Measuring IP Address Fragmentation from BGP Routing Dynamics

Xia Yin[2], Xin Wu[1,2], and Zhiliang Wang[3]

[1] Tsinghua National Laboratory for Information Science and Technology
[2] Department of Computer Science & Technology, Tsinghua University, Beijing, China
[3] Network Research Center of Tsinghua University, Beijing, China
yxia@mail.tsinghua.edu.cn, tun-x03@mails.tsinghua.edu.cn,
wzl@cernet.edu.cn

Abstract. Address Fragmentation plays a key role in the exponential growth of DFZ routing table, known as the scalability problem of current Internet. In this paper, we measure the severity of address fragmentation, and try to figure out the relationship between Prefix-Distance and Network-Distance of current Internet by taking Geographic-Distance as an approximation of Network-Distance. We focus our measurement on the prefixes with relatively small Geographic-Distance, and get Prefix Groups from BGP routing dynamics. This method reduces the number of active probes required in active measurement, and results a more detailed prefixes' distribution analysis. We find out that there are two extreme allocations of IP address blocks in current Internet. Some of the blocks with small Geographic-Distances have small Prefix-Distances, while others with small Geographic-Distances have rather big Prefix-Distances. This is the direct reason of the BGP routing table's inflation. We further conclude that by reallocating IP address blocks according to geography, we could significantly reduce the size of the global routing tables.

Keywords: Prefix Distance, Geographic Distance, BGP.

1 Introduction

The global Internet consists of tens of thousands of autonomous systems (ASes). The Border Gateway Protocol (BGP) [1] is the defacto inter-domain routing protocol, which transfers reachability information, i.e. update messages, between ASes.

It is commonly recognized that current Internet routing and addressing system is facing serious scaling problems. The growth of Default Free Zone (DFZ, the collection of all Internet ASes that do not require a default route to route a packet to any destination) routing table size is at an alarming rate [2]. Internet Architecture Board (IAB) claims that there are four main driving forces behind the rapid growth of the DFZ RIB (Routing Information Base, an electronic database storing the routes and sometimes metrics to particular network destinations) [2]: suboptimal RIR address allocations, multihoming, traffic engineering and business events such as mergers and acquisitions, among which address allocation policies contribute significantly to routing table size.

P. Mueller, J.-N. Cao, and C.-L. Wang (Eds.): Infoscale 2009, LNICST 18, pp. 46–57, 2009.

Address Fragmentation refers to the phenomena that the set of prefixes originated by the same AS cannot be summarized by one prefix. More than 75% of BGP routing table items are because of address fragmentation [3]. Analyzing the composition and distribution of address fragmentation can enhance our understanding of address allocation and aggregation.

In this paper, we propose a method to analyze address fragmentation. More specifically, we try to figure out the relationship between Prefix-Distance and Network-Distance of current Internet by taking Geographic-Distance as an approximation of Network-Distance. Instead of active probing, we use only passive monitoring of BGP message. By clustering prefixes based on similarities between their routing event start times, we cluster prefixes into groups. Prefixes in one group always have close logical relationships, e.g. they may be connected to the same ISP Point of Presence (PoP) or located in the same place. If prefixes in one group also have small Prefix-Distance, i.e. their IP address blocks are adjacent, they may get the chance to be well aggregated. However, according to our measurement result, there are two extreme allocations of IP address blocks in current Internet. Some of the blocks with small Geographic-Distances have small Prefix-Distances, while others with small Geographic-Distances have rather big Prefix-Distances.

The rest of the paper is organized as follows. We summarize some related work in Section 2, state the problem formally in Section 3. We propose a method to measure address fragmentation in Section 4, and analyze measurement results in Section 5. Finally, in Section 6, we summarize this paper.

2 Related Works

David G. Andersen [4] utilizes BGP dynamics to infer network topology. By clustering prefixes based upon similarities between their update times, He infers logical relationship between network prefixes within an Autonomous System (AS). Comparing with active method, this passive measurement method reduces the number of active probes required in traditional traceroute-based Internet mapping mechanisms. We also measure IP address distribution from BGP dynamics, however, we cluster prefixes into groups according to the similarities between their routing event start times. This significantly reduces the complexity.

Tian Bu [5] explore various factors contribute to the routing table size and characterize the growth of each contribution, including multihoming, load balancing, address fragmentation, and failure to aggregate. They find out that the contribution of address fragmentation is the greatest and is three times that of multihoming or load balancing. The contribution of failure to aggregate is the least. In our work, we further find out that there are two extreme allocations of IP address blocks in current Internet. Some of the blocks with small Geographic-Distances have small Prefix-Distances, while others with small Geographic-Distances have rather big Prefix-Distances. This is the direct reason of the BGP routing table's inflation.

M. Freedman [6] measures the geographic locality of IP prefixes using traceroute. He concludes that (1) address allocation policies and granularity of routing contribute significantly to routing table size, and (2) the BGP routing table can get aggregated significantly by reallocating IP addresses according to geographic locality. Our work not only validates Freedman's conclusions, but also enhances his measurement methodology and results. Instead of active probing, we only utilize passive measurement method, which reduces active probe monitors significantly. Besides Freedman's findings, we also identify two extreme allocations of address blocks in current Internet. Blocks with relatively small Geo-Distances have either very small or large Prefix-Distances.

Xiaoqiao Meng [7] quantitatively characterize the IPv4 address allocations made between 1997 to 2004 and the global BGP routing table size changes during the same period of time. 45% of the address allocations during that period were split into fragments smaller than the original allocated blocks. He claimed that without these fragmentations, the current BGP table would have been about half of its current size.

3 Prefix-Distance, Network-Distance and Geographic-Distance

In this section, we propose the basic assumptions of our measurement, state our purpose formally and define some related parameters which will be referenced frequently in the following of this paper.

***Assumption 1**: hosts within an IP prefix are topologically close.*

This assumption is adopted by many researches [6, 8]. Even though shorter prefixes (large address blocks) tend to comprise more geographic locations [6], for most prefixes (/24 and longer) this assumption is reasonable [8].

***Assumption 2**: Prefixes that always have routing events simultaneously have close logical relationship, i.e. they are always connected to the same ISP Point of Presence (PoP) or located in the same place.*

A similar assumption first appeared in [4]. It clustered prefixes based upon similarities between their update times, but not routing event start times. According to [4]'s evaluation, in more than 95% of the case, prefixes belonging to the same group share the same ISP PoP. Our assumption is a little different from [4]. Comparing with update arriving time, we believe taking routing event start time as clustering standard is also reasonable. We evaluate this assumption in Section 5.

***Definition 1**: D_p, D_n, and D_g*

$\boldsymbol{D_p}$: The *Prefix-Distance* of two address blocks. Firstly, we list the IP address from 0.0.0.0 to 255.255.255.255 alphanumerically. Then suppose there are two address blocks A and B, which cover two parts of the array. $D_p(A, B)$ refers to the number of addresses between A and B. Fig. 1 illustrates the definition of $D_p(A, B)$. If two blocks are overlapped, $D_p = 0$.

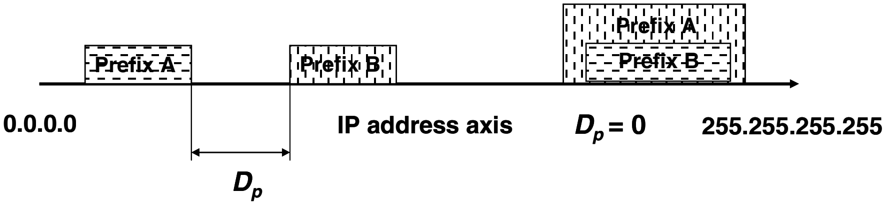

Fig. 1. Definition of D_p

D_n: The Network-Distance of two address blocks. *A* and *B* are two address blocks. i_A is one address in *A*, and j_B is one address in *B*. $t_{i_A j_B}$ stands for the round trip time (RTT) between i_A and j_B. n_A is the number of addresses in block *A*. n_B is the number of addresses in block *B*. The definition of D_n is shown in (1). Intuitively, it is the arithmetic mean value of all RTTs between *A* and *B* at the same time.

$$D_n(A,B) = \frac{1}{n_A n_B} \sum_{i_A \in A,\, j_B \in B} t_{i_A j_B} \tag{1}$$

D_g: The Geographic-Distance of two address blocks. *A* and *B* are two address blocks. i_A is one address in *A*, and j_B is one address in *B*. $d_{i_A j_B}$ stands for the shortest spherical distance between i_A and j_B. n_A and n_B are also the numbers of addresses in block *A* and *B*. The definition of $\boldsymbol{D_g}$ is shown in (2), which is the arithmetic mean value of all shortest spherical distances between *A* and *B*.

$$D_g(A,B) = \frac{1}{n_A n_B} \sum_{i_A \in A,\, j_B \in B} d_{i_A j_B} \tag{2}$$

We are trying to figure out the relationship between D_p and D_n. Ideally, the smaller D_p is, the smaller the D_n should be. That is to say, if two blocks have small Prefix-Distance, their Network-Distance should be small as well. This is good for aggregation. If two blocks have small Prefix-Distance, but their Network-Distance is large, these two blocks will not get aggregated in the routing table. As shown in Fig. 2, there are three districts in the D_p-D_n coordinates: Abnormal D_1, Normal D_2 and Abnormal D_3. If block *A* and *B*'s D_p-D_n point lies in Abnormal D_1, they have small Prefix-Distance but relatively big Network-Distance, and cannot get aggregated. On the other hand, if block *A* and *B*'s D_p-D_n point lies in Abnormal D_3, they have big Prefix-Distance but relatively small Network-Distance, and cannot get aggregated either. Only when *A* and *B*'s D_p-D_n point lies in Normal D_2, they may get the chance to get aggregated. The slopes of the two boundaries l_1 and l_2 are variable parameters in our method. We leave it for out future research.

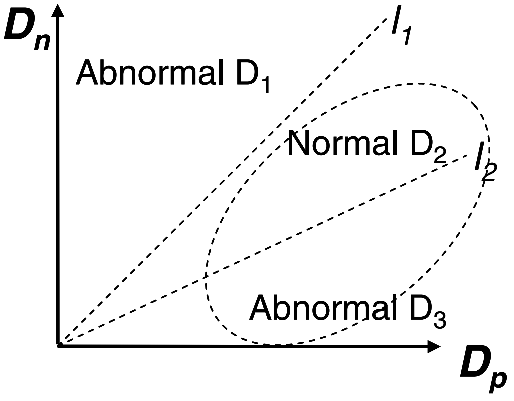

Fig. 2. D_p-D_n Coordinates

In this work, we focus on Normal D_2 and Abnormal D_3.(we leave Abnormal D_1 as our future work) The Network-Distance between any two address blocks is not trivial to obtain, thus we utilize Geographic-Distance D_g as an approximation of D_n. According to Oliveira [8]'s research, geographic distance is in direct proportion to end-to-end delay in most of the case, as a result this approximation is reasonable to some extend. There are also several other mechanisms estimating arbitrary end-to-end delays, such as IDMaps, GNP, Vivaldi and iPlane [11-14]. We plan to fit one of them into our system in the future.

4 Measuring D_p and D_g

4.1 Collect Prefix Groups

To collect prefix groups belonging to Normal D_2 and Abnormal D_3, we utilize the passive measurement method described in [4], but with a different prefix clustering algorithm. The basic assumption of this method is that prefixes that always have routing events simultaneously may have close relationship with each other (*Assumption 2*) . By grouping the prefixes that always have routing events together, we get the prefix groups that have relatively small D_n, which belong to Normal D_2 and Abnormal D_3.

There are two steps, as shown in Fig. 3, to get the prefix groups belonging to Normal D_2 and Abnormal D_3. The input is a time series of routing updates. An update is a BGP routing message that is specific to a prefix, such as an announcement or withdrawal. Each update contains a timestamp indicating the receive time and the prefix that is affected. The updates are ordered by timestamp. The first step is to use threshold T_1 to cluster BGP updates into *Routing Events*. Updates in one routing event are generated by the same reason, for example a link is down / up, or a router finds a better path to a destination. The second step is to group the prefixes that frequently have routing events in the same time window, whose width is T_2. The result of the clustering is groups of prefixes, each of which contains tightly correlated prefixes. Prefixes within one group always share administrative or topology features.

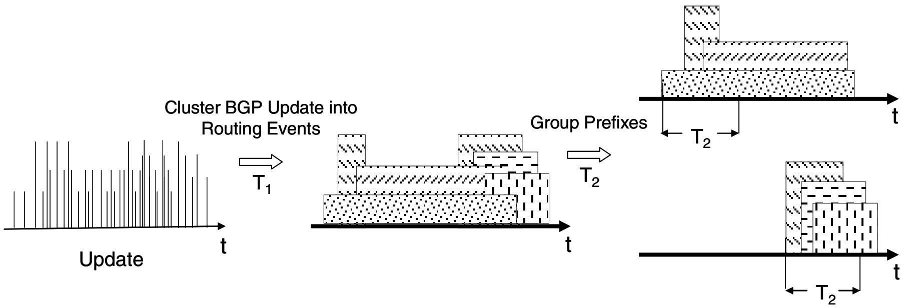

Fig. 3. Two Steps to Get Prefix Groups

Step 1: Cluster BGP Update into Routing Events

The basic idea of this process is to cluster consecutive BGP updates of the same prefix into one routing events if the updates are separated by a time interval less than a threshold. Fig. 4 illustrates this process.

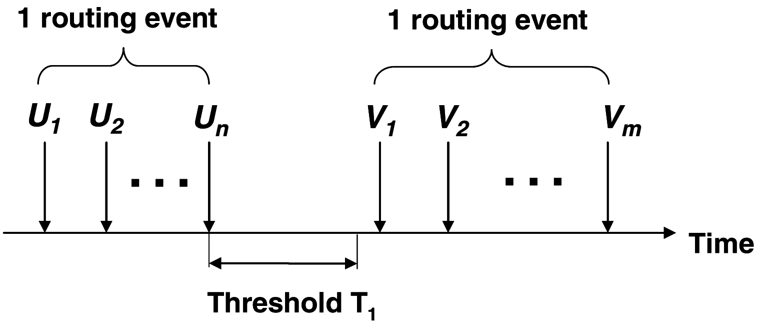

Fig. 4. The process of clustering BGP updates into routing events using threshold T_1. The horizontal axis stands for time. The vertical arrows stand for updates of the same prefix and the same observation point, which are represented by U_i (i = 1 … n) and V_j. (j = 1 … m). The threshold value in this example is T_1. The time intervals between consecutive U_is and V_is are smaller than T_1, while the time interval between U_n and V_1 is bigger than T_1, thus, all U_is constitute a routing events and all V_js constitute another.

The threshold T is crucial in this process. We utilize a dynamic threshold method proposed in [10] to cluster BGP updates into routing events. Comparing with static threshold method, this dynamic approach can get more accurate clustering results.

Step 2: Cluster Prefixes that Always Have Routing Events Simultaneously into One Group

As shown in Fig. 3, after the first step, updates are clustered into different routing events. Some of the events happen almost simultaneously, while others do not. We introduce anther threshold T_2 to indicate the width of the time window. If start times of different routing events locate within one time window, we cluster the prefixes that cause these routing events into the same group.

If prefix P_A and prefix P_B are clustered into one time window for only once, but for most of the time, P_A and P_B have their own routing events independently, we will not say P_A and P_B have relatively small Network-Distance. As a result, we introduce another parameter F in (3) to indicate the frequency of clustering different blocks into one time window. p_i stands for address block. W(p_i) stands for the number of time windows that contain p_i. $W(p_i \cap p_j)$ stands for the number of time windows that contain both p_i and p_j. If F is bigger than a certain threshold f, we say p_i, p_{i+1}…and p_{i+k} belong to the same *Prefix Group*. In our experiment f = 100%. That is to say, only those prefixes which have routing events within one time window all the time can be grouped into one prefix group.

$$F = \frac{W(p_i \cap p_{i+1} \cap ... \cap p_{i+k})}{W(p_i) \cup W(p_{i+1}) \cup ... \cup W(p_{i+k})} \tag{3}$$

4.2 Calculate D_p^g and D_g^g

Suppose k prefixes belong to the same Prefix Group, there will be C_k^2 points generated by prefix pairs in Fig. 2. However, information contained in these points is redundant. We are focusing on the features of a Prefix Group but not any pair in the Group. Only one point, instead of C_k^2 points, is required to characterize a Prefix Group. As a result, we further define D_p^g, D_n^g and D_g^g, where the superscript g stands for the name of the Prefix Group.

Definition 2: D_p^g, D_n^g and D_g^g of Prefix Group g

Prefixes $p_{i+1}...p_{i+k}$ belong to the same Prefix Group g. D_p^g, D_n^g and D_g^g are defined as the arithmetic means of D_P, D_n and D_g, as shown in (4), (5) and (6).

$$D_p^g = \frac{1}{C_k^2} \sum_{p_l \in g, p_m \in g, l \neq m} D_p(p_l, p_m) \tag{4}$$

$$D_n^g = \frac{1}{C_k^2} \sum_{p_l \in g, p_m \in g, l \neq m} D_n(p_l, p_m) \tag{5}$$

$$D_g^g = \frac{1}{C_k^2} \sum_{p_l \in g, p_m \in g, l \neq m} D_g(p_l, p_m) \tag{6}$$

We collect BGP updates from RouteViews [15]. Our dataset contains all updates from route-views.linx in January and February, 2009. The time span should not be too long in case of any significant topology change. We firstly cluster these BGP updates into routing events (we get around 60790000 routing events in total). Then we group prefixes with time window of 5 seconds. After calculating different groups' F values, we choose the groups whose F values are 100% as the Prefix Groups. In this work, we utilize D_g^g as an approximation of D_n^g, and compute Geographic-Distance of prefixes using GeoLite City database from Maxmind [16], which maps each IP address to a geographic location. . Finally, we put all the (D_p^g, D_g^g) points in Fig. 2 to find out the relationship between Prefix-Distance and Geographic-Distance.

5 Measurement Results

5.1 Clustering Result

Fig. 5 shows the result when time window is 5 seconds. There are around 50,000 points and two areas: ZONE 1 near the origin and ZONE 2 far from D_g. This measurement result indicates that there are basically two kinds of Prefix Groups. One is that with small Prefix-Distance and small Geographic-Distance, represented by ZONE 1. The other is that with small Geographic-Distance but significant Prefix-Distance, represented by ZONE 2. Situations in between are very rare.

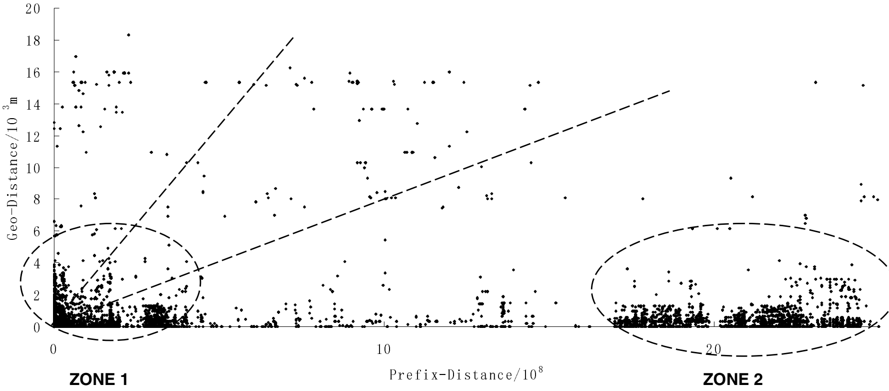

Fig. 5. Prefix-Distance and Geo-Distance when Time Window is 5 Seconds

Prefixes Groups in both ZONE 1 and 2 have relatively small Geographic-Distance. That is to say blocks in one Prefix Group are very likely to get connected to the Internet through the same Point of Presence (PoP). For ZONE 1, blocks in one Prefix Group have small Prefix-Distance, thus they are very likely to get aggregated. On the other hand, for ZONE 2, blocks in one Prefix Group are hardly to get aggregated because of big Prefix-Distance. This measurement result shows an interesting phenomenon: there are two extreme block allocations in current Internet. Some of the blocks with small Geographic-Distances have small Prefix-Distances, while others with small Geographic-Distances have rather big Prefix-Distances.

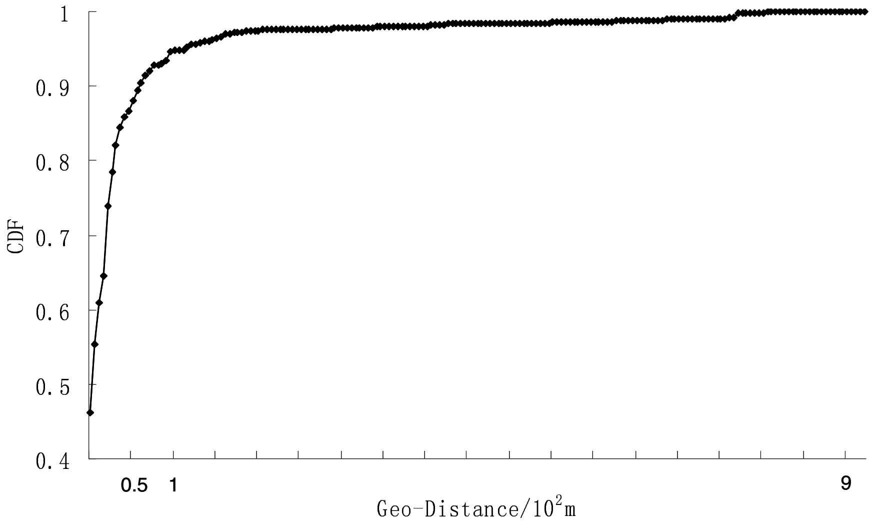

Fig. 6. Geo-Distance Distribution of Prefixes within One Group

We further calculate the Geo-Distance of each prefix group. Fig. 6 shows the Geo-Distance's distribution. More than 90% of the prefix groups' Geo-Distances are less than 100 meters. This indicates that prefix groups generated by our clustering method do have relatively small Geo-Distances. Thus, *Assumption 2* is reliable.

5.2 Prefix-Distance and Geo-Distance in BGP Routing Tables

We also calculate Prefix-Distance and Geo-Distance of current BGP routing table. Firstly, we choose one prefix from route-views.linx BGP routing table. Then, we calculate the Prefix-Distance and Geo-Distance of this prefix with all other prefixes appears in the routing table. In our experiment, we choose two prefixes.

24.143.8/24: This prefix is announced by AS6389, belonging to BellSouth.net Inc. According to CIDR-REPORT [18]'s statistic, AS6389 announces the largest number of prefixes, i.e. around 4322. 24.143.8/24 is one of the prefixes. It first appears in BGP routing table on Oct. 16th, 1999. Fig. 7 shows the Prefix-Distance and Geo-Distance of this prefix and other prefixes in March-4th-2009's BGP routing table download from RouteViews [15] (There are around 291003 prefixes). Even though AS6389 announced the largest number of prefixes, we find out that most of the points locate in Normal D_2. We also draw the Prefix-Distance and Geo-Distance graph of 24.143.8/24 and the other prefixes announced only by AS6389. However, as shown in Fig. 8, almost all the points are located in Abnormal D_3. This indicates the reason why AS6389 announces the largest number of prefixes: because its prefixes are near from each other geographically, but are separated address blocks.

59.252.0.0/16: This prefix is announced by AS37937, belonging to China eGovNet Information Center. According to CIDR-REPORT's statistic, AS37937 only announces this one prefix. It first appears on May 8th, 2007. Fig. 9 shows the Prefix-Distance and Geo-Distance of this prefix and other prefixes in March-4th-2009's BGP routing table download from RouteViews. Most of the points are located in Abnormal D_1, where Prefix-Distance is small but Geo-Distance is relatively large. This indicates

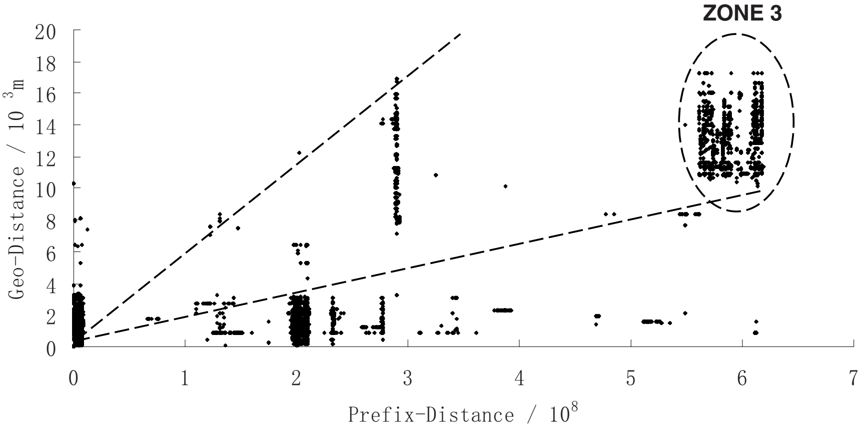

Fig. 7. Prefix-Distance and Geo-Distance of 24.143.8/24 and other Prefixes

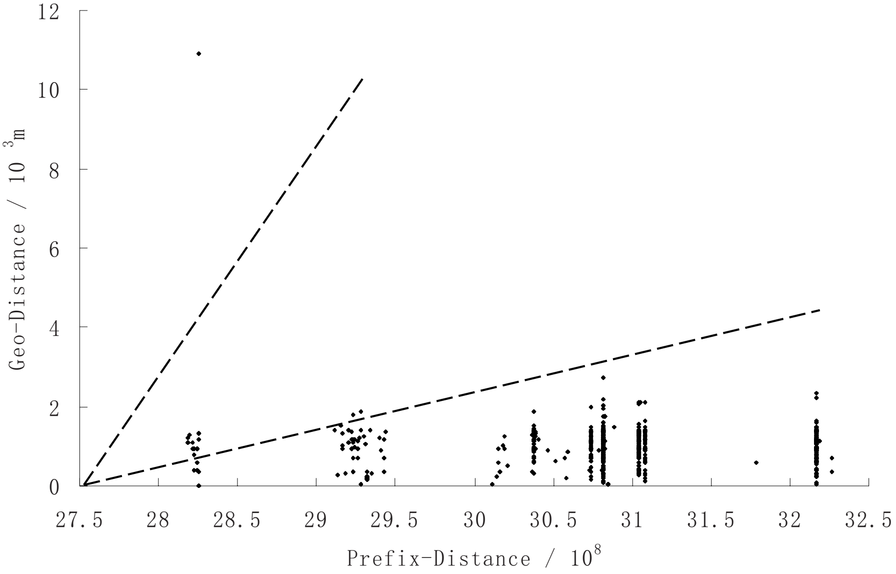

Fig. 8. Prefix-Distance and Geo-Distance of 24.143.8/24 and other Prefixes in AS6389

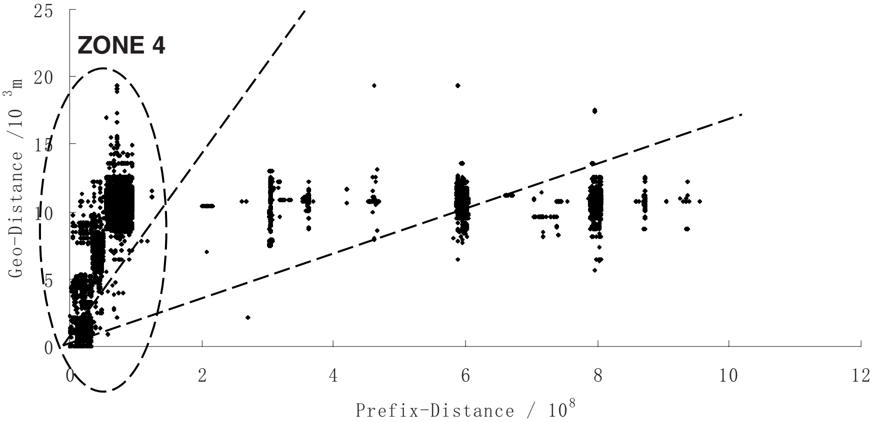

Fig. 9. Prefix-Distance and Geo-Distance of 59.252.0.0/16 and other Prefixes

that aggregatable prefixes are far away from each other, and thus cannot get aggregated. Comparing with prefix 24.143.8/24 (whose measurement result shows in Fig. 7), 59.252.0.0/16 first appears in BGP routing table eight years later. During these eight years BGP table size grows from 6000 entries to 290000 entries. According to Xiaoqiao Meng [7]'s conclusion that 45% of the address allocations from 1997 to 2004 were split into fragments smaller than the original allocated blocks, we speculate that failing to allocating IP address blocks according to geographic information is the most possible reason. This speculation is identical with M. Freedman [6]'s conclusion.

6 Conclusion and Future Work

In this paper, we try to figure out the relationship between Prefix-Distance and Network-Distance of current Internet by taking Geographic-Distance as an approximation of Network-Distance. We focus our measurement on the prefixes with relatively small Geographic-Distance, and get Prefix Groups from BGP routing dynamics. Comparing with previous active probing methods, our method 1) reduces the number of active probes required in active measurement. And 2) results a more detailed prefixes' distribution analysis. Our measurement result shows that 1) there are two extreme allocations of IP address blocks in current Internet. Some of the blocks with small Geographic-Distances have small Prefix-Distances, while others with small Geographic-Distances have rather big Prefix-Distances. 2) Failing to allocating IP address blocks according to geographic information in the past eight years is one of the driving forces of BGP tables' inflation.

There are two directions for our future work. Currently, the reliability of our measurement lies on the correctness of MaxMind [16]. Firstly, we will try to fit some tools that can estimate Network-Distance into our system. Secondly, we are trying to do some quantitative analysis about our measurement results.

Acknowledgments. We would like to thank Professor David Lee from the Ohio State University for his great suggestions on our problem statement, measurement and analysis. This work is supported by the National Basic Research Program of China (973 Program) (Grant No. 2009CB320502).

References

1. Rekhter, Y., Li, T.: A Border Gateway Protocol 4 (BGP-4). IETF, Request for Comments 4271 (January 2006)
2. Meyer, D., Zhang, L., Fall, K.: Report from the IAB Workshop on Routing and Addressing. In: Routing and Addressing Workshop, Amsterdam, Netherlands, December 15 (2006)
3. Bu, T., Gao, L., Towsley, D.: On Routing Table Growth. In: ICNP (2003)
4. Andersen, D., Feamster, N., Bauer, S., Balakrishnan, H.: Topology Inference from BGP Routing Dynamics. In: SIGCOMM IMW 2002 (2002)
5. Bu, T., Gao, L., Towsley, D.: On Characterizing BGP Routing Table Growth. Computer Networks: The International Journal of Computer and Telecommunications Networking 45, 45–54 (2004)
6. Freedman, M., Vutukuru, M., Feamster, N., Balakrishnan, H.: Geographic Locality of IP Prefixes. In: IMC 2005 (2005)
7. Meng, X., Xu, Z., Zhang, B., Huston, G., Lu, S., Zhang, L.: IPv4 address allocation and the BGP routing table evolution. Computer Communication Review 35(1), 71–80 (2005)
8. Oliveira, R., Lad, M., Zhang, B., Zhang, L.: Geographically Informed Inter-Domain Routing. In: ICNP 2007, Beijing, China (2007)
9. Chang, H., Jamin, S., Willinger, W.: Inferring AS-level Internet topology from router-level path traces. In: Proc. of SPIE ITCom, August 2001, pp. 19–24 (2001)

10. Wu, X., Yin, X., Wang, Z., Tang, M.: A Three-step Dynamic Threshold Method to Cluster BGP Updates into Routing Events. In: ISADS 2009 (2009)
11. Francis, P., Jamin, S., Jin, C., Jin, Y., Raz, D., Shavitt, Y., Zhang, L.: IDMaps: An architecture for a global internet host distance estimation service. In: Proc. IEEE INFOCOM, March 1999, vol. 1, pp. 210–217 (1999)
12. Ng, E., Zhang, H.: Predicting Internet network distance with coordinates-based approaches. In: INFOCOM (2002)
13. Shavitt, Y., Tankel, T.: On the curvature of the Internet and its usage for overlay construction and distance estimation. In: INFOCOM (2004)
14. Madhyastha, H.V., Anderson, T., Krishnamurthy, A., Spring, N., Venkataramani, A.: A Structural Approach to Latency Prediction. In: IMC 2006 (October 2006)
15. The RouteViews project, `http://www.routeviews.org/`
16. MaxMind GeoLite City, `http://www.maxmind.com/app/geolitecity`