Improving IP Geolocation Using Network Measurement Characteristics

Iaroslav Iaremko, Lukas Aron, Petr Hanacek
Faculty of Information Technology BUT, Bozetechova 1/2, Brno, Czech republic
E-mail address:{iiaremko, iaron, hanacek}@fit.vutbr.cz

Abstract

This paper describes the basic principles of evaluation of the physical location of the network focusing on passive IP geolocation methods. Creating a software application enables us to determine the position of the stations on the Internet using selected localization databases. The algorithm for locating the position of the station uses the combination of location databases and other available means of IP geolocation. The paper describes basic principles of TCP connection. It explains the activities of TCP protocol and shows how you can utilize TCP handshake in determining the distance between nodes. The distance between two or more computers is shown by the analysis of TCP, by monitoring TCP packets to calculate the TCP Round Trip Time and comparing the results from methods such as IATA, TTL and GEOIP for calculating distance measurement. An assessment of the created system and performance tests is enclosed as part of this paper.

Keyword: TCP, RTT, TTL, IATA, GeoIP, handshake, network node distance

1. Introduction

The number of internet users is growing every year, as well as the number of network devices owned by users and the number of applications that every modern user needs. Thanks to the internet, users from one part of the world for example from Sweden are able to connect with other users from different parts of the world i.e. from Australia using social networks like Facebook, Skype, Viber, etc.. There are questions related to the distance between users and their location. This location is usually connected with information about the local weather, local time and the services nearby. The position of an internet user is called geolocation - its related to the position on Earth. Geolocation is being increasingly used as a method for determining the position in many applications and services. Geolocation methods can be divided into two groups: active and passive. Passive methods do not perform any measurements, but are able, with the use of certain characteristics of the station, to determine its location in the different geolocation databases. Active IP geolocation techniques, typically based on delay measurements, may achieve desirable properties such as accuracy (i.e., active measurements provide better results compared to geolocation database in many cases). However, these properties come at the expense of lack of scalability, high measurement overhead, and very high response time ranging from tens of seconds to several minutes to localize a single IP address. This is several orders of magnitude slower than what is achievable with the passive approach, i.e., database-driven geolocation. And also the active geolocation methods are based on measurement of network path properties between target station and station with known position (landmark). Most popular proposed method based on this principle was GeoPing , where firstly the each active landmark measures round trip time (RTT) to a few passive landmarks and perform a delay vector. Target station also measure the

round trip time to the same passive landmarks. The geographic position of active landmark with the most similar delay vector to target delay vector is announced as position of target.
Constraint Based Geolocation (CBG) uses round trip time measurement between landmarks and target. The main principle of CBG is the usage of trilateration, where geographic distances from landmarks bound the area of possible location, the intersection of these boundaries is the area of estimated target position. The problem of this and similar methods is in round trip time mapping into geographic distance. CBG goes through the calibration phase before measurement. The calibration is the measurement of round trip times between all landmarks. Each landmark, based on calibration, calculates owned by Bestline, where slope m and intercept b are used for geographic distance g calculation by equation.

$$g = \frac{d-b}{m}$$

Eq 1: The measurement of round trip times between all landmarks

This principle is described by equation - see Eq 1. Where the *d* is measured round trip time (delay). Most of other geolocation methods are based on principle of CBG, however, they use different techniques how to calculate geographic distance. Speed of Internet (SOI) simplifies the geographic distance calculation. Method Octant brings in negative distances, which bound the area, where target cannot be. As in CBG the calibration is done firstly and then is measured pairs of values bounded with convex hull . This boundary is used for delay to geographic distance mapping [10, 12].
Many of the existing geolocation techniques use Whois databases, DNS LOC records or DNS names to determine the location of a given host. From Whois databases one can retrieve the name and street address of the organization which registered the address block. However, for a large ISP or a geographically dispersed organization the registered street address usually differs from the real location of its hosts. A similar problem arises in the use of DNS names, since the names can be both useful or misleading due to the naming conventions of the ISP. Other registry based techniques include commercial approaches, e.g. the gathering of user submitted location data from commercial websites, or network reconnaissance, where one obtains the description of the geographic layout of an ISP's network and internal routing policies.Another technique is where the topology information and latency measurements are used together in the location estimation. This method type is called topology based geolocation (TBG). TBG localizes all the intermediate routers between the landmarks and the target node. This approach is based on link-latency estimations and on precise topology discovery. The basic tools of this method are traceroute and interface clustering applications. Later works such as Octant enhance the accuracy of location approximations by combining TBG with various other techniques, including DNS and Whois lookups or clipping regions with negative geographic and demographic constraints. Europe using MaxMind database. MaxMind DB located all of the routers in Cambridge, UK, where the GÉANT operator is registered. This mislocation indicates that MaxMind relies on previously registered Whois data, which leads to unreliable results in this case.The Internet Assigned Numbers Authority (IANA) is a department of ICANN responsible for coordinating some of the key elements that keep the Internet running smoothly [8, 15].
Whilst the Internet is renowned for being a worldwide network free from central coordination, there is a technical need for some key parts of the Internet to be globally coordinated, and this coordination role is undertaken by IANA.Specifically, IANA allocates and maintains unique codes and numbering systems that are used in the technical standards (sometimes called protocols) that drive the Internet. IANA's various activities can be broadly

grouped into three categories:

- Domain Names - IANA manages the DNS Root, the .int and .arpa domains, and an IDN practices resource.
- Number Resources - IANA coordinates the global pool of IP and AS numbers, providing them to Regional Internet Registries.
- Protocol Assignments - Internet protocols' numbering systems are managed by IANA in conjunction with standards bodies [3].

2. Positioning though the internet network

2.1 Positioning using TCP RTT methods techniques and performance tests

The TCP flow from a server to client, called RTT, is estimated during the three-way handshake. Current trace starts with one SYN packet, followed by a SYN-ACK from the client to the server. Important note, that the trace does not include the SYN-ACK packet from the client to the server, because that packet is sent in the reverse direction flow. The basic idea is that the RTT can be estimated from the time interval between the last SYN and the first ACK packets that the client sends to the server. This principle is shown on Figure 1.

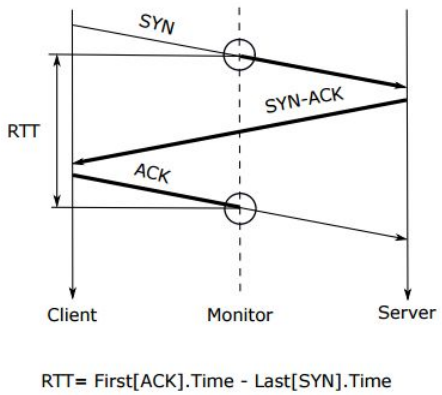

Figure 1: Distance measurement using RTT method

Examples of using the methods to evaluate the TCP Round Trip Time in the detection of the distance of the computers that are situated in different parts of the world are listed in the Tables 1 and 2. These tables show the basic examples that are possible to get by measuring the distance of various servers in different parts of the world. These results can help determine the accuracy or inaccuracy of the estimations of other geolocation methods [19].

Server	www.france.com	-
IPsrc/IPdst	147.251.209.52	96.126.117.104
RTT [ms]	131	-
TTL	64	-
srcCtry/dstCtry	czech republic	united states

Table 1: Estimation of the location of the web server www.france.com using the RTT method

Server	www.google.com	-
IPsrc/IPdst	147.251.209.173	64.233.166.147
RTT [ms]	16	-
TTL	64	-
srcCtry/dstCtry	czech repiblic	united states

Table 2: Estimation of the location of the web server *www.google.com* using the RTT method

These results show that the RTT value of the current IP address, which belongs to the Czech Republic and the IP address belonging to the server *www.france.com* is relatively large. After this investigation it is possible to determine that the server *www.france.com* it is not nearby. According to the GeoIP2 database this server situated in USA. The second example - the address of the server *www.google.com*. The GeoIP2 database determines that the IP address of the server belong to the US. In this situation our RTT value is small compared for the RTT value, which belonged to the server www.france.com. In this case, it is possible to determine that the server *www.google.com* is not situated in USA. We are probably communicating with the Google company located in Europe. Round Trip Time method is one of the most accurate of the all other methods, which are described in other articles Because RTT method can to provide a great accuracy to determine the probable location of a target node.

2.2 Positioning using the TTL method

Time to live (**TTL**) or **hop limit** is a mechanism that limits the lifetime of data in a computer or network. TTL may be implemented as a counter or timestamp attached to or embedded in the data. The basic principle of this method states that the packets are directed from the node A directly to the node B through several routers. This distance is measured by the number of hops from node A directly to the node B. Note, when we measure the TTL between the router IP which is *125.85.138.85* and node B in the direction from node A, in this case, the measurement of TTL values will be done correctly and the method correctly determines the number of hops from node A in the direction to the node B. Contrarily , when we perform a measurement of TTL values between the router, IP of which is *125.85.138.85* and node B in the direction from node B, the measurement of the TTL doesn't make any sense. In this case, the TTL value does not show, how much hops or nodes on the path are from the node B towards node A. In all cases, it is necessary to realize, in what direction to direct the packet. TTL method can be used as a supplement for the calculation of the distance together with other methods for getting more information about the location of the IP address, as well as roughly showing a topology of the network [9].

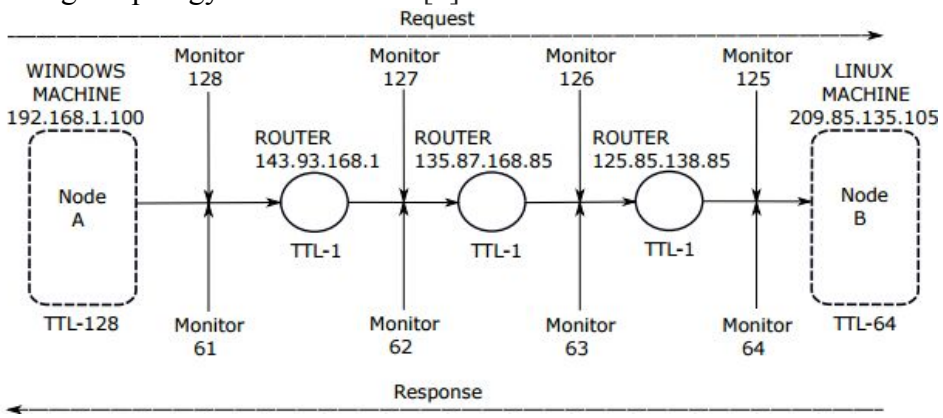

Figure 2: Distance measurement using TTL method

Server	www.ebay.com	-
IPsrc	147.251.209.173	66.211.160.87
IPdst	66.211.160.87	147.251.209.173
RTT [ms]	174	-
TTL	64	21
srcCtry	czech republic	united states
dstCtry	united states	czech republic

Table 3: Examination of the location of the web server *www.ebay.com* using TTL method

2.3 Positioning using IATA method

The basic principle of the IATA method for determining a positioning is shown in Figure 2. According to the IP address, a DNS query is entered in order to find the reverse DNS of the record, which in itself contains the IATA code which is used to determine the location of the server with the use of the IATA database. The following examples show the pros and cons of using this method to determine the location of the particular station [18].

Server	www.facebook.com	-
IPsrc	31.13.93.3	147.251.209.173
IPdst	147.251.209.173	31.13.93.3
RTT [ms]	-	19
TTL	43	21
srcCtry	ireland	czech republic
dstCtry	czech republic	ireland
IATA	ams	amsterdam

Table 4: Examination of the location of the server *www.facebook.com* using IATA method

The result of the measurement, which is shown in Table 4 demonstrates the positioning of the server *www.facebook.com*. In the result we have the IATA code of the reverse DNS record *edge-star6-shv-01-ams3.facebook.com*. From the ams code, which is found in the database of IATA-alpha codes, we were able to get the name of state. For additional information, such as getting the location of the network station directly from the domain name, we can also use the IATA method. The problem lies in the fact that not every domain name has its subdomain and not every subdomain has its IATA abbreviation. In some cases it is preferable to identify the location of the network station using the IATA code, which is encapsulated in a domain name than using the GeoIP database.

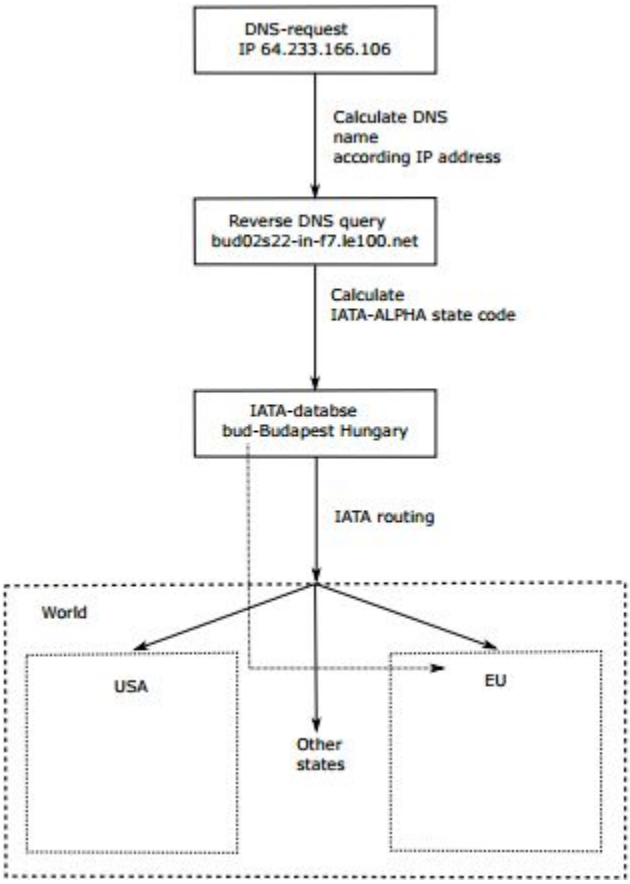

Figure 3: The measurement of distance using IATA method

2.4 Positioning using GeoLibc method

The basic principle of the GeoLibc method lies in the fact that we have a specific database that contains a set of IP addresses and corresponding countries or states, which includes the IP address and other values with which it is possible to determine the location in the network node with great precision. The fundamental problem lies in the fact that a lot of servers have their data service providers in other states and use the IP address of their service provider e.g. a server that has their provider in Ireland belongs to the United States, GeoIP database determines that the server and provider belong to the same state, but that is not true. This case illustrates our method using RTT estimation [14].

3. The algorithm for the implementation of the Geolocation functionality

```
pointer geoipv4 ← GeoIP_new(type og opening)
pointer geoipv6 ← GeoIP_open_type(the IPv6 database, opening type)
for all IP_adresa do
  if IP_adresa is IPv4 then
    DNS_NAME ← GeoIP_id_by_ipnum(geoipv4, IP_adresa)
    ctry_id ← GeoIP_id_by_ipnum(geoipv4, IP_adresa)
  else
    if IP_adresa is IPv6 then
      DNS_NAME ← GeoIP_id_by_ipnum_v6(geoipv6, IP_adresa)
      ctry_id ← GeoIP_id_by_ipnum_v6(geoipv6, IP_adresa)
    end if
  end if
  save iso3166_GeoIP_country_codes[ctry_name]alpha_2_code
  Calculate_RTT_among_network_hosts
  Calculate_TTL_among_network_hosts
  Calculate_IATA_code_for_DNS_name
end for
GeoIP_delete(geoipv4)
GeoIP_delete(geoipv6)
```

Explanation of the algorithm:

Step1: Get the IP address;
Step2: Analysis the IP address (for exp. is that current IP address IPv4 or IPv6 type);
Step3: Get DNS Name from current IP address;
Step4: Calculate RTT, TTL, IATA code from DNS and getting the current Geolocation according IP address and IATA code and compare this results;
Step5: Delete IP address from netflow cache;

4. Conclusion

On the diagrams we can see that the most accurate for identifying the location is the RTT method, less accurate is Geolibc, other methods do not give a precise result to locate some positions of the network node. For more accurate calculation of the distance it is possible to take advantage of the RTT and Geolibc methods. The TTL and IATA methods can be used as additional methods. RTT estimation shows the distance to a node, as well as the time that is required for the connection with another station. Geolibc using GeoIP2 database can be used as an auxiliary part of RTT method.

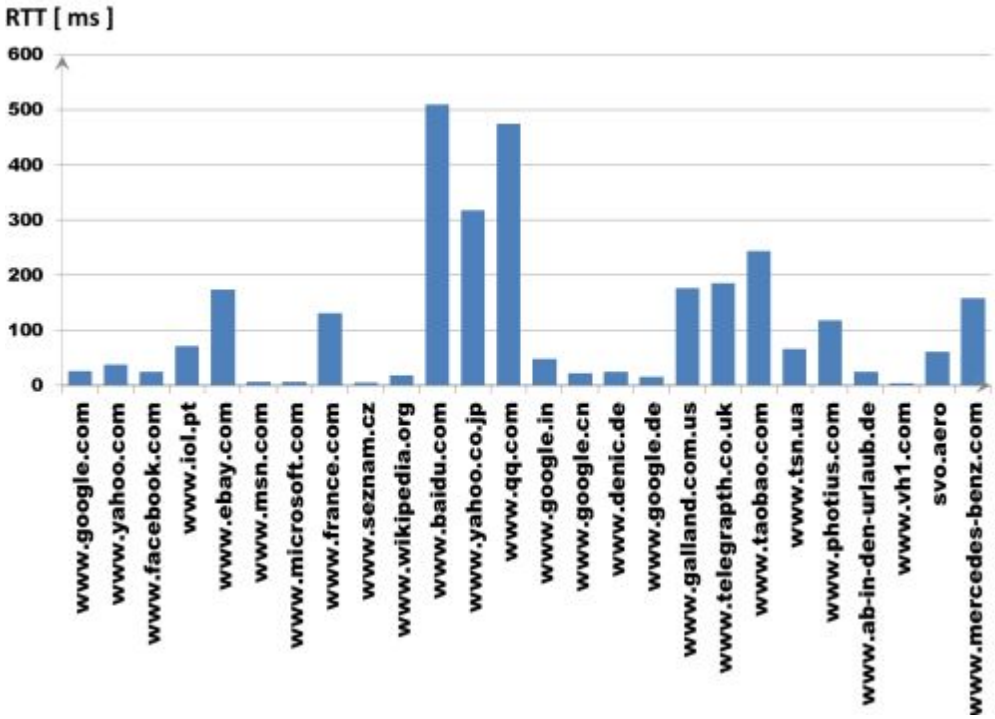

Figure 4: Results of the measurement methods RTT, TTL and Geolibc

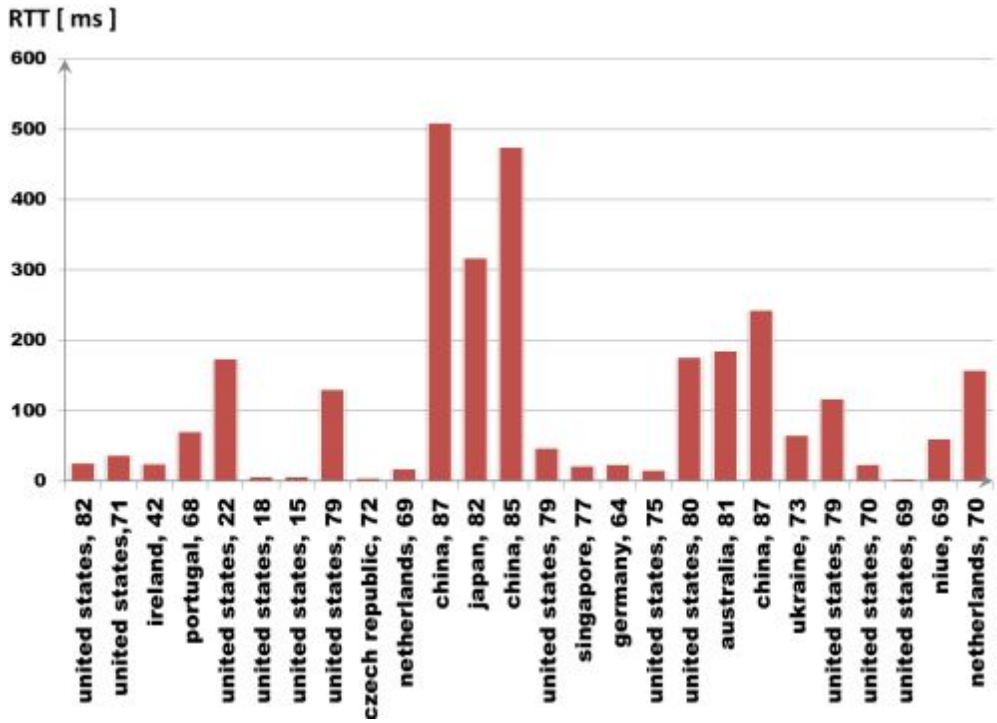

Figure 5: Results of the measurement methods RTT, TTL and Geolibc

5. Acknowledgments

This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070), by the project CEZ MSM0021630528 Security-Oriented Research in Information Technology and by project FIT-S-11-1 Advanced Secured, Reliable and Adaptive IT

6. References

1. B. Gueye, A. Ziviani, M. Crovella, and S. Fdida (2006). Constraint-based geolocation of Internet hosts. IEEE/ACM Transactions on Networking. vol. 14, no. 6, (pp. 1219–1232).
2. Bamba Gueye, Steve Uhlig and Serge Fdida (2007). Investigating the Imprecision of IP Block-Based Geolocation, PAM 2007, Louvain-la-neuve, Belgium. LNCS 4427(pp.237-240)
3. IANA - The Internet Assigned Numbers Authority.
4. P. Mátray, I. Csabai, P. Hága, J. Stéger, L. Dobos, G. Vattay (2007). Building a Prototype for Network Measurement Virtual Observatory. ACM SIGMETRICS - MineNet 2007, 12 June, San Diego, CA, USA.
5. S. Laki, P. Mátray, P. Hága, I. Csabai, G. Vattay (2010). A Model Based Approach for Improving Router Geolocation. Computer Networks, Volume 54, Issue 9 , 17 June 2010, (pp. 1490-1501).
6. B. Wong, I. Stoyanov, E.G. Sirer Octant (2007). A Comprehensive Framework for the Geolocalization of Internet Hosts. NSDI 2007 Symposium, Cambridge, Massachusetts, April 2007.
7. KOMOSNY, Dan a Lukáš VERNER. Geolocation of network devices in internet networks. 2011 -Elektrorevue. (2011).
8. Y. Shavitt, Y. Singer (2010). Limitations and Possibilities of Path Trading Between Autonomous Systems. IEEE INFOCOM. March 2010, San Diego, CA, USA.
9. Venkata N. Padmanabhan, Lakshminarayanan Subramanian (2001). An investigation of geographic mapping techniques for internet hosts. Proceedings of ACM SIGCOMM (pp.173-185). August 2001, San Diego, CA, USA.
10. M. Zhang, Y. Ruan, V. Pai, and J. Rexford (2006). How DNS misnaming distorts Internet topology mapping. USENIX Conference.
11. ERIKSSON, B., BARFORD, P., SOMMERS, J., AND NOWAK, R (2010). A learning based approach for IP geolocation. In Proceedings of the Passive and Active Measurement Workshop (April 2010).
12. GUEYE, Bamba, Artur ZIVIANI, Mark CROVELLA a Serge FDIDA (2010). Constraint-Based Geolocation of Internet Hosts. IEEE/ACM Transactions on Networking.
13. GUEYE, B., ZIVIANI, A, CROVELLA, M., AND FDIDA, S. (2006). Constraint-based geolocation of Internet hosts. IEEE/ACM Transactions on Networking.
14. IP geolocation using delay and topology measurements. In Proceedings of the ACM SIGCOMM Internet Measurement Conference (October 2006).
15. GILL, P., ARLITT, M., LI, Z., AND MAHANTI (2008). A the flattening Internet topology. Natural evolution, unsightly barnacles or contrived collapse. In Proceedings of the Passive and Active Measurement Workshop (April 2008).
16. N. Spring, D. Wetherall, and T. Anderson (2003). A public Internet measurement facility in Proc. USENIX USITS Symposium. March 2003
17. S. Siwpersad, B. Gueye, and S. Uhlig (2008). Assessing the geographic resolution of exhaustive tabulation for geolocating Internet hosts, in Proc. PAM Conference. April 2008.
18. B. Gueye, S. Uhlig, and S. Fdida (2007). Investigating the imprecision of IP block-based geolocation, in Proc. PAM Conference. April 2007.
19. Hao Jiang, Constantinos Dovrolis (2002). Passive Estimation of TCP RoundTrip Times. ACM Computer Communication Review.
20. B. Wong, I. Stoyanov, and E. G. Sirer (2005). Golocalization on the Internet through constraint satisfaction, in Proc. USENIX WORLDS Workshop, November 2005.