On the Geographic Location of Internet Resources

Short Abstract

Anukool Lakhina John W. Byers Mark Crovella Ibrahim Matta
Department of Computer Science
Boston University
{anukool, byers, crovella, matta}@cs.bu.edu

I. Introduction

One relatively unexplored question about the Internet's physical structure concerns the geographical location of its components: routers and links. We study this question using two large inventories of Internet routers and links. We first map each router to its geographical location using two different state-of-the-art tools. We then study the relationship between router location and population density and between geographic distance and link density.

The choice of this question is motivated by current problems in network topology generation. We turn to geography for inspiration because a number of unsolved problems in topology generation appear much easier to solve given an underlying geographical model. For example, an accurate geometric model of router placement and link formation would make the labelling of links with latency values a straightforward matter. Our goal is not to propose a new topology generation method in this paper, but to suggest a wider set of bases for the construction of topology generation tools.

II. Methodology

Our first topology dataset is a large collection of ICMP traceroute probes collected between December 26, 2001 and January 1, 2002 by Skitter [3], a measurement tool run by CAIDA on more than 20 monitors around the world to a predetermined destination list. This dataset has 563,521 router interfaces and 862,933 links. Our second dataset was collected during August 1999 by Mercator [4]. Unlike Skitter, Mercator is run from a single host to a heuristically determined destination address space. Our Mercator dataset is considerably smaller at 228,263 routers and 320,149 links. An important distinction between maps generated by Mercator and Skitter is that the former generates a map of routers, while the latter generates maps of interfaces. Despite this difference, our conclusions seem robust whether expressed in terms of routers or interfaces. We next draw on two different state-of-the-art geographic mapping tools to identify IP addresses with their geographical longitude and latitude: Ixia's *IxMapper* [6] and Akamai's *EdgeScape* [1]. IxMapper uses hostname based mapping techniques which exploit the router naming conventions used by ISPs. EdgeScape supplements hostname based mapping with Akamai's access to internal ISP geographical information.

Our principle findings are consistent across three economically homoegenous regions of the world (USA, Europe, Japan), across both sources of data, and across the two geographic mapping methods. However, due to space limitations, we only present results obtained from IxMapper and Skitter in the USA. A full version of this paper is available as [8].

This work was partially supported by NSF research grants CCR-9706685, ANI-9986397, ANI-0095988, and CAREER award ANI-0093296. Some of the data used in this research was collected as part of CAIDA's skitter initiative, http://www.caida.org. Support for skitter is provided by DARPA, NSF, and CAIDA membership.

III. Routers and Population

Focusing on the economically homogeneous regions allows us to ask how router density relates to population density. To answer this question, we subdivided each region into patches of size 75 arc-minutes $\times$ 75 arc-minutes. At the latitudes studied, this creates patches about 90 miles on a side. Within each patch, we tally the population (obtained from [2]) and the number of routers or interfaces.

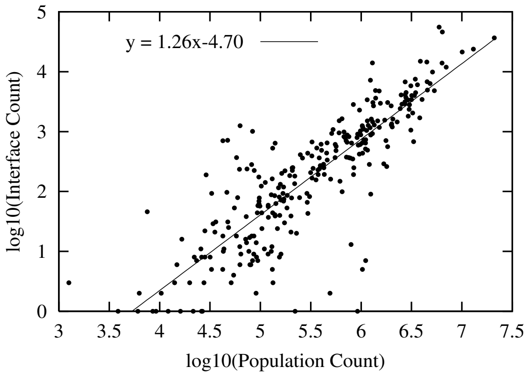

Fig. 1. Router Interface Density vs Population Density

For the USA region, this result is plotted as log-log scale in Figure 1 and shows a strong relationship between infrastructure and population density. Although the plot appears roughly linear on log-log axes, the precise functional relationship between population density and router density is difficult to identify from the data because of the significant amount of noise, and the relatively limited range of scales available. For example, it would be hard to distinguish a $n \log n$ relationship from a power law relationship for the data. Nonetheless, we conclude that router/interface density clearly bears a *superlinear* relationship to population density (slope of the fitted line is larger than 1). This surprising result indicates that the number of routers or interfaces per person is *higher* in areas of high population density (population centers). Furthermore, it seems reasonable to use a simple power law relationship as an approximation; that is, over the limited range of data studied, we can approximately model router or interface density R and population density P as related by $R \sim P^{\alpha}$ with α varying from 1.2 to 1.7 across the regions studied, based on the slopes of the fitted lines. This result may be interpreted as a consequence of simple scaling effects: as the number of network users n in a region grows, the number of potential connections between pairs of users grows via an n^2 law. If the capacity of individual switches does not scale accordingly, then in order to provide acceptable service it becomes necessary to add switches in a superlinear fashion. Thus, *e.g.,* multistage interconnection networks for multiprocessor computers are often designed to scale in $n \log n$ fashion [7], [5].

IV. LINKS AND DISTANCE

Given an understanding of how routers are distributed over the Earth's surface, we next examine the geographical properties of node-node links. To do so we measure the empirical probability that two routers separated by great-circle distance d, are directly connected. For any pair of routers separated by distance d let C be the event that the two routers are directly connected. Then we are interested in estimating the likelihood function: $f(d) = P[C|d]$. We call this the *distance preference function.* We estimate this function by placing the data into bins of size b. Then we form the empirical distance preference function as:

$$\hat{f}(d) = \frac{\text{\# links with length in } [d, d+b)}{\text{\# node pairs with distance in } [d, d+b)} \tag{1}$$

for values of d that are multiples of b.

Broadly speaking, we find that $f(d)$ shows two regimes: for short distances, $f(d)$ declines with distance; while for longer distances, $f(d)$ seems nearly constant. To explore this relationship further, we break the data up into two regions, "small d" and "large d".

For the US, this demarcation point is at approximately 250 miles. Focusing first on small d, we plot $\ln(f(d))$ vs. d in Figure 2. Surprisingly, this plot shows a linear tendency on the semi-log axes, suggestive of an exponentially declining function. In fact, the linear fit can be characterized in terms of Waxman's method for topology generation [10]. In the Waxman model, the probability that two nodes are connected $f_W(d)$ is: $f_W(d) = \beta \exp(-d/\alpha L)$ where L is the maximum distance between nodes, $0 < \alpha \leq 1$ is the sensitivity of link formation to distance, and $0 < \beta \leq 1$ controls link density. In terms of the Waxman model, we find estimates of $\alpha L \approx 140$ miles for the US. This is not to suggest that the Waxman model is a correct model for the growth of the Internet over these distance ranges, but rather that it is surprisingly descriptive of the end

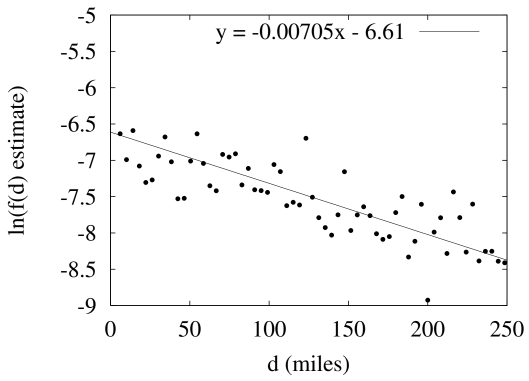

Fig. 2. Empirical Distance Preference Function, Small *d*, Semi-Log

result. In the other region (large d), the function $f(d)$ appears nearly constant, suggesting that the probability two routers are direcctly connected for large d is independent of their separation distance.

Furthermore, most links (from 75% to 95%) fall within the range of link lengths considered distance-sensitive. We conclude that distance sensitivity of router connectivity applies to the vast majority of router-router links in our datasets. On the other hand, we note that although a small fraction of routers are connected in a manner insensitive to distance, they are clearly not randomly connected, and their connections doubtless play an important structural role. In fact, work in [9] has shown that only a very small fraction of non-local links is needed to dramatically reduce the average diameter of an otherwise locally-connected graph.

V. CONCLUSIONS

Given the evident promise of geographically-based topology generation, we have presented in this paper a collection of results intended to bring that goal closer. These findings have many implications for the next generation of topology generators, which we envisage as producing router-level graphs annotated with attributes such as link latencies, AS identifiers and geographical locations.

REFERENCES

[1] Akamai Inc. At *http://www.akamai.com*.

[2] Center for International Earth Science Information Network (CIESIN), Columbia University. Gridded population of the world. Available at www.ciesin.org.

[3] Cooperative Association for Internet Data Analysis (CAIDA). The Skitter project. At *http://www.caida.org/Tools/Skitter*.

[4] R. Govindan and H. Tangmunarunkit. Heuristics for Internet Map Discovery. In *Proceedings of IEEE/INFOCOM'00*, March 2000.

[5] P. G. Harrison. Analytic models for multistage interconnection networks. *Journal of Parallel and Distributed Computing*, 12:357–369, 1991.

[6] IxMapper. At *http://www.ixiacom.com/products/*.

[7] C. Kruskal and M. Snir. The performance of multistage interconnection networks for multiprocessors. *IEEE Transactions on Computers*, 32(12):1091–1098, 1983.

[8] A. Lakhina, J. Byers, M. Crovella, and I. Matta. On the Geographic Location of Internet Resources. Technical Report BUCS-TR-2002-015, Boston University, 2002.

[9] D. J. Watts and S. H. Strogatz. Collective Dynamics of 'Small-World' Networks. *Nature*, pages 440–442, June 1998.

[10] B. Waxman. Routing of Multipoint Connections. *IEEE J. Select. Areas Commun.*, December 1988.