International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:7, No:8, 2013

Advanced Geolocation of IP Addresses

Robert Koch, Mario Golling and Gabi Dreo Rodosek

Abstract—Tracing and locating the geographical location of users (Geolocation) is used extensively in today's Internet. Whenever we, e.g., request a page from google we are - unless there was a specific configuration made - automatically forwarded to the page with the relevant language and amongst others, dependent on our location identified, specific commercials are presented.

Especially within the area of network security, Geolocation has a significant impact. Because of the way the Internet works, attacks can be executed from almost everywhere. Therefore, for an attribution, knowledge of the origination of an attack - and thus Geolocation - is mandatory in order to be able to trace back an attacker. In addition, Geolocation can also be used very successfully to increase the security of a network during operation (i.e. before an intrusion actually has taken place). Similar to greylisting in emails, Geolocation allows to (i) correlate attacks detected with new connections and (ii) as a consequence to classify traffic a priori as more suspicious (thus particularly allowing to inspect this traffic in more detail).

Although numerous techniques for Geolocation exist, each strategy is subject to certain restrictions. Following the ideas of Endo et al., this publication tries to overcome these shortcomings with a combined solution of different methods to allow improved and optimized Geolocation. Thus, we present our architecture for improved Geolocation, by designing a new algorithm, which combines several Geolocation techniques to increase the accuracy.

Keywords—IP geolocation, prosecution of computer fraud, attack attribution, target-analysis

I. Introduction

Today, information and communication technology (ICT) is of key importance in almost any economic sector: Health, mobility, education, entertainment, production, logistics, trade, finance or supply (e.g., energy, water) as well as public administration. The degree of dependence of modern industrialized countries on ICT - in the public as well as the private sector - has reached a dimension, which seemed unimaginable a few years ago. The Internet has become the "steam-engine" of the 21st century. It drives the economy and opens the door to new innovative business models. Consequently, the Internet has become one part of what is called the critical infrastructure [1]. A breakdown of the Internet would lead to a shortage of supplies, significant disruptions of public order or other dramatic consequences. As a recent study shows, 25 percent of all German companies would be bankrupt in case of a complete breakdown of their network infrastructure for only two or three days [2]. Because of the way the Internet works, attacks can be executed from almost everywhere. Therefore, for an attribution, knowledge of the origination of an attack is mandatory in order to be able to trace back an attacker.

Firstly, tracing and locating the geographical location (also called Geolocation) is a necessary precondition for the legal prosecution of law enforcement agencies (although sometimes different laws in different countries with specific extradition agreements may make the process difficult). As a recent study from the security company Mandiant [3] - claiming to analyze Chinas Cyber Espionage Units - proclaimed, "a large share of hacking activity targeting the US could be traced to an office building in Shanghai". Although the Chinese government has denied the accusations [4], the political pressure on China from the US continues. In return, it also seems that the US government has been hacking Hong Kong and China for years [5]. Both examples show how important an attribution in cyber space is and thus the rising importance of Geolocation to support attribution.

Secondly, Geolocation is also a necessary condition for identifying and examining the network structure of the opponent in order to (i) counterattack (for example in a Cyber Conflict) and to (ii) finally bring down the attack. Although numerous techniques can be used to scramble the real IP address of an attacker (e.g., NAT, proxies, anonymizing networks like TOR or the use of Bots, which are under control of the attacker), here, tracing and locating the geographical position can also support subsequent activities like isolating a system.

Thirdly, Geolocation may also be used *before* an Intrusion was successful. Based on attacks detected (e.g., by Intrusion Detection Systems such as Snort [6]), a correlation of these attacks with new connections is possible as well. Thus as a consequence, new connections originating from a location very close to where a recent attack was launched may be inspected in more detail in comparison to normal network traffic. Similar to greylisting in emails, this correlation of attacks with new connections may be performed to classify traffic a priori as more suspicious (thus particularly allowing to inspect this traffic in more detail, like for instance performing a deep packet inspection on this traffic while the regular traffic is only inspected flow-based (information derived from the packet headers [7], [8]). Other examples where Geolocation is used with regard to network security are for instance:

- **Online Banking Security:** E.g., PayPal uses Geolocation to protect against fraud, monitoring online payments for regional discrepancies excluding transactions that appear to come from countries imposed by international sanctions (according to the list of the Office of Foreign Assets Control).
- **Email Security:** DigitalEnvoy checks the emails for geographical plausibility. They compare the geography of the email header with the geography of the email body. Suspicious emails are possibly blocked or passed to routines to protect against phishing.

Besides network security, Geolocation can also be used for many other purposes, like:

All authors are members of the Research Center CODE (Cyber Defence), Faculty of Computer Science, Universität der Bundeswehr München, D-85577 Neubiberg, Germany
email: {robert.koch, mario.golling, gabi dreo}@unibw.de

- **Language/Currency Services:** Google and many other personalize their offers leading users automatically to the page of their language resp. their currency.
- **Advertisement:** Most online advertising firms now offer their customers nationally or even regionally differentiated advertising (*Ad Targeting*).
- **Content Delivery Networks** optimize the load balancing between their servers and provide better traffic management for downloads based on information gained through Geolocation.
- **Video-on-Demand**-Provider like CinemaNow or Disney are using Geolocation because sports associations and film publishers bind their content to territorial boundaries. Also YouTube has blocked some videos due to licensing issues in some countries.

The aim of this publication is to design a method for advanced Geolocation of IP addresses by taking into account the accuracy required and the potential detectability. Following the idea of Endo et al. [9], this publications tries to overcome shortcomings of existing approaches with a combined solution of different methods.

The paper is structured as follows: A short overview of state-of-the-art Geolocation techniques and tools is presented in Section II. Thereafter, our architecture is illustrated (Section III) as well as the corresponding Proof of Concept (Section IV), before an evaluation is performed in Section V. Finally, Section VI contains a conclusion and outlook.

II. Related Work

Over the past decades, several methods for Geolocation have been developed. According to Dahnert et al. [10] and Laki et al. [11], the different approaches can be classified into two categories. While the first is based on the semantic interpretation of prestored database-records, the second uses active latency and topology measurements. Since many of the approaches have active and passive components (*hybrid*), for the sake of clarity, this chapter is not divided into the two classifications.

Due to the brevity of this publication, only some approaches can be described in more detail. Thus, the selection of the methods presented is based on the criteria of publicity and reuse of certain aspects within our Proof of Concept.

A. Overview of Related Work

1) IP2Geo: is one of the first approaches for the allocation of a logical IP address to a physical location based on measurement. In "An investigation of geographic mapping techniques for internet host" [12] Padmanabhan et al. presented three different algorithms called GeoPing, GeoCluster and GeoTrack. Each of the three algorithms is hereby based on different methods to determine the location of the host.

GeoPing is a method that utilizes the correlation between latency values (such as Round Trip Time; RTT), and a geographical distance [12]. The existence of this relationship is a fundamental part of the GeoPing algorithm and at the same time represents a major challenge. In contrary to conventional opinions, that such a correlation does not exist, Ziviani et al. confirm their very existence [13]. The conclusion to the geographical position of a host is done using so-called landmarks (entities with known location). For this purpose, the minimum RTT of the client to the landmarks is measured and the results are then transferred to a map (see Figure 1). The granularity of the results depends largely on the amount and location of usable landmarks [13]. Also distortion caused by, for example, routing loops, the last mile and safety aspects represent fundamental problems in locating an IP address. The accuracy of the results provided in GeoPing is limited to a discrete solution space, which in this context means a concrete landmark and not a region.

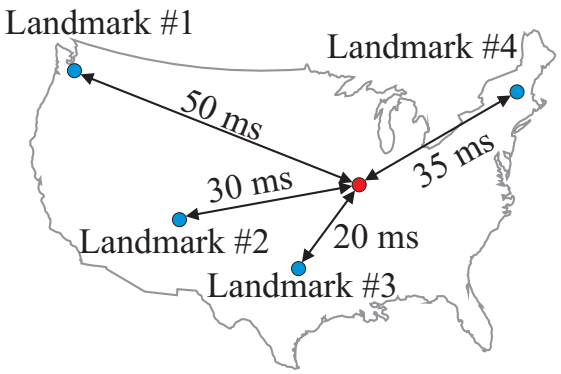

Fig. 1. Functionality of GeoPing

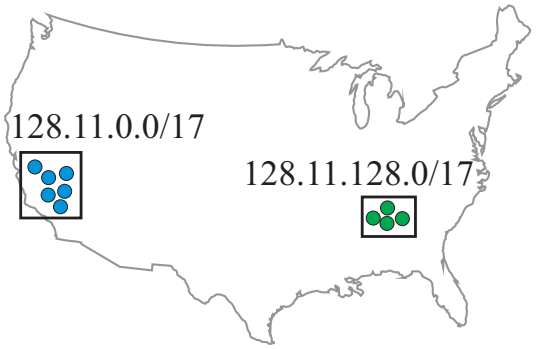

Fig. 2. Affiliation of IP addresses to clusters in GeoCluster

GeoCluster divides the entire IP address space into blocks or clusters. The basic assumption is that all IP addresses of a cluster can be found in the same region (see Figure 2). Thus, based on the allocation of a cluster to a geographic region, the actual location of the destination system is suggested. Consequently, in order to assign a logical address to a cluster, extensive information on the general distribution of the IP portfolio is required. This information is obtained by the evaluation of Border Gateway Protocol (BGP) routing tables/BGP address prefixes, whois databases and information gathered from other sources such as Internet Service Providers or registry data from Service Providers. Due to the fact that the records of the databases are usually not checked intensely for correctness, a deliberate falsification is possible. The same applies to the

whois protocol. A mapping of a single IP to a precise location is also not easily possible, because usually only the address of the headquarters of the owner is deposited. This in turn won't bring a benefit if the corresponding autonomous system is geographically widely distributed.

GeoTrack tries to infer the geographic location of a host by examining its Full Qualified Domain Name (FQDN) for geographical indications. This is due to the fact that many network operators provide references to the location of the individual nodes within the FQDN, for the purpose of simplified network administration [12]. Furthermore, GeoTrack uses so-called tracerouting to determine all intermediate stations (routers) including the corresponding FQDN on the way to the host. With the use of regular expressions and pattern matching each hostname (target host and all routers on the way) is examined for geographical indications. In addition to city and country names, the underlying databases also contain airport codes, as they are in accordance with an empirical study mostly used for labeling [12]. GeoTrack faces the problem of DNS misnaming according to Zhang et al. [14]. Another problem is, that the use of geographical pattern as part of the hostname is not a standard, and thus extent and type of application depends on the provider or network operator. Serious in terms of accuracy is the effect of the use of traceroute, since it relies on UDP or ICMP, which is dropped frequently on routers.

2) Constraint Based Geolocation (CBG): was developed to deal with the problems of a discrete solution space for the localization, using landmarks (see GeoPing) [9]. In "Constraint-based geolocation of internet hosts" [15] Gueye et al. provide an approach based on multilateration (see Figure 3), where the position of a host is also determined based on the distance to known landmarks. Here, a continuous solution space is achieved by using two values: A minimum and a maximum distance. Based on latency measurements of signals in fiber optic cables as well as the assumption that up to the last mile (respectively satellite links) almost all lines are made of fiber, the theoretical minimum distance is assumed to be $min = \frac{2}{3}c$; where c is the speed of light [16] (represented in Figure 3 by complete circles). The maximum distance is represented by the maximum speed of signals in fiber optic cables, which is $max = c$ (represented in Figure 3 with the use of dashed circles). The intersection over all discovered circular functions (minimum and maximum distance) is used to determine a geographic region whose center is assumed to be the exact position [15]. CBG deliberately makes an overestimation of the upper limit to ensure that the solution space is not empty. This, however, at the same time increases the intersection and thus the potential target area. Accuracy is influenced by the number of available landmarks [13] and their positions. A fundamental problem in this case are firewalls, proxies and Intrusion Detection Systems. Since CBG exclusively uses Ping-based methods to determine the delay, it can be assumed that many measurements are faulty.

3) Topology Based Geolocation (TBG): is an evolved variant of CBG, also taking topological aspects into account and thus increasing the accuracy significantly. TBG is only an extended version of the CBG-Algorithm and thus raises the same problems. In addition, the reduction of errors is done at the expense of performance.

4) Octant: is a modular framework for Geolocation, which uses a variety of geometric curves, known as Bézier curves, to determine the physical location of a target system, as well as positive and negative conditions [9], [17] (see Figure 4). The framework, developed by Wong et al. [17], was built on the results of TBG and extends this approach by using network nodes of the path towards the client as additional landmarks. The modular design enables Octant to formulate additional constraints that can limit a possible geographic region significantly. These constraints are based on collected demographic data, for example, and may limit the location of a possible site to inhabited areas. Other possibilities are the introduction of information from Regional Internet Registries (RIRs) and the use of Geoservice providers such as MaxMind or Quova. Nevertheless, also Octant has the same problems TBG or CBG has.

Table I provides a brief overview of methods and approaches for Geolocation. The second column indicates whether the approach relies on the use of landmarks or simply the requested IP address. The third column indicates, whether the approach uses active methods (landmarks) or passive methods (IP) or a combination (*hybrid*). The fourth and the fifth columns provide information about the solution space and whether the program uses other auxiliary programs. Finally, the last column shows whether the project is actively maintained or discontinued.

B. Evaluation of Related Work

Performing an evaluation of the approaches presented is not easy (see Table II). This is mainly due to the fact, that (i) not all of them publish information about the corresponding accurateness and (ii) the source code is not publicly available.

Since NetGeo was officially discontinued in 1999 and thus is no longer developed and - as a consequence - is no longer fully available as a web-based solution, this approach is no longer considered for further considerations within this paper [18], [19].

Due to the lack of access to the required information, GeoCluster and GeoTrack are also not considered within the architecture and the corresponding proof-of-concept [12].

III. Overview of our Architecture

The previous section has shown that each strategy is subject to certain restrictions. E.g., the accuracy is too low, a complex infrastructure is needed for the execution of the programs or no selection of active or passive measurements is possible.

This is also reflected in a direct quote from Endo et al. [9]:

> One can see that each strategy suffers from certain restrictions. Therefore the use of a well thought out hybrid technique may improve the geographic location estimation. As a result, a strategy that combines different locations inferences would offer better results, since these strategies obtain information from different sources to estimate Geolocation.

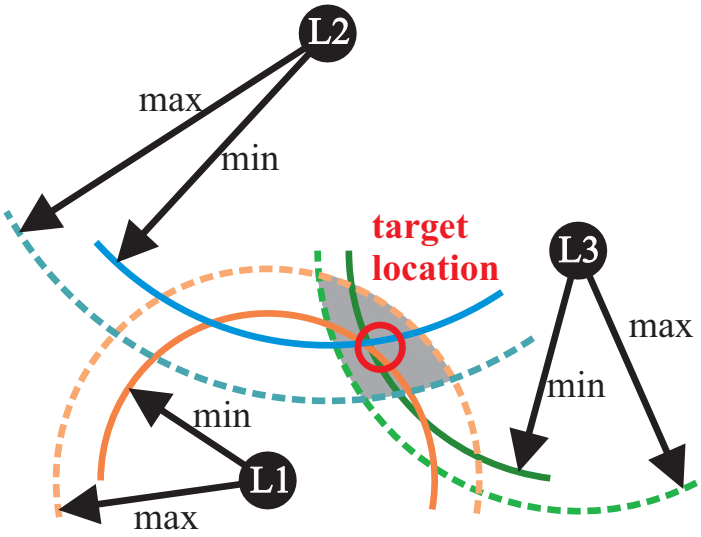

Fig. 3. Multilateration used in Constraint Based Geolocation [16], [11]

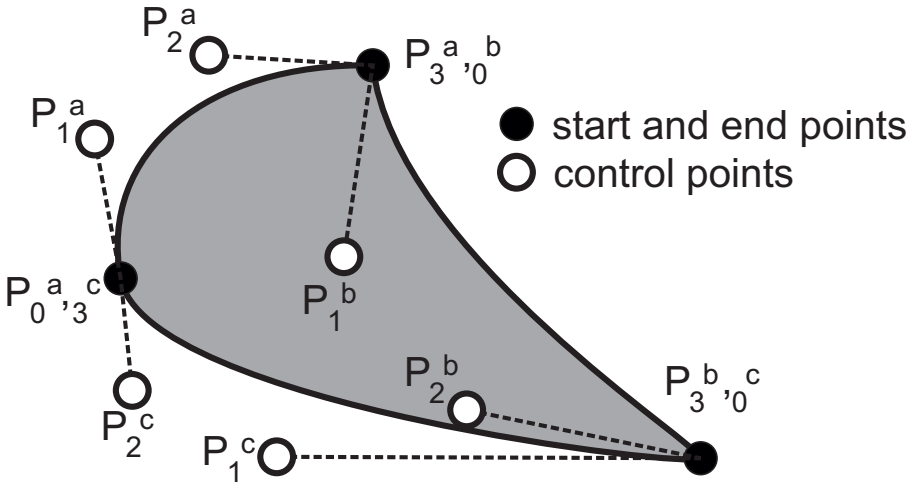

Fig. 4. Octant - Bézier Graph [17]

Following this idea, we present a new algorithm which combines several Geolocation techniques to increase the accuracy on the one side and which can be configured regarding analyzing techniques used (active or passive) on the other.

A. Components of the Algorithm

For the development of a Geolocation strategy, the use of databases of geoservice providers marks the first step for further investigations. Although the accuracy and credibility of geodatabases is considered questionable [23], [24], they can still be used for an accurate geographic position indication at the country level. Thus, the allocation of the requested IP address to a country - with the use of geodatabases - provides a solid base for further steps, for example, the restriction to a specific language/code (used for pattern matching). Furthermore, the use of multiple geospatial databases creates a large dataset, on which the corresponding results can be mutually verified. The result is weighted based on the quality of the used data (see Section II) and verified in two further steps. For that, databases of the RIRs are used first. After that, an additional verification process is done based on code databases. The setting of the weighting factors is done based on the analysis of empirical studies [25], [23], [26], [27]. In detail, the weighting and verification is done as follows:

1) Geodatabases: In order to draw conclusions about the accuracy of the geodatabases, only those approaches are considered that have already been investigated by empirical studies [23], [27], [25], [26]. Out of the available geodatabases, four have been selected and are used for the localization process: MaxMind, HostIP, IP2Location and IPInfoDB. According to [23], MaxMind comprises the larges number of cities including longitude an latitude information. Therefore within our architecture, this dataset has the highest influence of about 50 percent. The database of the provider IPInfoDB follows

TABLE I
OVERVIEW OF GEOLOCATION APPROACHES

METHOD	LANDMARK/ IP	ACTIVE/PASSIVE/ HYBIRD	SOLUTION SPACE	AUXILIARY PROGRAM	MAINTAINED
IP2Geo - GeoPing	Landmark	active	discret	-	yes
IP2Geo - GeoCluster	IP	passive	discret	-	no
IP2Geo - GeoTrack	IP	passive	discret	traceroute	yes
CBG	Landmark	hybrid	continuous	-	yes
TBG	Landmark	hybrid	continuous	-	yes
Octant	Landmark	hybrid	continuous	-	yes
NetGeo	IP	hybrid	discret	whois	no
Geoservices	IP	passive	discret	-	yes
Whois	IP	passive	discret	whois	yes
DNS LOC	IP	passive	discret	-	yes

TABLE II
EVALUATION OF RELATED WORK

METHOD	LOCALISATION LEVEL	ACCURACY
IP2GEO	Country/ISP	98%
(GeoPing,GeoCluster, GeoTrack) [20]	Region	75%
	City	63%
CBG [15]	Western Europe *(Median Error)*	below 25 km
	U.S. *(Median Error)*	below 100 km
OCTANT [21]	*Median Error*	22 miles
STRUCTON [22]	Province	93,5%
	City	87,4%

with 25 percent. According to Poese et al. and Huffaker et al. [23], [25] the use of the MaxMind Lite version is the reason therefore. Because of the higher error rates when determining address blocks [23], IP2Location is used with only 15 percent weighting. Finally, HostIP enters with 10 percent based on their purely voluntary listings which can hardly be verified by the provider. Also, the localization is limited to /24-blocks at HostIP.

2) Databases of RIR: After the analysis of the IP address based on geodatabases is performed, data of the RIRs is used for the verification of the results. Therefore, the cost-free *whois*-service is integrated into the algorithm. Unfortunately, the different RIRs are using different query and output schemes; it needs to be differentiated where the address is registered to execute a direct query. Pattern matching and regular expressions are used to analyse and extract the geoinformation from the result sets.

3) Code Databases: For the further analysis of the FQDN, code databases are used by our algorithm. Four types of databases are available: city-, regional-, airport- and radiobeacon-codes. For the implementation, cumulated databases of the International Air Transport Association (IATA) and Very high frequency Omnidirectional Radio range (VOR) are used. In addition, beacon codes are considered, too. The network entity, the primary DNS server identified by the Start of Authority (SOA) Record and the hops in the catchment area of the target address identified by route tracing are examined. Therefore, the country code of the targets queried from one of our databases is used to narrow the results. Each FQDN is split into its individual segments with the help of pattern matching and regular expressions. The verification of the Geolocation - based on code-databases - is the last step of our weighting algorithm.

In order to merge the output of the three components of the algorithm, different weightings are used based on the concrete usage of the algorithm. In particular, three basic modes are implemented, which are presented in more detail in Section IV., where the specific weights for this merging are illustrated, too.

B. Sources of Error

Because of the heterogeneity of the different databases and information sources, different errors are possible. On the one side, the transfer of million of records into a common format is nontrivial. On the other side, other aspects have to be considered, e.g., failures when querying the RIR databases or errors in the `traceroute` runs. The different sources of error in the weighting and verification process are:

1) Geodatabases:
As different empirical studies have shown, geodatabases are not supplying complete results [25], [23], [26], [27]. Deliberate or unwanted falsifications are possible within the data sets. E.g., the completely voluntary filled database of HostIP is particularly a risk for corruption. Therefore, it has the lowest weight in our evaluation. With multiple verification, as used in our algorithm, different results can be recognized and attenuated.

2) Databases of RIRs:
Of course, also the databases of the RIRs can have errors or can be manipulated. Also, as already mentioned, there is no standardized query which hampers the automatic evaluation of addresses.

3) Code Databases:
The most important problem arising by the use of code databases is the overlapping of information. Especially the airport- and radiobeacon-codes can have many overlaps among each other. Table IV gives an example for Frankfurt, Germany.

As illustrated in the example, all correct results can be assigned to Germany. Therefore, it is recommended to use

TABLE III
OVERVIEW OF THE NUMBER OF RECORDS WITHIN EACH GEOGRAPHIC DATABASE

PROVIDER	ADDRESS BLOCKS	LAT/LONG	NUMBER OF COUNTRIES	NUMBER OF CITIES	INFLUENCE WITHIN THE ARCHITECTURE
HostIP	8 892 291	33 680	238	23 700	10 %
IP2Location	6 709 973	17 183	240	13 690	15 %
InfoDB	3 539 029	169 209	237	98 143	25 %
MaxMind	3 562 204	203 255	244	175 035	50 %

TABLE IV
CODE DUPLICATION AS SEEN FOR FRANKFURT ON THE MAIN

CODE	LOCATION	DATABASE	COUNTRY	COUNTRY CODE
FW	Frankfurt on the Main	VOR	Germany	DE
FFM	Frankfurt on the Main	VOR	Germany	DE
FFM	Minnesota (Fergus Falls)	IATA	USA	US
FRA	Frankfurt on the Main (Airport)	IATA	Germany	DE
FRA	Fora	VOR	Brazil	BR
FRD	Frankfurt on the Main	VOR	Germany	DE
FRD	Washington (Friday Harbor)	IATA	USA	US
ZRB	Frankfurt on the Main (Central Station)	IATA	Germany	DE
EDDF	Frankfurt on the Main (Airport)	IATA	Germany	DE

the country code as the precondition before the verification process (with the code databases) is performed. Of course, this may lead to additional errors, because a preliminary containment is required. Therefore, the influence of the code databases is set to low in the weighting algorithm.

IV. PROOF OF CONCEPT

For the further evaluation of our architecture, a Proof of Concept (PoC) was implemented. Because of the requirement of an easy-to-use of the tool in a shell and the possibility to pipe the results into other programs, the PoC was implemented as a shell-script. The stealthiness and correctness of the results are two main aspects for the use of our tool. Therefore, different program modes have been realized.

A. Program Modes

Three basic modes are available to cover the requests for assignment, namely:

- *Paranoid:* Passive localization techniques are used; therefore no direct connection to the host under examination is needed. A combination of geoservice and code-databases as well as whois-queries are the basis for the realization. The IP address of the host as well as the primary DNS server of the zone are analyzed. The paranoid mode has a limited amount of verification techniques of the determined location and a short execution time.
- *Regular:* In contrast to the paranoid mode, more verification techniques are used to control the result of the localization process. Therefore, Route Tracing is used to identify the path to the target IP address. *TCPTraceroute* and *Paris Traceroute* are used for this purpose which offer different functionalities, e.g., regarding the supported protocols. Beside the target IP address, the last two hops before the target or the last resolvable hop are used for the verification of the location. Based on the assumption of the last mile [28], a matching is done with the results of the target. This mode is more precise, but the execution takes more time because of the active components involved in the analysis.
- *Aggressive:* This mode integrates the network scanner *Nmap* and executes a detailed scan of the target (in order to perform the first steps to identify and examine the network structure of the opponent; see introduction). The results of the scan are included in the output of the localization process.

Beside the different modes of operation, several options can be defined by switches on the command line:

- *Level of precision:* The user can select between most detailed information, output on city level or output on country level.
- *VPN:* The tool is able to use a configured OpenVPN installation to cover up the real origin of the analysis. This is particularly useful to hide the origin of ones one location (especially useful with *Regular* and *Aggressive* mode).
- Instead of a single IP address on the command line, a file can be defined for an automated localization of all addresses included in the file. The output is also redirected to another file.

B. Adaption of the Algorithm

Based on the different program modes, more or less information is available for the verification of the localization result. Especially the use of Route Tracing techniques and the verification of the surrounding hops is useful for the verification process. The verification information is used to calculate the probability of localization results. Because the maximum probability must be reachable in every mode - but with a different number of verification techniques - the weights have to be adapted accordingly. Therefore, the exact weights of our algorithm are adapted to the different program modes as shown in Table V.

In more detail, the basis for the localization process are the geoservice databases. Depending on the program mode and the further verification possibilities, their influence for the

TABLE V
WEIGHTING ALGORITHM - INFLUENCE OF THE DIFFERENT PROGRAM MODES

	GEO-DATA-BASES	RIR-DATA	CODE-(DATABASES) IATA, VOR, ...
COUNTRY (REG./AGGR.)	∼ 65%	∼ 23%	∼ 12%
CITY (REG./AGGR.)	∼ 50%	∼ 29%	∼ 21%
COUNTRY (PARANOID)	∼ 81%	∼ 10%	∼ 9%
CITY (PARANOID)	∼ 63%	∼ 20%	∼ 17%

localization differs between 50 and 81 percent. The geoinformation extracted from the RIR databases is matched with the results of the Route Tracing in the program modes *Regular* and *Aggressive*, therefore having a higher influence in these modes. Finally, the code databases have the lowest influence on the localization process. Due to the higher number of city codes compared to the country codes, this must be reflected in the localization calculation. Therefore, the weight of information about the identification of cities is higher weighted than information about countries.

C. IP Localisation Tool

The PoC is implemented as a shell script for the Bash. After the tool is started, a configuration file is read out, which contains access data for a MySQL database. If the configuration file cannot be found, an initial file will be generated, which can be adapted to the respective system and services. If the database does not contain tables, an initialization is started where the tables are generated. After that, the databases of IATA, MaxMind, etc., are downloaded and the tables are generated. If the tables are already available in the database, they could be updated after the start of the program. Regular updates are recommended, because significant changes can be observed in the databases over a specific period of time [27]. Therefore, this option is included in our tool.

When the databases are available, the actual program core with its three program modes can be executed. First, if the option for using a VPN was set on the command line, a VPN connection is initialized based on an existing OpenVPN installation. After that, it is checked if only one IP address is given on the command line or if a complete set is given in an additional file. Based on that, the localization process starts in the program mode requested, returning the results to the console or to an output file. Details about the program run of the regular mode are shown in the flowchart in Figure 5.

First, the required arrays and variables are deleted. After that, the validity of the input IP address is checked. If the address is not valid or if the IP address belongs to a reserved address range, the program aborts. If the address is valid, the traceroute operations are started. In order to increase the likelihood of a complete resolution of the path, different protocols and traceroute programs are combined. Therefore, *TCPTraceroute* and *Paris Traceroute* are used in combination with the protocols TCP, UDP and ICMP. After the traces have been finished, the output is verified and truncated. Unnecessary characters and unresolved hops are deleted (see Table VI).

After that, the analysis of the hops can be done. If it was possible to resolve the complete path, the last two hops are

TABLE VI
OUTPUT VERIFICATION AND TRUNCATION

```
tcptr -w1 -q1 -m30 $1 2> /dev/null | sed '/^ *$/d' > t0
paris-tr -Q -m30 -T1000 -pudp -q1 $1 | sed '/^ *$/d' > p10
paris-tr -Q -m30 -T1000 -ptcp -q1 $1 | sed '/^ *$/d' > p20
paris-tr -Q -m30 -T1000 -picmp -q1 $1 | sed '/^ *$/d' > p30
```

examined. Otherwise, the last hop that can be resolved is examined. Finally, if a VPN connection had been used, the connection is closed down and the program run is completed.

D. Known Shortcomings

The current prototypical implementation has some shortcomings: first, errors can arise by the aggregation of heterogeneous information sources to a homogeneous database. For example, an error in the aggregation process can result in an empty outcome. Another aspect is the use of route tracing for the verification of localization results. E.g., firewalls can hamper the resolution of the path. This problem can be reduced by the use of different tracerouting programs and protocols, but it cannot be eliminated. In addition, it is possible that the primary as well as the secondary DNS server is not assignable to the country of the target system, therefore affecting the evaluation. The already mentioned non-standardized queries of the different RIRs can hamper the automatic processing of addresses when trying to extract geographic information. At least, temporary traffic loads can result in the termination of a traceroute, which can result in a devaluation of the localization result because of the missing verification process.

At the moment, only services and databases free of charge are used for the evaluation. It is possible that the detection accuracy can be improved by the use of commercial geodatabases because of their possibly larger datasets.

V. EVALUATION

For the evaluation of our Proof of Concept, testruns with known *(IP-Address, Location)* tuples have to be fulfilled. In practice, it is difficult to obtain this groundtruth. To generate a corresponding database, the method described below has been used.

A. Collective Test Data

The website *NeedMoreCookies* was used to collect the tuples required [29]. This website is a content crawler which integrates selected content. Based on the evaluation of the *Real Time Web Analytics Service Clicky* which is implemented in the backend of *NeedMoreCookies*, 3629 datasets had been acquired with the distribution shown in Table VII.

Because the service *Clicky* is using the commercial variant GeoIP of the provider MaxMind, the ground-truth has a success rate of 96% to 98% on country level [23].

B. Testing Procedure

The collected data of the website *NeedMoreCookies* was extracted and transformed to be processable by our localization

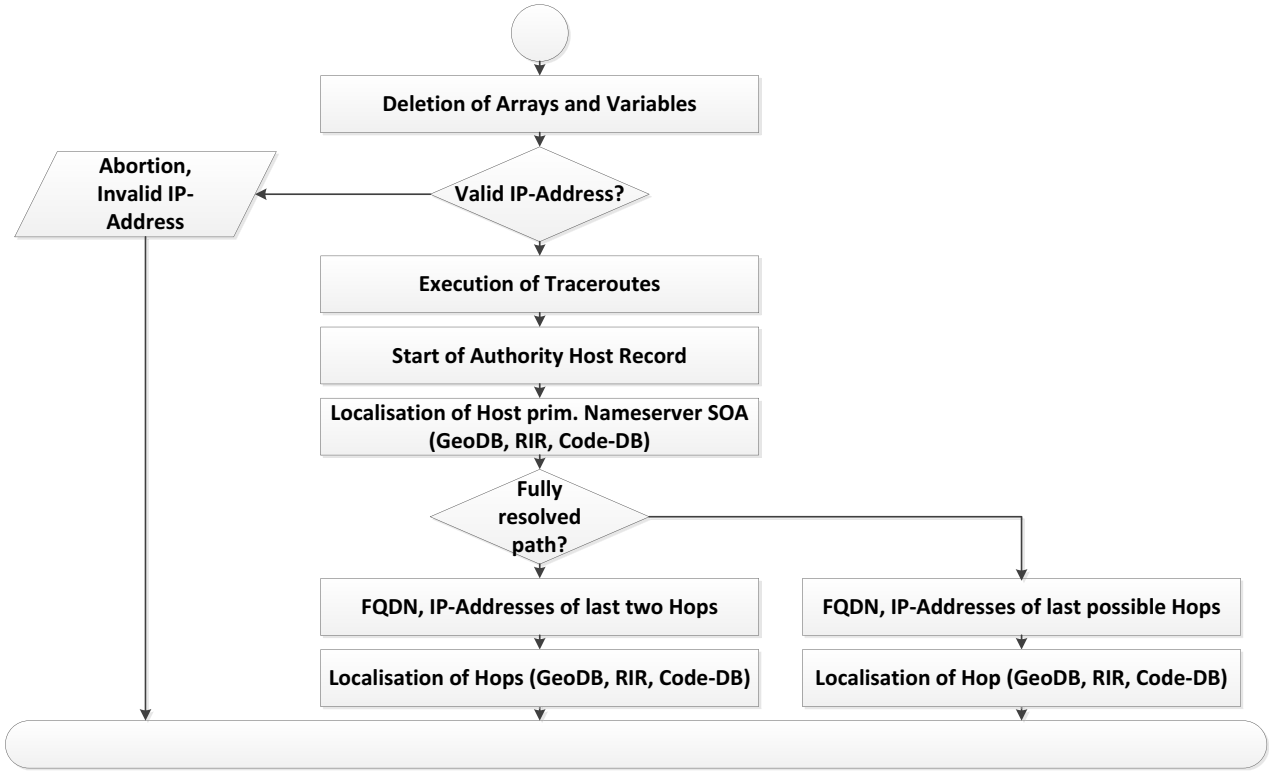

Fig. 5. Program flowchart for IP-Localisation, Regular Mode

TABLE VII
ACQUIRED DATASETS AND THEIR PERCENTAGE DISTRIBUTION

COUNTRY	NUMBER OF ADDRESSES	PERCENTAGE
USA	1725	47,53%
Germany	1018	28,05%
UK	354	9,75%
Canada	334	9,20%
Australa	74	2,04%
Ireland	45	1,24%
Netherlands	42	1,16%
India	37	1,02%

tool. Therefore, a data manipulation language has been used to extract the IP addresses which subsequently have been put into our tool. Based on this addresses, an assignment to geographic locations has been done and the results have been stored in the database for an automated comparison. For the evaluation of the correctness on city level, a multilevel analysis has been done. This is necessary because of inaccuracies due to incorrect identifiers (e.g., *Kastellaun*, *Kastel Laun* and *Castelaun*) or slight alterations of longitude and latitude. First, all datasets are analysed for a concensus of the city names. If no correlation is possible, longitude and latitude will be evaluated. The next step is an analysis based on pattern matching and regular expressions to detect deformed identifiers (e.g., *Bernkastel-Kues* vs. *Bernkastel*). If no matching is possible, a staggered change of the longitude and latitude of about $\mp\frac{1}{2}^\circ$ respectively $\mp 1^\circ$ (1° corresponds on average an deviation of approx. 100 kilometers) is used to narrow the geographic location. Only if one of these steps was successful, a positive rating is done on city level.

C. Evaluation Results

Because of the limited space, only an excerpt and summary of our evaluation results is given. Table IX presents the evaluation results on country level for the regular and paranoid mode, while Table VIII is giving the overview of our results.

As illustrated, the paranoid mode is good enough to gain a high evaluation accuracy on country level; no improvement was possible by using the regular mode. In contrast, the results on city level can be improved by using the regular mode. Because the improvement is modest, the paranoid mode is sufficient to reach a high location accuracy down to city level. Of course, at the moment all evaluation data was obtained from industrialized countries. For the evaluation of our subsequent implementation, we are trying to use a more distributed dataset, including developing countries.

VI. CONCLUSION AND OUTLOOK

Geolocation can contribute to network security in many ways; both in advance of an Intrusion (*ex ante*) to ward off an attack, as well as after an Intrusion was successful (*ex post*) to either support finding out from where the attack was launched or to examine the network structure of the opponent in order to counterattack.

TABLE VIII
EVALUATION OF RELATED WORK

	LOCALISATION LEVEL	ACCURACY
Existing Approaches		
IP2GEO	Country/ISP	98%
(GeoPing,GeoCluster, GeoTrack) [20]	Region	75%
	City	63%
CBG [15]	Western Europe *(Median Error)*	below 25 km
	U.S. *(Median Error)*	below 100 km
OCTANT [21]	*Median Error*	22 miles
STRUCTON [22]	Province	93,5%
	City	87,4%
Our Approach		
PARANOID MODE	Country	99,78%
	City	87,57%
REGULAR MODE	Country	99,78%
	City	90,49%

TABLE IX
EVALUATION OF THE REGULAR MODE ON COUNTRY LEVEL ($\mathcal{C}_{Land}$ ARE THE TUPLES TO EVALUATE REGARDING THE COUNTRY, $\mathcal{M}^{reg}_{Country}$ THE CORRECT EVALUATED TUPLES)

Country	$\mathcal{C}_{Country}$	$\mathcal{M}^{reg}_{Country}$	ERROR
USA	1725	1724	1
Germany	1018	1016	2
UK	354	353	1
Canada	334	332	2
Australia	74	74	0
Ireland	45	43	2
Netherlands	42	42	0
India	37	37	0
$\sum$	3629	3621	8

Several techniques have been proposed for the Geolocation of IP addresses. Although all of them are usable in principle, each of them has specific restrictions, e.g., the accuracy is low or special infrastructure is needed. Therefore, we propose an architecture which combines several approaches to increase the accuracy and the efficiency of the localization process (see Table VIII). Verification techniques are used to evaluate the probability of the solution. Based on that, we are able to identify locations of IP addresses with an accuracy of about 99% on country level and about 90% on city level. At the moment, we are developing the next generation of our algorithm which will include further active measurement techniques. Because of performance reasons, this time, we are using Perl for the implementation. For further evaluations of our prototypes, a broader database will be used. In particular, we are acquiring additional *(IP-Address, Location)* tuples, focusing on developing countries. Also, the inclusion of commercial databases would be of interest regarding the accuracy of the detection results.

ACKNOWLEDGEMENT

The authors wish to thank the members of the Chair for Communication Systems and Internet Services at the Universitt der Bundeswehr Munich, headed by Prof. Dr. Gabi Dreo Rodosek, for helpful discussions and valuable comments on previous versions of this paper. The Chair is part of the Munich Network Management Team.

In addition, special thanks go to Lars Stiemert and Kay Gottschalk, who made a significant contribution for this publication with their high-grade bachelor's thesis [30].

This work was partly funded by Flamingo, a Network of Excellence project (ICT-318488) supported by the European Commission under its Seventh Framework Programme.

REFERENCES

[1] T. Lewis, "Index," *Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation*, pp. 463–474, 2006.
[2] Symantec, "Symantec 2011 SMB Disaster Preparedness Survey - Global Results," 2011, http://www.symantec.com/content/en/us/about/media/pdfs/symc_2011_SMB_DP_Survey_Report_Global.pdf.
[3] Mandiant, "APT1 - Exposing One of Chinas Cyber Espionage Units," 2013, http://intelreport.mandiant.com/Mandiant_APT1_Report.pdf.
[4] Chen Jie, Ministry of National Defense, The People's Republic of China, "China has no cyber warfare troops: spokesman," 2013, http://eng.mod.gov.cn/Press/2013-03/01/content_4434894.htm.
[5] Lana Lam, South China Morning Post, "Edward Snowden: US government has been hacking Hong Kong and China for years," 2013, http://www.scmp.com/news/hong-kong/article/1259508/edward-snowden-us-government-has-been-hacking-hong-kong-and-china.
[6] M. Roesch *et al.*, "Snort-lightweight intrusion detection for networks," in *Proceedings of the 13th USENIX conference on System administration*. Seattle, Washington, 1999, pp. 229–238.
[7] J. Quittek, T. Zseby, B. Claise, and S. Zander, "Requirements for ip flow information export (ipfix)," *IETF RFC3917, Oct*, 2004.
[8] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, and B. Stiller, "An overview of ip flow-based intrusion detection," *Communications Surveys & Tutorials, IEEE*, vol. 12, no. 3, pp. 343–356, 2010.
[9] P. Endo and D. Sadok, "Whois based geolocation: A strategy to geolocate internet hosts," in *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*. IEEE, 2010, pp. 408–413.
[10] A. Dahnert, "Hawkeyes: an advanced ip geolocation approach: Ip geolocation using semantic and measurement based techniques," in *Cybersecurity Summit (WCS), 2011 Second Worldwide*. IEEE, 2011, pp. 1–3.
[11] S. Laki, P. Mátray, P. Hága, T. Sebok, I. Csabai, and G. Vattay, "Spotter: A model based active geolocation service," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 3173–3181.
[12] V. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for internet hosts," in *ACM SIGCOMM Computer Communication Review*, vol. 31. ACM, 2001, pp. 173–185.
[13] A. Ziviani, S. Fdida, J. de Rezende, and O. Duarte, "Improving the accuracy of measurement-based geographic location of internet hosts," *Computer Networks*, vol. 47, no. 4, pp. 503–523, 2005.
[14] M. Zhang, Y. Ruan, V. Pai, and J. Rexford, "How dns misnaming distorts internet topology mapping," in *Proceedings of the annual conference on USENIX'06 Annual Technical Conference*, 2006.
[15] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. ACM, 2004, pp. 288–293.

[16] B. Gueye, S. Uhlig, A. Ziviani, and S. Fdida, "Leveraging buffering delay estimation for geolocation of internet hosts," *NETWORKING 2006. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, pp. 319–330, 2006.
[17] B. Wong, I. Stoyanov, and E. Sirer, "Octant: A comprehensive framework for the geolocalization of internet hosts," in *Proceedings of the NSDI*, vol. 7, 2007.
[18] D. Moore, R. Periakaruppan, J. Donohoe, and K. Claffy, "Where in the world is netgeo. caida. org." INET, 2000.
[19] "Cooperative Association for Internet Data Analysis. NetGeo." http://www.caida.org/tools/utilities/netgeo/.
[20] Jgsoft Associates, "IP2Geo: Frequently Asked Questions, How accurate is IP-Country-Region-City-ISP database?" 2013, http://www.ip2geo.net/ip2location/ip-country-region-city-isp-faq.html.
[21] B. Wong, I. Stoyanov, and E. G. Sirer, "Geolocalization on the internet through constraint satisfaction," in *Proceedings of the 3rd conference on USENIX Workshop on Real, Large Distributed Systems*, 2006, pp. 1–1.
[22] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang, "Mining the web and the internet for accurate ip address geolocations," in *INFOCOM 2009, IEEE*. IEEE, 2009, pp. 2841–2845.
[23] I. Poese, M. A. Kaafar, B. Donnet, B. Gueye, and S. Uhlig, "Ip geolocation databases: Unreliable?" Deutsche Telekom Lab./TU Berlin, Technical Report, March 2011.
[24] S. Laki, P. Mátray, P. Hága, I. Csabai, and G. Vattay, "A model based approach for improving router geolocation," *Computer Networks*, vol. 54, no. 9, pp. 1490–1501, 2010.
[25] B. Huffaker, M. Fomenkov, and K. Claffy, "Geocompare: a comparison of public and commercial geolocation databases," Technical Report, May 2011, network, Mapping and Measurement Conference (NMMC).
[26] S. S. Siwpersad, B. Gueye, and S. Uhlig, "Assessing the geographic resolution of exhaustive tabulation for geolocating internet hosts," in *Passive and Active Network Measurement Workshop (PAM)*. Springer-Verlag, 2008, pp. 11 – 20.
[27] Y. Shavitt and N. Zilberman, "A study of geolocation databases," School of Electrical Engineering, Technical Report, July 2010.
[28] M. Dischinger, A. Haeberlen, K. Gummadi, and S. Saroiu, "Characterizing residential broadband networks," IMC, Technical Report, 2007.
[29] K. Gottschalk, "NeedMoreCookies: The Funstuff Crawler," Website, http://needmorecookies.com/.
[30] K. G. Lars Stiemert, "Geolocalization and Verification of IP-Adresses; German: Geolokalisation und Verifikation von IP-Adressen," Master's thesis, Institut für Technische Informatik, Universitt der Bundeswehr Mnchen, Germany, 2012. [Online]. Available: https://www.unibw.de/inf3/forschung/dreo/publikationen/ba-und-ma/2012_Stiemert-Gottschalk_Geolokalisation.pdf

Robert Koch is a Research Assistant at the Universität der Bundeswehr München (UniBwM). He received his Diploma in Informatics in 2002 and his PhD in 2011 from the UniBwM. His main areas of research are network and system security with the focus on intrusion and extrusion detection in encrypted networks, security of COTS products, security visualization and the application of artifical intelligence. He has several years of experience in the operation of high security networks and systems.

Mario Golling is a PhD student at the Universität der Bundeswehr München (UniBwM), where he graduated in business informatics in 2007. His key aspects of research activity are network security, cyber defence, intrusion detection and next generation internet. He has many years of experience in running operational networks as well as teaching and training network administration/security. Among other things, he is a member of the Working Group IT Security of the UniBwM.

Gabi Dreo Rodosek holds the Chair of Communication Systems and Internet Services at the Universität der Bundeswehr München. She received her MSc. from the University of Maribora and her PhD from the Ludwig-Maximilians University Munich. She is spokesperson of the Research Center Cyber Defence (CODE), which combines skills and activities of various institutes at the university, external organizations and the IT security industry (for instance Cassidian, IABG or Giesecke & Devrient).