalance in the number of web servers that are located in the United States versus the numbers located in other countries. Furthermore, since the trace was taken in the San Francisco Bay Area, one would expect California to be the number one destination state in the US, but it was edged out by Virginia.

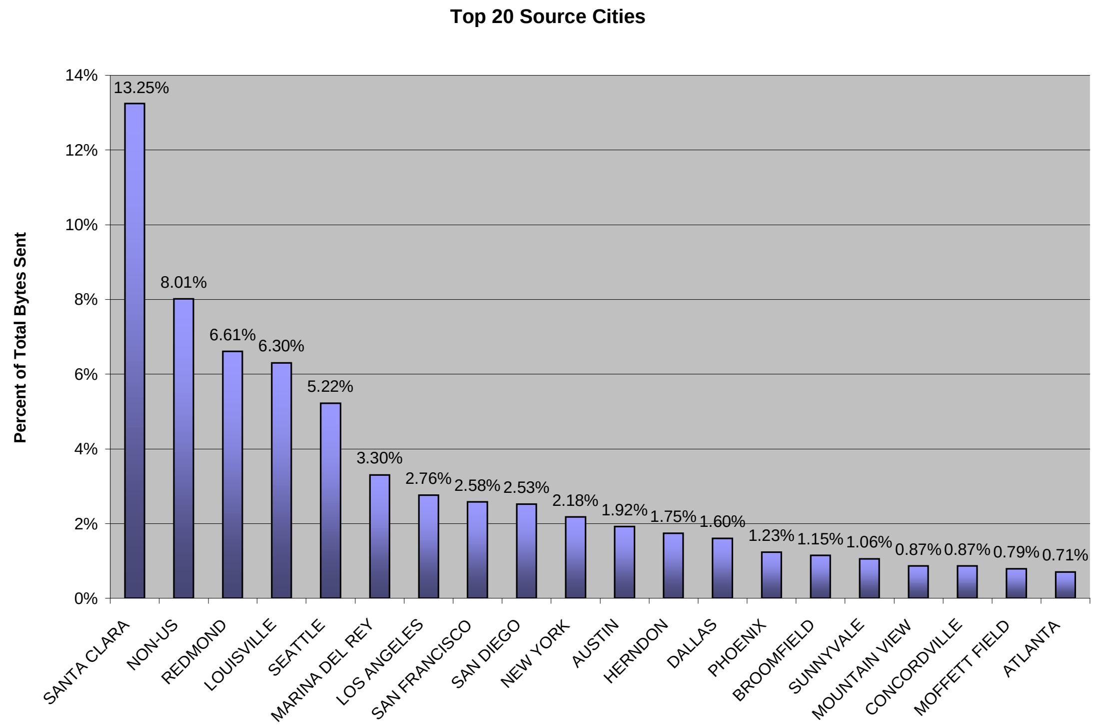
Top 20 Destinations by State



3.5 Top 20 Source Cities in the United States

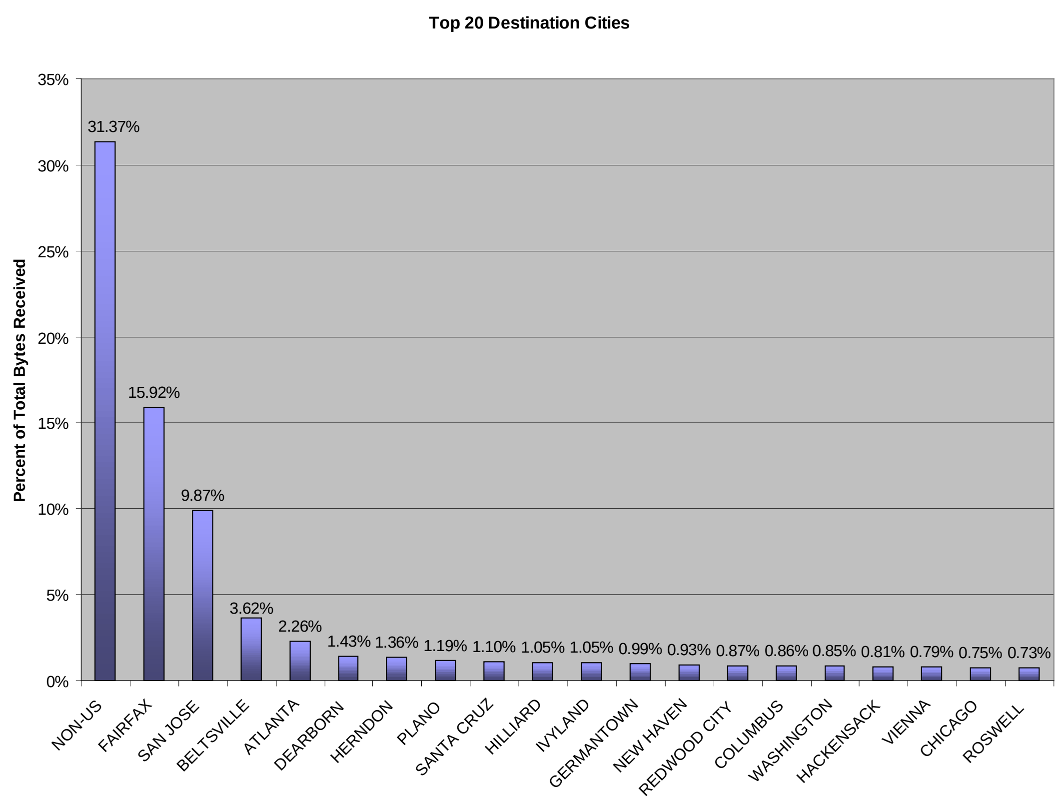
By looking at the source state graph above, you would expect California cities to dominate the top 20 list of US cities. In the graph below, we can see that California does dominate the traffic when broken down by city. San Francisco Bay Area cities are the most dominant, as is expected given the source of the trace. Again, Washington is near the top due to high volumes of traffic from Redmond and Seattle.





3.6 Top 20 destination cities in the United States

Looking at the destination state graph gives a good idea for what the top 20 destination cities graph will look like. Fairfax, Virginia is at the top (just as in the state graph) of the list, followed by San Jose and other random cities. California makes a strong showing again, but it does not dominate this graph as much as it does others.



If we assume that the trace is representative of all Internet traffic flowing through Internet Exchange Points, then the above data seems to point at a very unbalanced Internet for inter-AS traffic. Hosts in the United States are much more likely (by an order of magnitude) to be the source or destination of byte-traffic than all other countries combined. Furthermore, the United States itself is much more likely to be the source than the destination of byte-traffic.

4 Geographic Traffic Flow Analysis

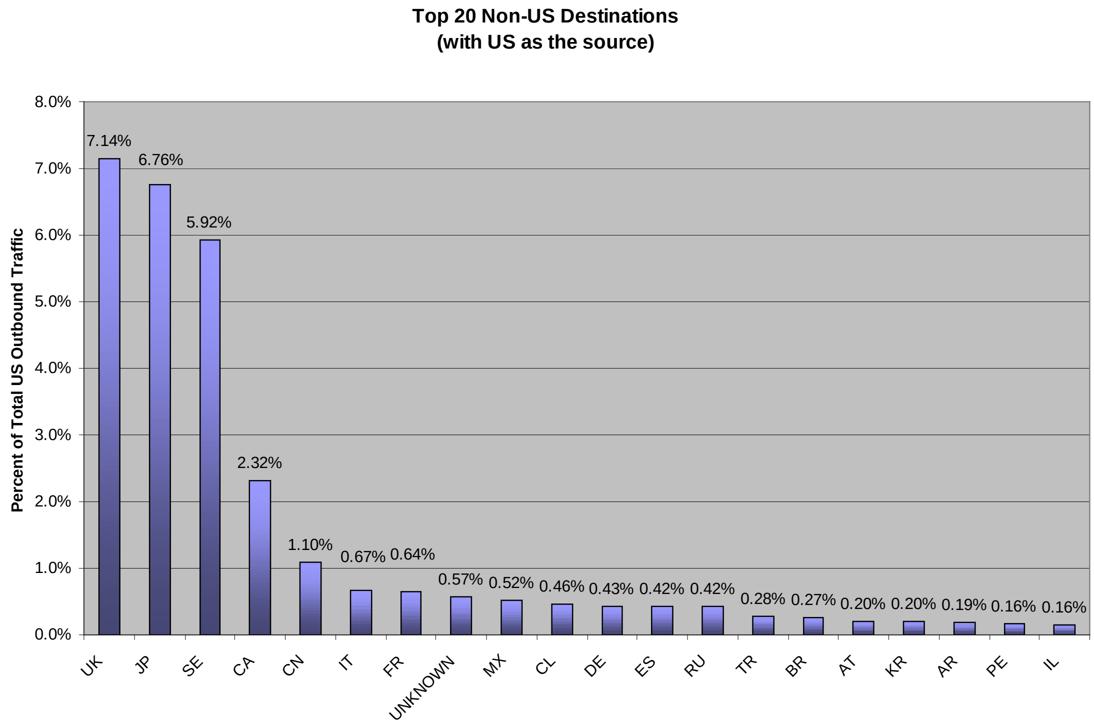
Geographic traffic flow analysis is similar to the 80/20 studies from AT&T. The 80/20 study from AT&T claims that 80% of the traffic within an AS does not cross the AS boundaries. There are two major differences in our study compared to AT&T study. First, the AT&T study was based on traffic originating from within the AT&T AS, and that study only analyzed how one AS would behave. In contrast to that study, we analyzed traffic behavior at city, state, and country levels. Second, instead of looking at source and destination IP addresses, we looked at the data from the viewpoint of geographic location, giving us a different view of behavior of Internet. In our view, a geographical region such as a state could be compared to the AS of the 80/20 study. In this study, an AS is treated as a single host within a geographical region since we are only looking at inter-AS traffic.

4.1 Source analysis

In source analysis we take a look at the distribution of destinations from a particular geographical region. In this way, we can observe where the data is flowing from the chosen region. We can then see if the given region follows or violates the 80/20 argument where we treat the region as a psuedo-AS. Specifically, at the country level we focused our attention on the United States, Japan, and Korea as individual AS's. At the state level, we focused on California and New York. Finally at the city level we focused on San Francisco and Santa Clara. In each case, the region was chosen using two factors: 1) The amount of traffic from the region in our trace, and 2) geographically how close (or far) it was from where our trace was taken.

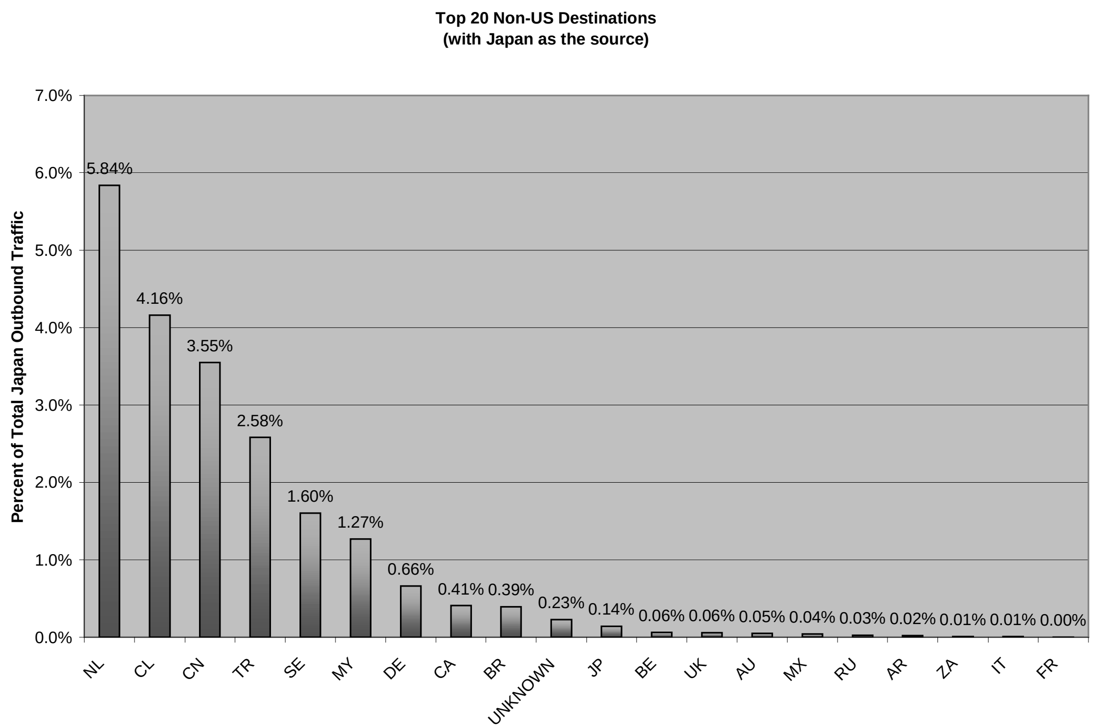
4.1.1 United States

As expected in the 80/20 argument, the United States sends 70% of its data back into the United States. The other 30% is dominated by three nations, Japan, the United Kingdom, and Sweden. After that most other countries are just in the noise. Still, a significant majority of the traffic from the United States was destined to stay within the United States.



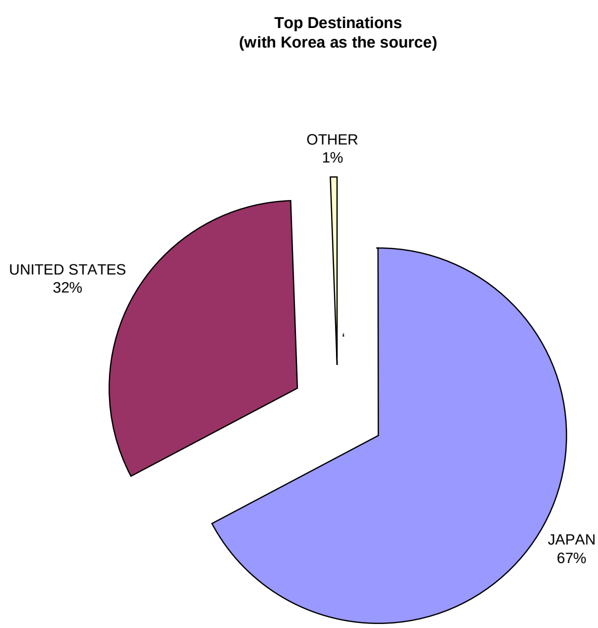
4.1.2 JAPAN

80% Japan's traffic that went through the Internet Exchange Point of our trace was actually destined for the United States. Given the fact that we sampled the data in California, we were expecting a slightly higher percentage. What is odd is that 3% of traffic from Japan in the trace was headed towards China (through California!), with a very small amount of the traffic was destined to end up back in Japan where source and destination only 22 miles apart physically.



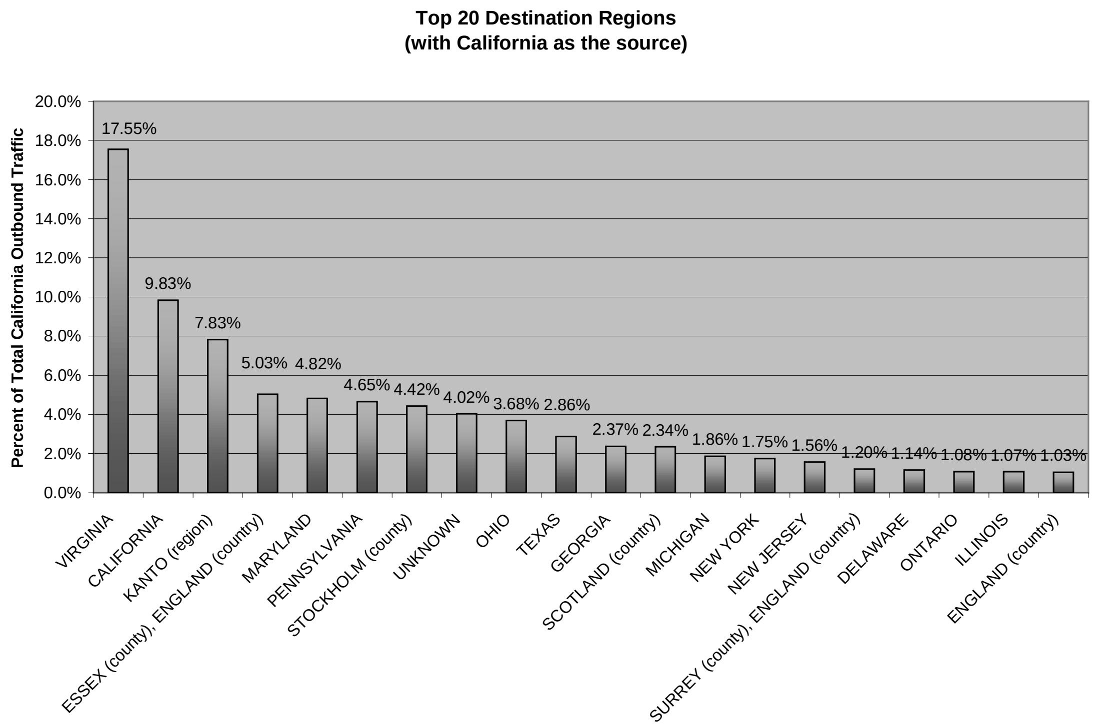
4.1.3 Korea

Nearly 70% of traffic from Korea in the trace was actually headed towards Japan. This is interesting in that it creates a most likely unnecessary burden on routers in California to handle traffic that should not even be on this continent.



4.1.4 California

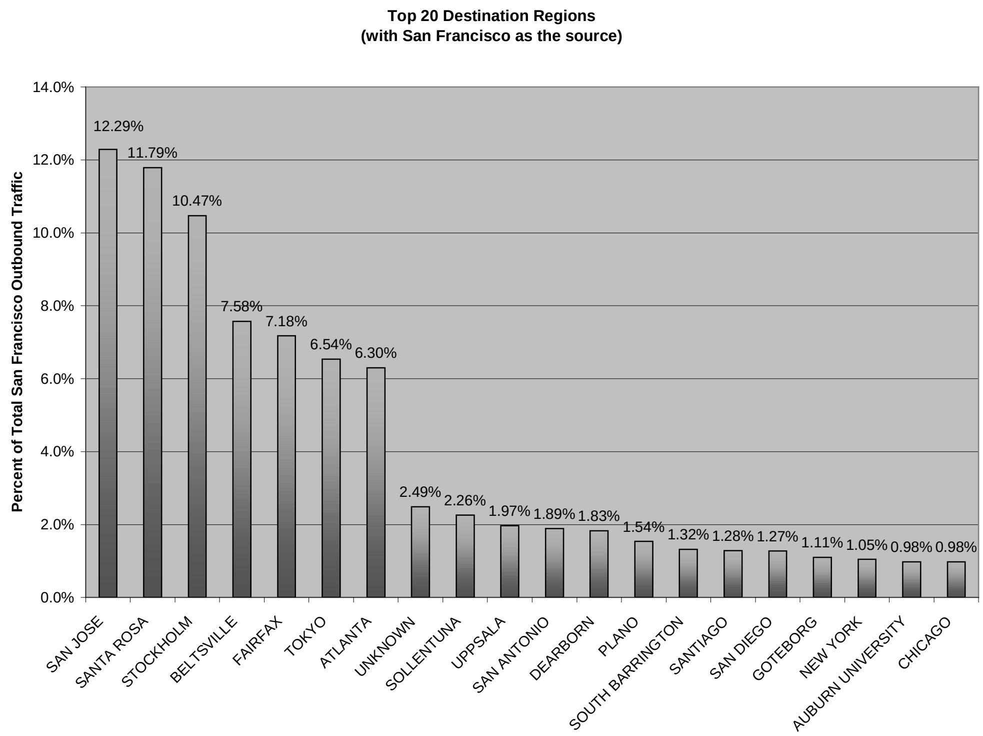
In the state level version of the 80/20 argument, we narrowed our focus on all the traffic generated by California. In this state-level pseudo-AS, we see that 17.5% of California's traffic is moving towards Virginia, while only 10% stays within California. Assuming our trace is representative of a border router of the California pseudo-AS, then we see that the 80/20 argument does not hold at this level.



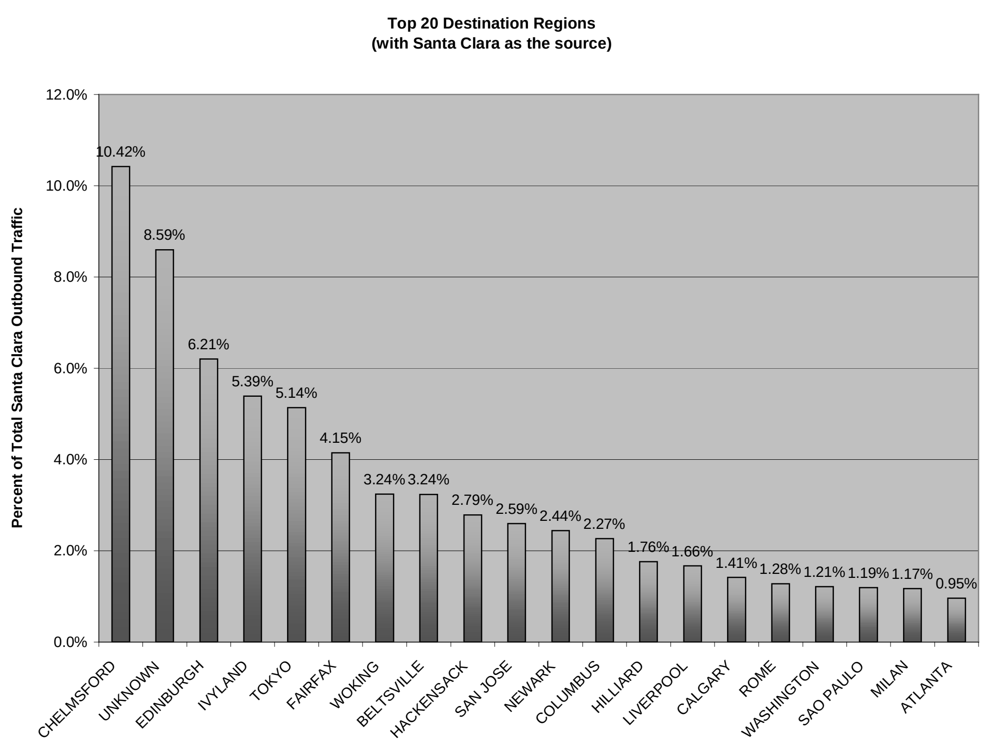
4.1.5 Bay Area

Looking at the graphs of destinations for byte-traffic from San Francisco Bay Area hosts we see a more even distribution then we saw in the previous graphs. In the previous graphs, we see domination of the distributions in just a few other geographical regions. Down at the city level, this distribution is more spread out. Also, we can see a trend, where nearby geographical neighbors are the ones at the top of the lists as we narrow our focus down to the city level.

For San Francisco, only 0.15% of traffic generated stayed within the city. Only 20% the traffic ended up in neighboring cities such as SANTA ROSA and SAN JOSE, where the average distance between source and destination is around 111 miles. Most of the traffic was headed towards other cities that are about 1000 to 4000 miles away. Seattle and Redmond had a similar pattern to that shown below, with most of their traffic heading for destinations over 1000 miles away.



Santa Clara was the single largest source city in our trace. Not a surprise since the trace was taken at NASA AMES. This distribution is all over the geographical map. Again, the 80/20 argument does not seem to apply to inter-city traffic.

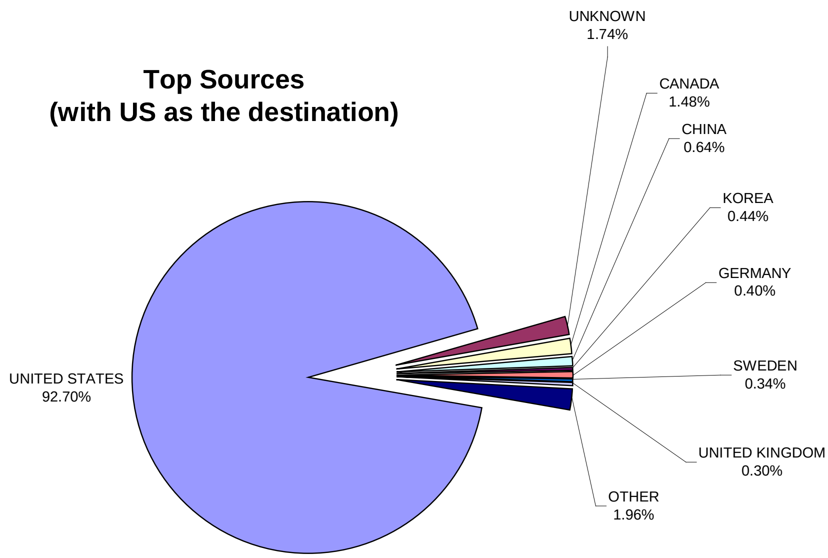


4.2 **Destination Analysis:**

In destination analysis we observed the distributions of sources from a particular geographical region. We can then combine this with the above source analysis to get a complete picture of the traffic patterns for a particular geographical region.

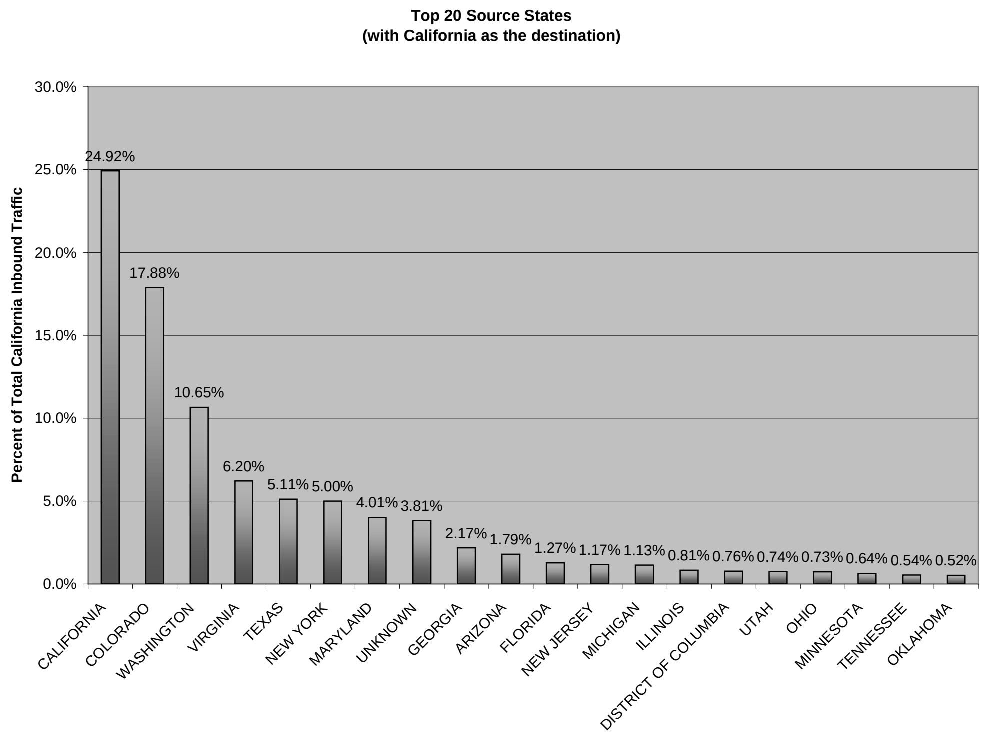
4.2.1 United States

Looking at the US as a large pseudo-AS, we find that most of its traffic did not leave the AS (93% stayed within the US). Like the source based analysis on the United States above, this follows the 80/20 argument and perhaps shows that an even stronger statement of 90/10 could be made at the country level.



4.2.2 California

25% of the bytes destined for California originated in California. 17% of California's traffic in the trace came from Colorado and 10% from Washington. It is evident from this graph that at most 10 states could be picked out as the sources of nearly all traffic bound for California. If we assume our trace is representative of traces for border routers at the state level in the US, then this could point at a small subset of states within the United States as being the most active in the Internet.



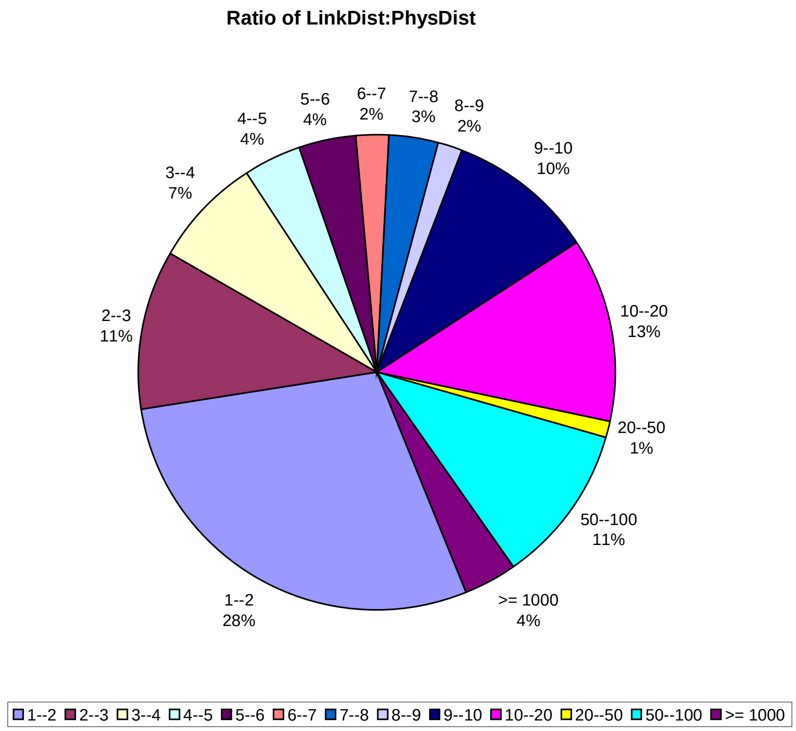
4.3 Traffic Flow Analysis Conclusions

Overall, we can see that at country level it is obvious that most of the data generated by United States, stays within the boundaries of the United States. Therefore, in a sense, the 80/20 argument holds at country level. Furthermore, from the trace we observed that perhaps a stronger 90/10 argument could be made at the country level. At the state and city level, the distribution of bytes is not dominated by certain states/cities. We focused on California (for state level) and cities in the San Francisco Bay Area (for city level) to determine their traffic behavior. From this we observed that the 80/20 argument did not hold at those levels.

There was a noticeable trend as we focused our observations down from the country to state to city levels. This trend was that traffic from or to a region tended to be dominated by only a few regions at the country level, and that the number of regions in the distributions increased as the focus was narrowed down to the city level. Perhaps this can be explained away by the fact that there are simply a greater number of cities connected to the Internet than states or countries.

5 Wrong Way

After using NetGeo to retrieve estimates of the source and destination geographical locations (latitude, longitude, city, state, and country), we found some interesting pairs of sources and destinations such as a New York source to a New York destination. This is rather surprising considering that the trace was captured in the Bay Area. This prompted a closer look at how much distance each flow was actually traveling versus how far the physical distance between the source and destination was. The figure below is a break down of the ratios of link distance (actual distance traveled) and physical distance. Roughly a quarter of the traffic was traveling 10 times the distance it should have traveled!



The data above was collected by calculating the ratio between minimum link distances and minimum physical distances for each flow. The minimum link distance is simply the sum of the geographical distances between the point that the trace was taken and the source and destination endpoints of each flow, while the minimum physical distance is simply the geographical distance between the source and destination of the flow. It is not surprising to see that all the traffic flows traveled more than their physical distances. This is expected due to the fact that traffic does not get routed in a straight line. By taking a closer look at the source and destination pairs, we are able to rationalize this surprising distribution.

From the previous section, we discovered that the US by far is the largest source during this trace. Within the US, we discovered that more than half of the top 16 largest sources are eastern states. By looking into each of these eastern states (VA, NY, PENN, NJ, FL, GA, NC, MA, OH), we find that there is a large number of flows coming from these states that travel more than 10 times the physical distance between the source and destination. (The table below illustrates how much of the traffic is traveling.)

State	**% of flows traveling ≥ 10 times physical distance**
NY	46.5
VA	36.0
NJ	26.3
PENN	23.2
NC	7.7
OH	4.7
GA	4.0
MA	3.9
FL	0.4

In particular, we find that most flows coming out of eastern states and going through our trace point have destinations in eastern states; some even have both source and destination in the same city! A possible explanation is that we are seeing the effects of hot potato routing. Perhaps each AS is simply passing traffic on to the next AS and so on until it reaches the West Coast and has to travel all the way back to the East Coast. Another explanation may be due to the estimation of the geographical locations based on DNS registries, which is the method used by NetGeo. DNS registries are often not up-to-date and the data entered is not well specified. It is not uncommon to see large ISP's, based in a city on the East Coast but with routers at various locations across the country, register all their sub-nets as being located in the base city. In this case, NetGeo sees traffic coming from VA and going back to VA via a router in the Bay Area. In reality, maybe the traffic comes from a router in CA and is routed to a destination router physically in CA but both routers are registered as being in VA.

Another interesting discovery is related to the distance ratios for the traffic within Asia. By looking at traffic generated from China, Japan, Korea, and Taiwan (these 4 countries are picked because they have a significant amount of traffic), we find that out of a total of 152297 flows generated from these countries, 18243 of them are eventually routed to a country among them, through our trace-point in the Bay Area. In other words, in the traces we analyzed, roughly 12% of the traffic coming out of Asia is routed to a destination within Asia. In this case, the arguments for the extra distance traveled for traffic originating from eastern states perhaps does not apply here. The more plausible reason is a less-than-perfect connectivity within Asia. For traffic with both source and destination in Asia, perhaps it takes longer to route the traffic within Asia than routing it through the United States due to poor connectivity within Asia.

6 Conclusions

It was rare to see a state/city to be dominated by its own traffic either in sources or in destination, which leads us to believe that 80/20 comparison does not apply at city and state levels. We did notice that California was one of the states that directed its data back to California relatively high amount of the time (around 30%). We know that the trace were taken from California, so that could be a factor in that result. When looking at the United States as a pseudo-AS, however, we can see strong indications that the 80/20 argument is valid. In fact there is a possible 90/10 argument that can be made as well.

After analyzing the traces, we have observed that for any given geographical region (city/state/country) there is a small number of regions that are the dominant sources/destinations of the given region. Furthermore, after looking at all the destinations and sources of a few well-chosen regions, we have observed that state and city level traffic tends to travel outside its own state and city. However, at the country level, most of the traffic tends to stay within the country. So even though 80/20 did not seem to hold at the city and state levels, there seemed to be very small subsets of states/cities that tended to dominate the traffic flow to and from a particular state or city.

This trace also brought to light some issues with inefficient routing. In many cases we observed flows which traveled more than 10 times the distance than they needed to. The longer the distance a packet travels in the Internet, the longer it stays in the network. This increases the probability that the packet will be dropped and decreases the amount of traffic that can be flowing from source to destination. Further study needs to be done to determine if these are true inefficiencies that cause wasted resources.

7 References

1. *Trends in Wide Area IP Traffic Patterns – A View from the Ames Internet Exchange*, Monterey, May 2000, Sean McCreary and KC Claffy, University of California, San Diego, CAIDA.
2. *Measuring the Internet*, IEEE Internet Computing Online, Jan 2000, kc claffy, UC San Diego Supercomputer Center.
3. [CoralReef] CoralReef home page, http://www.caida.org/Tools/CoralReef
4. [NetGeo] NetGeo home page, http://www.caida.org/Tools/NetGeo