Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080351

Forensics Tracking for IP User using the Markov Chain Model

Feng-Yu Lin[1,*], *Yeali S. Sun*[1] *and Meng Chang Chen*[2]

[1] Department of Information Management, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan
[2] Institute of Information Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan

Received: 22 Jun. 2013, Revised: 27 Oct. 2013, Accepted: 28 Oct. 2013
Published online: 1 May. 2014

Abstract: Increasingly severe cybercrime results in heavy impact and loss for society and public security of various countries, and even influences Homeland security. Based on previous experiences, the most effective method to resist cybercrime and reduce its impact and damage is that law enforcement agencies (LEA) identify and arrest criminals within the shortest time possible after a crime occurs. Therefore, IP location and IP Individualization play key roles. In view of this, this study proposes an IP user tracking forensics mechanism, based on the concepts of IP location and computational forensics, to develop forensics tracking on the Internet. The proposed mechanism can instantly trace the "physical location" of cybercriminals when cybercrimes occurs, analyze the probable "identity" associatively, and reconstruct the historical physical path of the cybercriminal. The proposed mechanism was implemented for verification. The results showed that the accuracy of IP location is 0 m error on the fixed network, while the mobile network could reach Cell-ID covered range (150–500 m radius in urban area). The identity of cybercriminal could be successfully reasoned out, with segmental paths (e.g. tracking breakpoints) reconstructed by algorithm, thus, obtaining the complete path of the target. The average success ratio (predictable ratio) was 90.91%, and the accuracy rate was 88.70%.

Keywords: Cybercrime, IP location, Computational Forensics, Tracking Forensics, Law Enforcement Agency (LEA), Homeland Security

1 Introduction

Increasingly severe cybercrime has resulted in heavy impact and loss for society and the public security of various countries, and even influences Homeland security. Based on previous experiences, the most effective method to resist cybercrime, and reduce its impact and damage, is that the law enforcement agencies (LEA) identifies and arrests criminals within the shortest time possible after a crime occurs. However, Internet network hosts present their logical locations by IP addresses, which are not spatial locations associated with physical points or regions on the surface of the globe. Furthermore, as the subject of IP has the issues of DHCP and NAT, it is difficult to individualize users. Therefore, developing an IP location method/mechanism, which considers how to map IP addresses with physical location to determine who used the IP address (i.e. individualize IP user), becomes an emergency issue for homeland security.

Reviewing related works, most studies [1,2,3,4,5,6,7,8,9,10,11,12,13] focus on IP Traceback, and discuss the identification of the sources of attacks and instituting protection measures for the Internet. IP location has not yet been thoroughly studied.

IP location is the process of finding the geographic location of an Internet host [15]. In an effort to offer this service, researchers have proposed various IP location mapping schemes, such as TBG [14], IP2Geo [16], CBG [17], Octant [18] and GeoGet [19]. They are primarily delay-measurement based solutions, which measure the delay from a targeted client to landmarks, and then map the targeted client to a location inferred from the measured delays. However, there are some issues that still require solutions, as follows: 1) based on the end-to-end delay measurement (from a set of landmarks) method, there remain common weakness in that they cannot outperform much simpler techniques, and the errors of such approaches is determined by the distance to the nearest landmark, even when triangulation is used to combine estimates from different landmarks; 2) they do not consider the issues of DHCP/NAT; therefore,

* Corresponding author e-mail: d95725003@ntu.edu.tw

users of IP communications cannot be individualized; 3) such research focus primarily on fixed networks, without considering the problems on nomadicity and mobility, that is, users are able to be nomadic and access the Internet from multiple and relative arbitrary locations (Nomadicity). Further, there is a category of Internet access technologies that supports full mobility of the user allow a user to connect to the Internet and access services even while traveling in a car at high speed. The main impact is that static, predefined, database cannot to be use for obtaining user location based on IP [23].

Second, we claim that tracking a particular suspect may, in most cases, be necessary in criminal investigations; indeed, proving the location of a suspect at a particular time (or during a particular period of time) can be a significant piece of evidence in its own right. We call this class of tracking forensic tracking, which is basically conducting a tracking procedure for forensic purposes.

In the course of target trace reconstruction by IP location, the raw data extracted are usually broken data, namely, in trace reconstruction, the target is given a specific trace route. However, the extracted raw target trace often has multiple trace breakpoints. Therefore, how to convert these broken paths into a useful trace route, for use as evidence, is another urgent topic to be solved.

The IP user tracking forensics mechanism, as proposed in this study, is based on the concepts of IP location and computational forensics, which develops forensics tracking on the Internet, and aims to instantly trace the "physical location" of cybercriminals using this mechanism when cybercriminal behavior occurs, in order to analyze the probable "identity" associatively, and reconstruct the historical physical path of the cybercriminal.

The remainder of this paper is organized as follows. Section II reviews related works; Section III, introduces the proposed IP user tracking forensic mechanism; Section IV presents the empirical evaluation results for the proposed IP user tracking forensics mechanism, and discusses its strengths and weaknesses; Section V offers conclusions and suggestions to future research.

2 Related works

During the last decade, various IP location techniques have been proposed. In [16], V. N. Padmanabhan and L. Subramanian proposed three distinct techniques, GeoTrack, GeoPing, and GeoCluster, to perform IP location in 2001. GeoTrack is based on the DNS names of target host, or other nearby network nodes, using a traceroute tool to analyze IP location. GeoPing is based on delay-measurement from geographically distributed locations to target hosts and estimate the possible coordinates of the target host. GeoCluster infers target hosts' geographic location by combining partial host-to-location mapping information and BGP prefix information, and is the most promising method, with various median errors from 28 km for well-connected university hosts to hundreds of kilometers for more heterogeneous set clients; however, mapping information must be updated manually and periodically.

In 2004, B. Gueye et al. proposed a Constraint-Based Geolocation (CBG) approach, which employs a triangulation-like technique based on multilateration with distance constraints to infer geographic location. For accurate results, CBG estimates and removes additive delay distortion by self-calibrating the delay measurements. However, the median error distance product by the CBG approach can be reduced to less than 100 km, at the 25th percentile, with 15 to 25 km to landmarks [17]. In addition, in 2006 Ethan Katz-Bassett et al. employed network topology information to improve location accuracy, but the estimation errors for topology measurement are more than tens of kilometers [14]. Wong et al. [18] proposed a comprehensive framework in 2007, called Octant, for determining possible region by positive and negative constraints within 22 miles (around 35.4 km). In 2009, Dan Li et al. proposed an IP location mapping scheme, GeoGet, which involves moderately connected Internet regions by HTTP/Get probing for delay measurement, and the results show that it can accurately map 35.4% of targeted clients to the city level, with a median error distance of approximately 120 km [19]. In addition, in 2011 Sandor Laki et al. proposed a probabilistic location approach, Spotter, for estimating the geographic location of Internet devices by handling all calibration points together to derive a common delay-distance model. Therefore, with dataset of COGENT, and a large Tier-1 ISP, Spotter improves the median error to 30 km [20].

For [14,15,16,17,18,19,20], no matter the dataset used, they are primarily delay measurement based approaches, which try to measure the delays from a target client to landmarks, and then map the targeted client to a location inferred from the measured delays. As the delay problem of traffic congestion on a network is inevitable, an inherency feature of the Internet, it causes delay measurements with unpredictable errors in the round trip time for probing, which has the negative effect of random errors regarding geographic location. In addition, IP addresses behind a proxy or firewall cannot be detected for location.

There are some IP location approaches, e.g. the WHOIS-based [22] and DNS-based approaches [21], which perform IP location by querying the information on databases that store relative data or location regarding users when they registered with ISPs. However, there are some limitations, as the databases must be manually updated periodically and is for fixed users only. They can be more accurate than the delay-measurement based approach, if the registry information offered by users is correctly updated; however, the problem of private IP addresses cannot be solved by a database based approach.

Table 1: Termination Information vs. Access Information

	Termination Information	Access Information
Definition	The information identifies a particular terminal is called Terminal Information	The information related to how the terminal accesses the network is called Access Information.
Variability	Data seldom change	Data often change, especially in the environment of dynamic acquisition of IP
Acquisition mode	Provided by network administrator, also known as Provisioning Data.	Obtained from the existing network communication equipment or communication protocols

3 IP user tracking forensics mechanism

The proposed IP user tracking forensics mechanism is based on the concepts of IP location [23] and computational forensics, which develop forensics tracking on the Internet. The logic of IP location and individualization is that in an environment of an integrated fixed network, mobile telecommunication network 3/3.5G, WiMAX, and next generation network IP Multimedia Subsystem (IMS), each related node in the network is confirmed, the DPI (Deep Packet Inspection) is in charge of accessing, copying, decoding, and saving necessary data retention for IP location, in order to analyze the location of the IP Address and associatively identity IP users according to an IP Address and time and the information record left from using IP services on the Internet.

Secondly, this study uses the Markov chain to construct a trace reconstruction algorithm, in order to reconstruct segmental paths (e.g. trace breakpoints) by predicting the next step, thus, obtaining a complete path for the target (Computational Forensics).

3.1 IP location and individualization

In the IP user tracking forensics mechanism, as proposed in this study, the IP location mechanism includes two key component equipments, Distributed DPI Agents (DDA) and IP2Location Database (IPLD), and two XML-based Protocols, DDA to IPLD and IPLS to IPLD protocols. The IP location logic correlates the Access Information, as gathered by DDA and stored in the IP2Location Database (IPLD), with the Terminal Information obtained by the network management database, where the IP user and physical location at that time can be reasoned out through the correlation information chain, as shown in Figure 1, Figure 2, Figure 3 and Figure 4.

3.1.1 Termination Information

In this study, the information identifies a particular terminal is called Terminal Information, e.g. MAC Address, dial-up account, or auxiliary number. The Termination Information is provided by a network management database, also known as Provisioning Data. The ISP provider must provide the information related to user location (including geo information), which is transferred via Firewall to the Data Provisioning module inside the IP user tracking forensics mechanism system.

3.1.2 Access Information

The information related to how the terminal accesses the network is called Access Information. The DDA is responsible for providing Access Information when given Terminal Information, which data often changes, especially in the environment of dynamic acquisition of IP, and can be obtained from the existing network communication equipment (DHCP and/or RADIUS Server, AAA Server) or communication protocol packets.

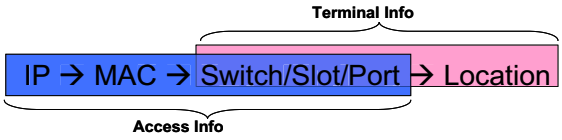

Fig. 1: Correlation information chain of enterprise LAN network

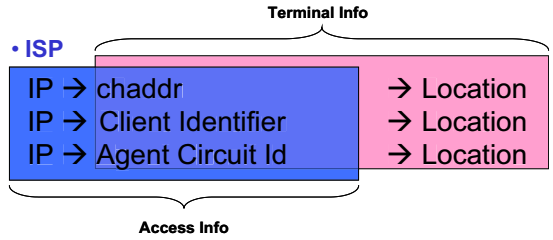

Fig. 2: Correlation information chain of xDSL network

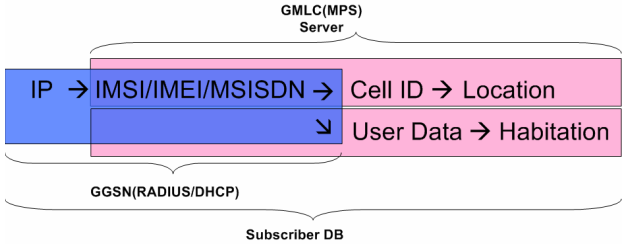

Fig. 3: Correlation information chain of 3G/3.5G network

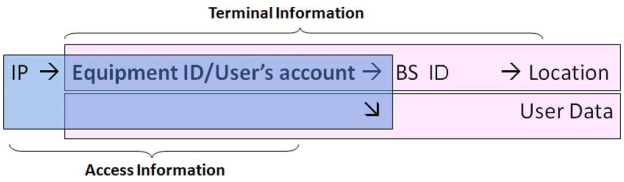

Fig. 4: Correlation information chain of WiMAX network.

3.2 IP user trace reconstruction

After the IP location and individualization analysis in Section 3.1, which ISP provider issued the IP can be known. If it is from a fixed network, there is no trace reconstruction. If it is from a mobile network, the historical track of the IP user (e.g. track formed of Cells) can be reconstructed, as based on User data obtained by individualization analysis, as well as the historical CDR (Call Data Record) of the billing system of the ISP provider.

3.3 Computational forensics for IP user

3.3.1 Data preprocessing

According to the reconstructed path data in Section 3.2, many records are incomplete and repeated, which is due to cell overlap at the location of the IP device and signal strengths of similar base stations in the region. When the user stays or passes by this location, the mobile phone has constant handover among the base stations with equivalent signal strength, forming a ping pong change that results in inconsistent data retention.

In order to address the above problem, uncertain data retention is filtered, such as base station jitter, short move, base stations in different directions, etc., in order to avoid unnecessary calculations and solve the problem of different positioning errors, as resulted from different Internet connection technologies (e.g. 3/3.5G, WiMax, LTE) used by the IP user. This study attempts to use spatial clustering for data preprocessing, where the data of an IP location results in the same area coverage (i.e. base station location (cell) of all types of access obtained by IP location in the range) and are recorded in a single record. On one hand, the repeated data record of base station jitter and short move can be filtered; on the other hand, various access network location results can be normalized.

The concept of spatial clustering is to express the space as tile matrix set, and the tile center point coordinates are used as the area coordinates after clustering (Figure 5). This study splits the tile based on splitting geographic space, as defined by the "OpenGIS® Web Map Tile Service Implementation Standard" [24]. The Web Map Tile Service Implementation Standard is the GIS standard proposed by "Open Geospatial Consortium, Inc.". Based on this standard, we can designate the IP location result coordinates (base station position coordinates) as our tile coordinates in order to normalize the base station location and the actual position of the end IP device. Meanwhile, as we can flexibly change the space unit size, repeated data sources can be effectively removed under different requirement conditions.

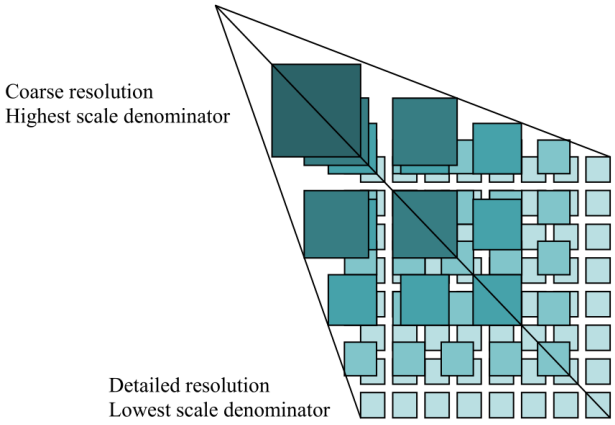

Fig. 5: Schematic diagram of tile matrix

Time clustering is conducted after spatial clustering. The concept of time clustering regards a tile repeatedly occurring within a fixed time interval as occurring once.

Finally, this study combines locations, and different sections are combined for convenient prediction if the connecting portion is in the same tile position.

Algorithm 1. Spatial clustering algorithm

```
1:  Input: S: Series of the cells' coordinates,
    L: Level of the standard
2:  Output: C: Set of clusters
3:  procedure GetClusterTile(S,T)
4:      C:= {}
5:      for s in S do begin
6:          c := TileStandard.GetTile(s,L)
7:          C.Add(c)
8:      end
```

3.3.2 Markov Chain

The Markov Chain refers to a discrete-time random variable set within the Markov process chain. Therefore, the Markov Chain is also a random model. The movement tracking (behavior) of individuals can be regarded as a multi-state process. For example, the movement behaviors of individual's varies with time, such as daily rest locations, travelling to a company for business, meals in familiar locations, etc. Such behaviors can be simulated by the Markov Chain; therefore, it can be used to infer the changing states of movement behaviors in different periods.

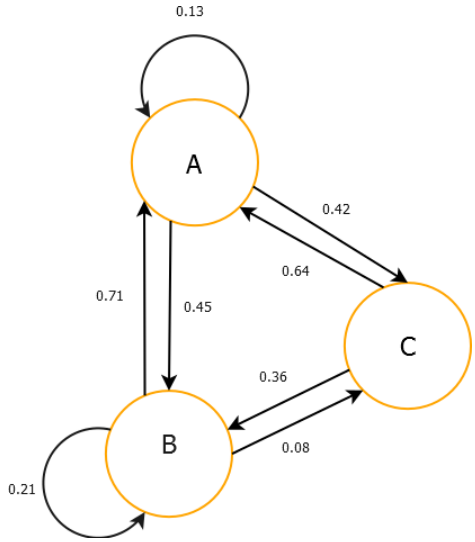

Fig. 6: Markov Chain state transition diagram. This diagram describes the probability of transition of States A, B, and C.

In the Markov Chain, each state transition (from current state to future state or maintaining the current state) is determined by current information or knowledge. The previous information (i.e. historical state before now) is unrelated to predicting the future state. In other words, the Markov Chain uses (first-order) conditional probability to predict the future state. The mathematical definition, and important particularities of the Markov process, are described, be as follows. For example, random process $\{X_t\}(t = 0,1,2,\ldots)$, where X_n refers to the state of x at time n. If the following equation is met, this random process has the Markov process.

$$P\{X_{t+1} = j | x_t = i, x_{t-1} = i_{t-1} = i_{t-1}, \ldots, x_0 = i_0\} = P\{X_{t+1} = j | x_t = i\}$$

Where i, j represent different states. This equation describes that the (future) state at time $t+1$ is only correlated with the state at time t (previous time), and is unrelated to the previous state $t-1,\ldots,0$ (former $n-1$ time).

$P\{X_{t+1} = j | x_t = i\}$ can be expressed as p_{ij}, p_{ij} is called transition probabilities, meaning in state i, the probability of the next state entering in state j. In addition, as the probability is not negative, for any i and j, $p_{ij} \geq 0$

$$P_{ij} = P\{x_{t+1} = j | x_t = i\}$$

The transition matrix is the set of all of changes in the random process. It records the transition probabilities of all states in the random process. A Markov Chain, with $n \times n$ states, can be expressed as $n \times n$ transition matrix P, containing transition probabilities $P_{ij}(1 \leq i \leq n, 1 \leq j \leq n)$

$$P = \begin{vmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{vmatrix}$$

In the transition matrix, each row represents the various transition probabilities of a specific state. As the sum of probabilities is 1, the sum of each row in the transition matrix is 1.

$$\sum_{i=1}^{n} p_{i1} = 1$$

The transition matrix of the example in Figure 6 is, as follows:

$$P = \begin{vmatrix} 0.13 & 0.71 & 0.64 \\ 0.45 & 0.21 & 0.36 \\ 0.42 & 0.08 & 0 \end{vmatrix}$$

In terms of Row 1, it means the probability of transition of State A to states B, C, and itself, is 0.45, 0.42, and 0.13, respectively. In addition, the transition matrix describes an important feature, in which the random process exists in only one state at any time.

The transition probabilities and matrix can help with path prediction. The transition matrix of a path file is calculated first, and then the position of a real-time path in the transition matrix is matched, in order to determine the most probable position (maximum probability). Finally, it is matched with the actual path to calculate the accuracy rate. To sum up, the computing process of the Markov Chain is described, as follows:

1.Analyze historical path and calculate the transition matrix.
2.Analyze path, pick the next position of maximum probability for current position as the forecast group.
3.Select unique forecast result from the forecast group.

3.3.3 Markov chain prediction model

Before prediction, the Markov chain prediction model calculates the probability of various Tiles (spatial clustering) moving to different Tiles (spatial clustering), as based on historical paths, and determines the probability of the current path point moving during prediction. If the time between two points exceeds a specific time (1 hour at present), it is regarded as a noncontinuous path, and is excluded from the calculation of movement probability (Figure 7).

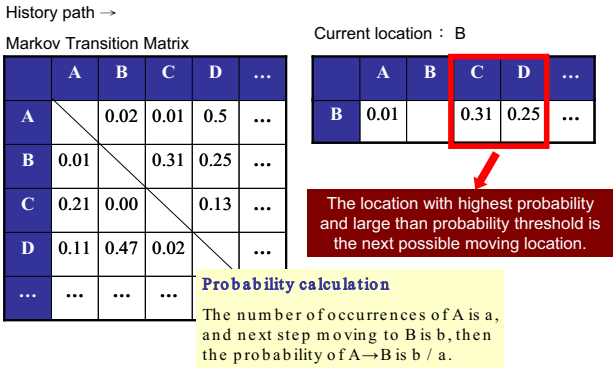

Fig. 7: Path prediction (Markov chain prediction model)

Algorithm 2 Markov Chain Computational Forensics Model

Input: H: History Route, L: Current Location
Output: P: Prediction Location
procedure MarkovModelPrediction(H,L)
  r := {from, to, count} // r: Record the times from one to other.//
  c := |H| // c: History Count//
  for $i := 1$ **to** c **do begin**
    if r.contains(from.equal(H[i]) **and** to.equal(H[$i+1$])) **then**
      r.count := r.count + 1
    else
      r.Add(H[i], H[$i+1$], 1)
    end
  end
  l := where r.from.equal(L)
  l' := sortByCount(l)
  p' := maxCount(l')
  if $|p'| > 1$ **then**
    p' := getBestResult(p')
  end
  $P = p'$

3.4 Efficiency evaluation index

In the evaluation of an algorithm, we use three indices, including "accuracy rate", "predictability", and "uncertainty", in order to evaluate the effect of each prediction. The meaning of each index is detailed, as follows.

i. Accuracy rate

The "accuracy rate" is used to evaluate the percentage of correct prediction. Its equation is defined, as follows.
Accuracy rate = (total number of predictions − number of wrong predictions) × 100 (%)/total number of predictions

The IP location result of the proposed IP user tracking forensics mechanism is measured by the "accuracy rate" index.

ii. Predictability

The "predictability" is applied to multiple prediction results in the same prediction. If one result can predict the actual path, there is a chance of successful prediction.

In the evaluation of computational forensics results of the proposed IP user tracking forensics mechanism, as the Markov chain prediction model generates multiple forecast results, we use "predictability" to measure the performance of forecast results.
Predictability = number of chances of success × 100 (%)/ total number of predictions

iii. Uncertainty

Besides "accuracy rate", and "predictability" we use entropy to evaluate the "uncertainty" of data. Entropy is a random description, which describes all uncertain and disorderly conditions in a situation. Entropy is defined, as follows [25]: if there are n events in the range of random variable X, i.e. $X = \{x_1, x_2, \ldots, x_n\}$, and let the probability of occurrence of each event be $P = \{p(x_1), p(x_2), \ldots, p(x_n)\}$, the information content of No. i event is $-\log_b p(x_i)$, and b is the logarithmic base. The entropy value $H(X)$ of random variable X is defined as: $H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$.

Meanwhile, in terms of the characteristics of entropy, we can use the following characteristics for further judgment.

I. Entropy is greater than or equal to zero, i.e. $H(X) \geq 0$.
II. If N is the total number of events in random variables, $H(X) \leq \log_2 N$. If, and only if, the event probabilities are equal, $H(X)$ has the maximum value. Herein, $\log_2 N$ is the maximum entropy.

The "uncertainty" is used to evaluate the degree of disorder of a historical path. The historical path is the data source of prediction. If the chances of occurrence of various base stations in the data are almost equal to each other, meaning the data are disorderly, they are more difficult to predict. The ratio of maximum entropy to entropy can help us evaluate the forecast result of a current historical path.
Uncertainty = entropy of historical data × 100 (%)/ maximum entropy of historical data

4 Empirical evaluation and discussion

4.1 IP user tracking forensic mechanism to existing framework

This study implemented the proposed IP user tracking forensic mechanism in four ISP providers providing xDSL Internet connections (Chung Hua, TFN, Sparq, SeedNet), one 3/3.5G Internet connection (FETnet), and four WiMax Internet connections (Global Mobile Corp., FETnet, Tatung, VeeTIME) WISP providers. This study selects an ADSL network and a 3G network for description.

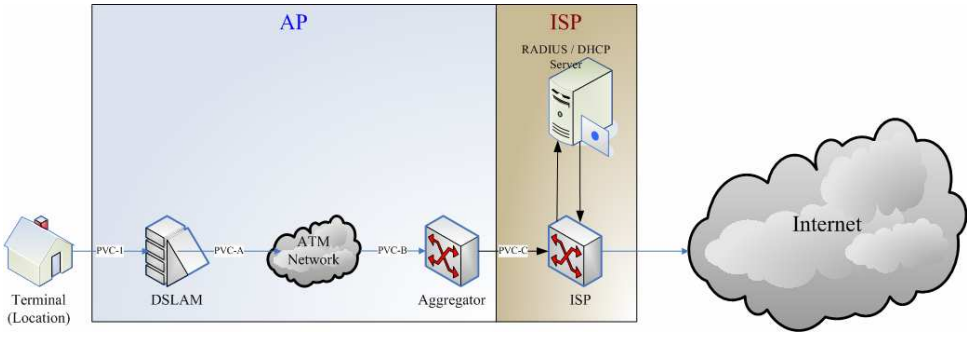

Fig. 8: Implementation of IP user tracking forensic mechanism in ADSL network

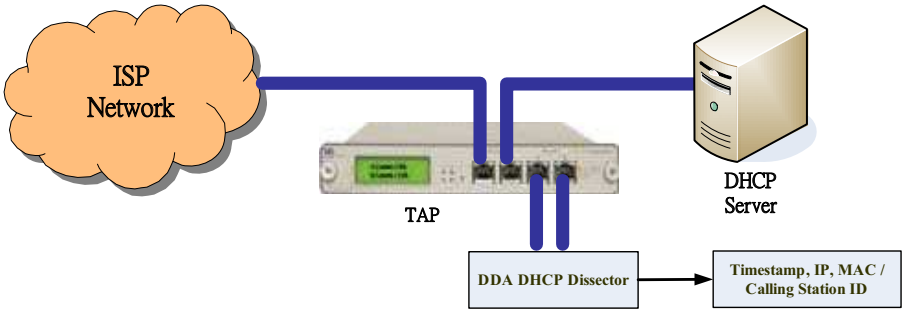

Fig. 9: Implementation of DDA in ADSL network-DHCP access architecture to obtain Access Information

4.1.1 Implementation in ADSL network

Generally, the ADSL Internet connection mostly uses PPPoE or DHCP to obtain dynamic or fixed dynamic IP, as shown in Figure 8.

This study takes an ADSL network environment DHCP access architecture as an example to describe how to obtain Access Information. As shown in Figure 9, in this architecture, when DDA obtains Access Information, e.g. IP mapping chaddr (client hardware address, i.e. MAC Address), when it is associated with the mapping of a MAC Address stored in the Termination Information, and the user's application address, the IP location is complete (Figure 2).

4.1.2 Implementation in 3G UMTS network

When a mobile phone connects to Internet via a GPRS network, the GGSN obtains dynamic IP Address mapping of the mobile phone from DHCP and a RADIUS Server. Therefore, if the DDA module is installed in DHCP and RADIUS Server, the IMSI and MSISDN of the mobile phone and the corresponding IP address information can be obtained. These data will be fed back by DDA via XML-based protocol to the IPLD module.

The GGSN and Charging Server also transmit CDR (Call Detail Records) information to each other; therefore, if a DDA module is installed between GGSN and Charging Server, the IMSI and MSISDN of the mobile phone and the corresponding IP address information can be obtained, and these data will be fed back by DDA, via the XML-based protocol, to the IPLD module.

When the IPLD carries out IP location, the actual position of the mobile phone can be obtained from GMLC, via the MLP (Mobile Location Protocol) interface, the IP location information chain corresponding to the IP and actual position can be concluded from the mobile phone IMSI and MSISDN and the corresponding IP address information, as shown in Figure 3.

4.2 The results evaluation and discussion

This section will use 10 actual Internet fraud cases provided by the High-Tech Criminal Center, Criminal Investigation Bureau of Taiwan, to evaluate the feasibility of the IP user tracking forensics mechanism, as proposed in this paper. The criminals in the cases used IP as means of communication to avoid investigation (e.g. E-mail, VoIP, and MSN) (Figure 11). We select 32 criminals whose locations (tracking monitoring) have been mastered by LEA for verifying the feasibility of objects' IP location, IP user individualization, and historical path reconstruction.

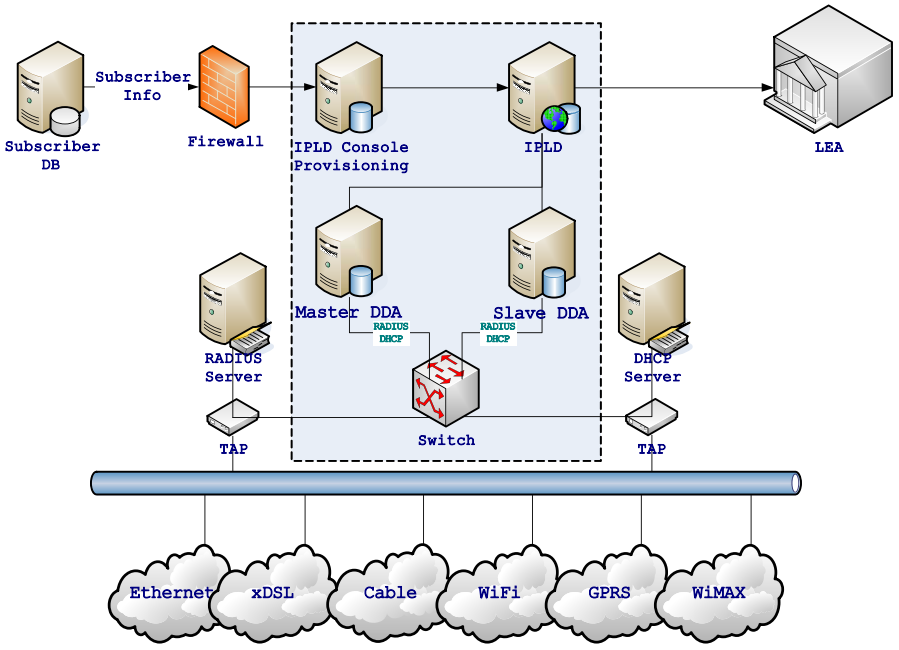

Fig. 10: Implementation of IP user tracking forensic mechanism in 3/3.5 network

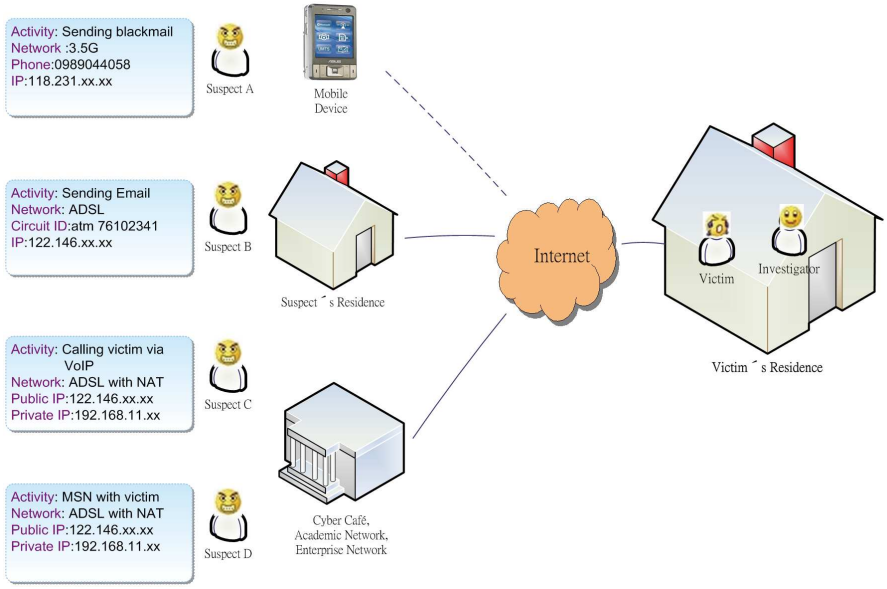

Fig. 11: Schematic diagram of IP communication dodging detection

4.2.1 IP individualization and location

About 122 IP Addresses obtained from actual cases are used to validate the location results of the IP user tracking forensic mechanism, and to match the actual location in actual crime. It is observed that 32 IPs connected to the Internet via ADSL, the IP location results can locate the actual position of an end IP device, and the "accuracy rate" is 100%. There are 15 IPs connecting to the Internet via WiMAX, and 75 connected to the Internet via a 3G/3.5G mobile network. The IP location results show the "accuracy rate" is 100%, the base station coverage when a criminal connects to the Internet can be located. The error range depends on the planned coverage of the base station. For example, the 3G/3.5G base station covers a radius of 150–500 m in an urban area, it is 1–5 km in the suburbs. The WiMAX base station covers a radius of about 500 m in an urban area (WiMAX is installed only in urban areas in Taiwan). The empirical results show that the IP user tracking forensics

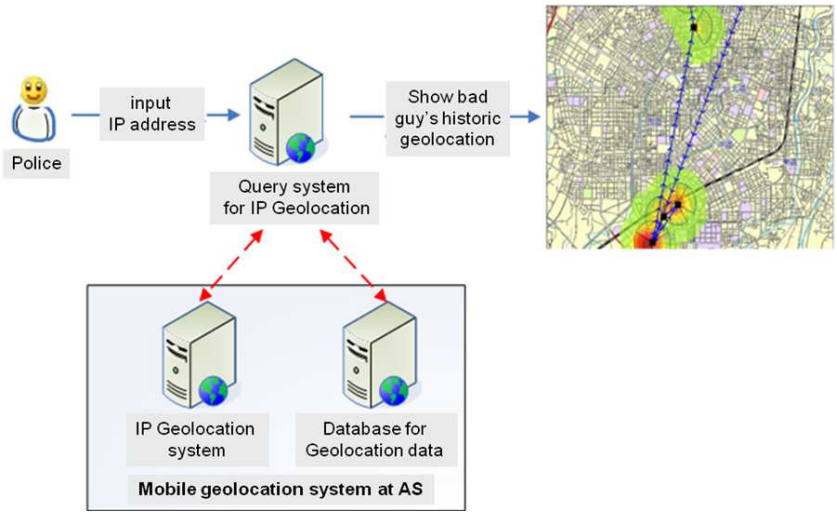

Fig. 12: IP location and Visualization.

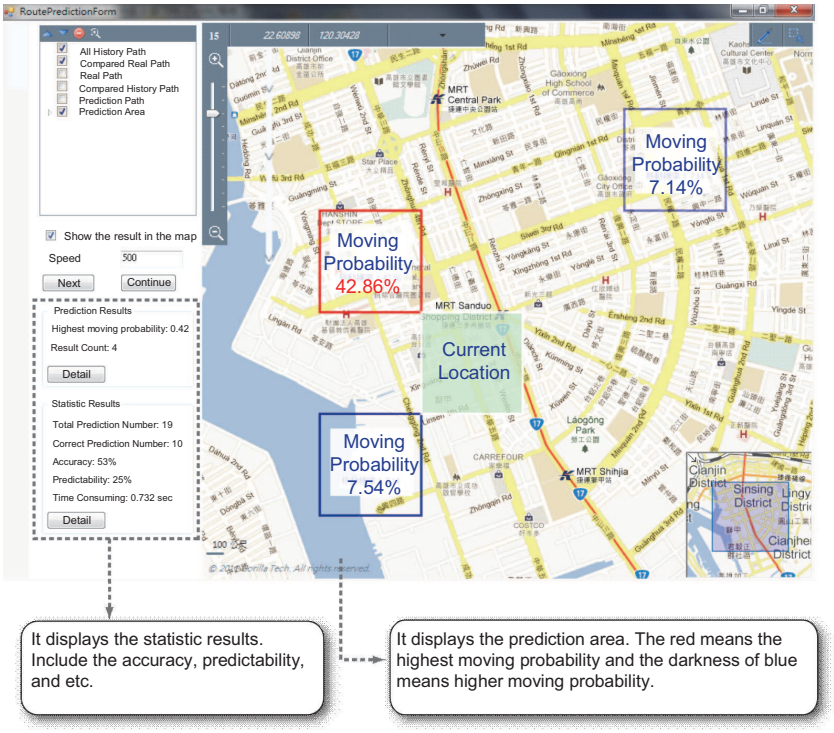

Fig. 13: IP user tracking forensics and Visualization.

mechanism can successfully conduct accurate location, and can incident out an IP user (criminal) (Figure 1, Figure 2, Figure 3, and Figure 4).

4.2.2 Computational forensics

The historical physical paths of the 122 IP Address users (i.e. cybercriminals) are successfully reconstructed using an IP user tracking forensic mechanism (this study adopts the CDR of last year), with 122 physical traces and approximately 19,854 base station records, as shown in Figure 12.

There were 30 breakpoints made in each trace of the 122 physical traces, and then the IP user tracking forensic mechanism was used for computational analysis. The results showed that the average "predictability" was 90.91%, and the accuracy rate was about 88.7%. The average entropy was 4.55, the average maximum entropy was 8.42, and the average uncertainty was 54.03%.

4.3 Discussion

The proposed IP user tracking forensic mechanism reasons out the spatial error of IP location. If the User connects to the Internet via a fixed network, e.g. xDSL, the IP location result can directly locate the actual position of end IP device (Figure 2). If they connect to the Internet via a mobile network (3G/3.5G, WiMAX), the base station coverage (Cell) when the IP user/cybercriminal connects Internet can be located (Figure 3 and Figure 4). The error range depends on the planned coverage of the base station. For example, the radius is 150–500 m in urban area, the radius is 1–5 km in the suburbs.

Secondly, the empirical results of computational analysis show that, when the average uncertainty of reconstructed trace data is 71.83%, the average "predictability" is 90.91%, meaning the proposed IP user tracking forensic mechanism has good forecast ability.

5 Conclusions and future works

The proposed IP user tracking forensics mechanism is based on the concepts of IP location and computational forensics, and develop forensics tracking on the Internet. In an environment of an integrated fixed network, mobile telecommunication network 3/3.5G, WiMAX, and next generation network IP Multimedia Subsystem (IMS), each related node in the network is confirmed, the DPI (Deep Packet Inspection) is in charge of accessing, copying, decoding, and saving necessary data retention for IP location, in order to analyze the geolocation of the IP Address associatively, and according to an IP Address, time, and the information record left from using IP services on the Internet, and "identity" can be individualized. The segmental paths (e.g. trace breakpoints) can be reconstructed by using an algorithm, in order to obtain complete path of the target.

In terms of contributions of this study, the IP location mechanism of the proposed IP user tracking forensic mechanism can be used for cybercrime investigation, and can be applied to targeted marketing, restricting digital content sales to authorized jurisdictions, another security applications, such as credit card fraud, and could serve as part of an E-911 system for voice over IP. The accuracy of IP location has been increased from the current ISP-level and city-level to a minimum 0 m error (Internet access via fixed network), the maximum error is Cell range (Internet access via mobile network), which is low cost without constructing reference nodes (landmarks). This mechanism also can individualize the IP user, as well as use an algorithm to reconstruct segmental paths (e.g. trace breakpoints) in order to obtain a complete path. The average success ratio (predictability) is 90.91%, thus, performance is perfect.

In terms of Implications of Practice, the proposed IP user tracking forensic mechanism does not need to modify the protocols of an existing network, redesign a new router, or set numerous reference points, as it can be directly applied to the existing network, and can provide excellent accuracy. The research findings can be used as reference for various countries to develop Internet forensics tracking.

In terms of the limitations of this study, the proposed IP user tracking forensics mechanism is only an applicable for domestic IP. In order to apply it to global Internet location, various countries should have the same mechanism, and a centralized IP2Location Database (IPLD) should be established, in order to meet the requirements for multinational IP user tracking forensics.

References

[1] Ma M.: Tabu marking scheme to speedup IP traceback. Computer Networks, **50**, 3536-3549 (2006).
[2] Gao Z., Ansari N.: A practical and robust inter-domain marking scheme for IP traceback. Computer Networks, **51**, 732-750 (2007).
[3] Lai G. H., Chen C. M., Jeng B. C., Chao W.: Ant-based IP traceback. Expert Systems with Applications, **34**, 3071-3080 (2008).
[4] Wang X. J., Wang X. Y.: Topology-assisted deterministic packet marking for IP traceback. The Journal of China Universities of Posts and Telecommunications, **17**, 116-121 (2010).
[5] Aljifri H., Smets M., Pons A.: IP traceback using header compression. Computers & Security, **22**, 136-151 (2003).
[6] Liu J., Lee Z. J., Chung Y. C.: Dynamic probabilistic packet marking for efficient IP traceback. Computer Networks, **51**,866-882 (2007).
[7] Hilgenstieler E., Duarte Jr. E. P., Glenn M. K., Shiratori N.: Extensions to the source path isolation engine for precise and efficient log-based IP traceback. Computers & Security, **29**, 383-392 (2010).
[8] Castelucio A., Tadeu A., Gomes A., Ziviani A., Salles R.M.: Intra-domain IP traceback using OSPF. Computer Communications, **35**, 554-564 (2012).
[9] Hsu H. M., Lin F. Y., Sun Y. S., Chen M. C.: A novel protocol design and collaborative forensics mechanism for VoIP services. Journal of Communications, **7**, 132-142 (2012).
[10] Li L., Shen S. B.: Packet track and traceback mechanism against denial of service attacks. The Journal of China Universities of Posts and Telecommunications, **15**, 51-58 (2008).
[11] Luo J., Wang X., Yang M.: An interval centroid based spread spectrum watermarking scheme for multi-flow traceback. Journal of Network and Computer Applications, **35**, 60-71 (2012).
[12] Kim Y., Helmy A.: CATCH: a protocol framework for cross-layer attacker traceback in mobile multi-hop networks. Ad Hoc Networks, **8**, 193-213 (2010).
[13] Durresi A., Paruchuri V., Barolli L.: Fast autonomous system traceback. Journal of Network and Computer Applications, **32**, 448-454 (2009).

[14] Katz-Bassett E., John J., Krishnamurthy A., Wetherall D., Anderson T., Chawathe Y.: Towards IP geolocation using delay and topology measurements. Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, 71-84 (2006).
[15] Muir J. A., Oorschot P. C.: Internet geolocation and evasion, Citeseer, (2006).
[16] Padmanabhan V., Subramanian L.: An investigation of geographic mapping techniques for internet hosts. Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications, 173-185 (2001).
[17] Gueye B., Ziviani A., Crovella M., Fdida S.: Constraint-based geolocation of internet hosts. IEEE/ACM Transactions on Networking, **14**, 1219-1232 (2006).
[18] Wong B., Stoyanov I., Sirer E.: Octant: a comprehensive framework for the geolocalization of internet hosts. Proceedings of the 4th USENIX conference on Networked systems design & implementation, 313-326 (2007).
[19] Li D., Chen J., Guo C., Liu Y., Zhang J., Zhang Z., Zhang Y.: IP-geolocation mapping for involving moderately-connected internet regions. Project participation from Microsoft Research, (2009).
[20] Sarangworld Traceroute Project. http://www.sarangworld.com/TRACEROUTE/.
[21] Vixie D. C., Goodwin P., Dickinson T.: A means for expressing location information in the domain name system. RFC 5280, IETF Network Working Group, (1996).
[22] Harrenstien K., Stahl M., Feinler E.: NICNAME/WHOIS. RFC 954, IETF Network Working Group, (1985).
[23] Dawson M.: The internet location services model. Computer Communications, **31**, 1104-1113 (2008).
[24] Masó J., Pomakis K., Julià N.: OpenGIS web map tile service implementation standard. Open Geospatial Consortium Inc., (2010).
[25] http://en.wikipedia.org/wiki/Entropy_information_theory.

Feng-Yu Lin received his Ph.D. degree from the National Chiao Tung University, Taiwan, Republic of China, in 2004. Currently, he is working towards the second Ph.D. degree in the Department of Information Management, National Taiwan University. His research interests include communication/network forensics, data mining, and information security.

Yeali S. Sun received her BS from the Computer Science and Information Engineering department of National Taiwan University in 1982, and MS and Ph.D. degrees in Computer Science from the University of California, Los Angeles in 1984 and 1988, respectively. From 1988 to 1993, she was with Bell Communications Research Inc. (Bellcore; now Telcordia). In August 1993, she jointed National Taiwan University and is currently a professor of the Department of Information Management. Her research interests are in the area of wireless networks, Quality of Service and pricing, Internet security and forensics, scalable resource management and business model in cloud services and performance modeling and evaluation.

Meng Chang Chen received his B.S. and M.S. degrees in Computer Science from National Chiao Tung University, Taiwan, in1979 and 1981, respectively, and the Ph.D. degree in Computer Science from the University of California, Los Angeles, in 1989. He was with AT&T Bell Labs from 1989 to1992. He is a Research Fellow of Institute of Information Science, Academia Sinica, Taiwan and have served as Deputy Director of the institute for 5 five years. His current research interests include wireless access network, QoS networking, computer and network security, information retrieval, and data and knowledge engineering.