The Carmen Sandiego Project

Don Bailey
donb@isecpartners.com

Nick DePetrillo
nick@depetrillo.net

July 4th, 2010

1. Definitions

MCC	Mobile Country Code
MNC	Mobile Network Code
ME	Mobile Equipment, such as a cellphone or a device such as the iPad
MSC	Mobile Switching Center
HLR	Home Location Register
CNAM	Caller Display Name (Caller ID)
HBI	High Business Impact
PII	Personally Identifiable Information
NPA	Numbering Plan Area (Area Code)
NXX	Area Code Exchange
MSISDN	Mobile Subscriber ISDN (NPA-NXX-NXXX)
IMSI	International Mobile Subscriber Identity
GSM	Global System for Mobile Communications

2. Executive Summary

This white paper purports to detail the methodologies and techniques for collecting intelligence on both individuals and organizations using novel methods of access and correlation of data from various publicly available sources. These sources focus on data collected by various telecommunication networks and databases, including global GSM networks and the caller ID system within the United States and Internationally.

Collecting the data is the easiest part. Making the data useable in its practicality, ensuring the information is verifiable, and correlating the data into actionable items, is a distinct challenge taking time, resources, and finances. This paper will cover techniques employed to process and analyze the data and will explain the results of this analysis. Mitigation strategies will be discussed that provide individuals and organizations with the means to detect and protect themselves from these threats.

3. Leveraging Resources

Prior to Tobias Engel's (Engel, 2008) research on the Home Location Register (HLR) leveraging information from the cellular network to track and profile subscribers was complex, difficult, and expensive. Security researchers were focused on attacking Mobile Equipment in order to obtain tracking information for specific subscribers. The exposure of the HLR by for-profit entities overseas opened new doors for researchers, providing them with a window into the telephone network. With access to this new resource researchers were able to retrieve subscriber specific data that not only described the current general physical location of a subscriber, but the information related to a subscriber's mobile network.

If an individual knows the phone number (MSISDN) of their target, HLR access can easily provide details pertinent to an attack. But, what if the MSISDN isn't known? Obtaining a cellular phone number for high profile targets is not always an easy task. While a phone number isn't specifically a private or protected value from an engineering perspective, it is often treated as such from a social perspective. If the MSISDN is the key to private subscriber data, obtaining it is an imperative for the attacker. One technique that has been proven to work well is the use of the caller ID service.

The caller ID service, a well known telephony resource, can be used to obtain billing name information for both landline and cellular MSISDN. Acquiring a specific individual's MSISDN can be accomplished by scanning all numbers within the region that the individual is presumed to live or work. Candidate matches are matches within the caller ID database that look similar to the name or organization of a target subscriber. Once potential candidate matches are obtained from the scanning process, a researcher can use other techniques to determine which of the candidate MSISDNs are owned by the target subscriber.

With the MSISDN and HLR, attackers may implement further attacks on the privacy of a subscriber's Personally Identifiable Information (PII) or High Business Impact (HBI) data. By using the HLR to associate a subscriber with a specific cellular provider, an attacker may tune their attacks to those known to work for that particular provider. This allows the attacker to increase the potential for success while minimizing the probability of exposure.

Though the tuning of attacks to specific providers is interesting from an academic perspective, those attacks will not be covered in this document as they are out of scope with respect to the Carmen Sandiego project.

In the following sections, methods for obtaining location data and identity data using HLR and Caller ID will be described in detail.

4. Leveraging Location Data within the United States

Access to the Home Location Register will often yield information related to a user's location. This information is obtained in the form of an opaque series of digits known as the Mobile Switching Center (MSC). Mobile Switching Centers are comprised of the infrastructure for a particular physical location and provide routing, payment, and other services for subscribers within that physical boundary. The MSC generally corresponds with a general physical location with varying granularity. Some MSCs have been found to be as large as cities and their suburbs. In other cases, multiple MSCs comprise a single city.

Despite its association with a physical location, known MSCs within the United States do not reveal the corresponding location. Some MSC values found overseas, particularly in and around Germany, encode the zip code into the MSC. At this time, there is no known encoding for MSCs within the United States. Also, it is currently believed that there is no public database that translates an MSC into its corresponding geographic bounding box. Therefore, it is necessary for a researcher to reverse engineer the MSCs into a particular location in order to use these values for tracking purposes. There are several steps that must be taken to accomplish this goal.

Finding Candidate MSISDN

Building an MSC to geographic location association database requires a large sampling of MSISDN for a particular provider. Since MSC codes are provider specific, each provider must be mapped within a given geographic location. Thus, the first step of the process is to isolate a set of mobile providers within a geographic region. The method for doing this, however, is out of scope of this document. Once obtained, all area code and exchange pairs (NPANXX) for the providers found within the geographic region must be catalogued (NANPA, 2010). For each block of MSISDN within the NPANXX ranges, random HLR queries should be made. Thus, if an NPANXX pair is (877)555, valid MSISDN within this range would be (877)555-0000 through (877)555-9999.

The result of the HLR query must validate two pieces of information. First, the HLR data must properly resolve, indicating that a valid subscriber owns Mobile Equipment associated with the MSISDN. Second, the Mobile Country Code (MCC) and Mobile Network Code (MNC) tuple must equate to the provider's tuple. For example, if HLR data *D* is retrieved for a MSISDN *M, D.MCCMNC* must be representative of the provider that owns *M.NPANXX*. Codes that match any other provider may be catalogued for later use, but must not be included in the current findings.

Correlating an MSC to a Location

Once a large set of MSISDN, typically in the order of several thousand, are found to be valid with respect to the above requirements, observations pertaining to location can be made. HLR queries for this set of MSISDN should be made at least three times of day. These times must correspond with societal norms oriented to location. In other words, the researcher must query the HLR database when it is likely that the largest set of individuals is home, at the office, and back home again for the evening. Of course, this pattern is recognizable as the typical nine to five work-day adhered to by most Americans.

By sampling HLR data at these times, one may infer a correlation between the resulting MSC code and the presumed location of the MSISDN. Over a period of several days, a pattern can be seen that distinguishes one MSC code as the probable code for the general location of the subscribers. Since the NPANXX corresponds with a known location, an association can be made between one particular MSC code and a section of a city, a city, or a city and its surrounding area.

Once this process is performed on cities throughout the United States, clear associations between MSC codes and physical locations can be made. Now, when a particular subscriber's MSISDN travels from one location to another, the researcher can make accurate predictions as to the path of travel and the destination based solely on observation of the changing MSC.

Building Physical Bounding Boxes

Building associations between MSCs and general geographic locations is important, but it provides high level data describing a subscriber's location. The observer cannot make inferences regarding the location of a subscriber within an opaque MSC. However, there are strategies that will provide insight into the geographic boundary of these MSCs.

The Caller ID database, described later in this document, associates a billing record name with a particular MSISDN. This resource is well known in the telephony world and is typically associated with "landline" systems. However, the Caller ID database also contains mobile subscriber data. By leveraging the information contained within the Caller ID database, bounding boxes (or polygons) can be built that purport to describe the physical boundary of an MSC. The ability to approximately define a physical MSC boundary enables a researcher to make clearer observations as to the location of a particular subscriber.

Building physical boundary information can be done using whitepages.com (White Pages, 2010). Whitepages.com provides a web API that allows searching records within the White Pages database. This record search can be restricted to a particular state and city and augmented with a name. Using the Caller ID record for a particular MSISDN, a query can be built that attempts to locate potential matches within a city. Records that match these parameters will be returned in an XML list. Iterating through the list, matches that contain a physical location can be isolated. If a physical location is provided, the whitepages.com API will also include geographic coordinates (latitude and longitude) describing the location. These match candidates can be stored in a database, associating them to one or more MSISDN.

After amassing a large amount of potential matches, a bounding polygon can be built. Using the geographic coordinates, a polygon can be auto-generated using Google Maps KML file format. This file format, described on Google's website, will allow the overlay of a polygon on top of specific geographic coordinates. MSISDN with associated physical locations can be pulled from an internal database and rendered in the KML file. A boundary for the MSC can be generated by detecting geographic locations on the outer edge of the map. This outer edge, or bounding polygon, describes the approximate physical boundaries of the MSC. By overlaying the bounding polygon on top of a geographic location in a visualization platform such as Google Maps or Google Earth, further location information may be inferred.

Using Bounding Box Data

By defining an approximate physical bounding polygon for an MSC, means of ingress and egress through the particular MSC can be pinpointed. Major expressways, airports, rivers, or ports, can be detected and described as travel mechanisms that may affect the quality of a cellular signal. Subscribers traveling in watercraft or aircraft are more likely to be out of service and render within the telephone network as an 'Absent Subscriber'. If an HLR query detects an 'Absent Subscriber' condition, and the subscriber has not been detected moving into an adjacent MSC, one can hypothesize that the subscriber is traveling via aircraft or watercraft if that subscriber does not have a habit of powering off their Mobile Equipment.

Observations may also be made as to the likelihood of a subscriber's location within a physical boundary. By viewing the polygon describing an MSC, it can be determined whether the area is primarily residential or primarily business in nature. If one or the other, a researcher may be able to make strong assumptions related to the subscriber's function within an MSC during specific hours. Observations such as this can help differentiate between a subscriber working the night shift in an industrial area and a subscriber working the day shift.

Overlaying Cell Tower Data

One last technique can be used to augment physical location data within an MSC. The OpenCellID (OpenCellID, 2010) database is a project that aims to catalogue cell towers worldwide. Each tower is documented along with its geographic coordinates, and which network provider (MCCMNC) operates the tower. By overlaying the geographic coordinates for towers specific to the provider that owns a particular MSC, observations may be made with respect to which tower a subscriber is likely to be associated with during travel. One cell tower may be more likely to govern subscribers traveling over a main interstate, while another tower may govern subscribers arriving via watercraft. Assumptions can be made with respect to the signal strength of the cell towers relative to their physical location. This knowledge augments the ability for the attacker to leverage more technical attacks, such as IMSI catching.

Summary

Location data begins with the MSISDN, is defined by the MSC, and is augmented by Caller ID data paired with White Pages results and OpenCellID information. Though MSC values are opaque within the United States, the techniques described above can not only define a general physical location, but may define approximate physical boundaries of that general location. The information can be leveraged to monitor potential travel as well as ascertain what specific cellular towers are likely to be used by the subscriber. Together, these items provide a powerful window into the daily behavior and travel of a subscriber.

5. Leveraging Identity Information

As described in the last section, creating a database of caller ID records for each phone number (MSISDN) within a region is imperative when attempting to construct physical bounding boxes for Mobile Switching Centers. There are several challenges related to this process, as described here.

Primarily, in order to access the caller ID database, you must enable the caller ID lookup capability (CNAM) on your connection to your provider. With most modern Voice over IP (VoIP) systems, this is an exceptionally easy task and usually costs a fraction of a cent per lookup. The low per-lookup rate enables research as the overall cost is negligible. Modern software PBX systems, such as Asterisk, enable the CNAM capability by default. Typically, accessing the data on a per-call basis only entails reading a variable during the processing of an extension script.

Old Tricks, New Tricks

The most prevalent problem with the use of caller ID information is that in order to request the caller ID database for a particular MSISDN/caller-name record, the PBX system initiating the request must in the process of accepting an incoming call from the target MSISDN. Essentially, an incoming phone call from the target is required to access the database. There is a workaround, however. The ability to spoof the phone number originating an outbound call is a widespread issue within VoIP networks. This functionality may be leveraged to trick the network into querying caller ID databases without the need for the actual owner of the target MSISDN to call the researcher's PBX system.

A PBX user may make originate an outgoing call with a forged MSISDN, impersonating the target of the caller ID request. However, the PBX calls a phone number that it has been programmed to accept calls for. In other words, it spoofs a phone call to itself. When the PBX acknowledges an incoming phone call from the target MSISDN, the caller ID data is queried. Once the data has been retrieved from the caller ID database, the call is dropped without being answered and the caller ID data is stored in a database with the associated MSISDN.

This simple technique allows for the mass retrieval of caller ID entries for a negligible cost. Because the call is never answered, the call is never billed, leaving the resulting charges associated with the call at simply the cost to initiate a call, acknowledge an incoming call, and the cost of the caller ID request. Thousands of requests can be issued for caller ID data at the cost of tens of dollars.

At the time of writing, the authors believe that this technique stays within the legal boundaries with respect to initiating outbound calls with a forged MSISDN. While this technique does leverage the ability to spoof an MSISDN, there is no attempt to deceive. The subscriber receiving the spoofed call is aware of the forgery because the subscriber initiated it and as a result, there can be no deception involved. Monetarily, there is no potential for fraud since there is no money exchanged except the cost to initiate the call, acknowledge the call (if any), and to initiate the caller ID request. These billable functions are not occurring or changing due to the introduction of a forged MSISDN. These billable functions would occur no matter what MSISDN was used to complete this process. As a result, no fraud can be established.

Finding Nemo

Once the caller ID information has been retrieved for an entire region, the researcher may be able to scan for entries that look like a specific individual. Caller ID records are typically one of four types, a private individual name, an organization name, a privatized record, or an unallocated record. Each entry typically holds up to eighteen characters.

Records containing names of private individuals are usually easily recognizable. They are stored in common formats such as “Firstname Lastname”, “Lastname, Firstname”, and “FirstnameLastname”. Despite the wide variation in formatting, records containing a particular surname or partial name can be easily identified through SQL queries. Organization names are typically the name of the business, government branch, etc. These are also readily identifiable records and are simple to search for. Privatized records vary by provider but typically contain a blanket name, such as “WIRELESS CALLER”. While these records are sometimes set for subscribers that do not wish to have their billable name published, they are more often the result of a phone number ported from one wireless carrier to another where the previous carrier did not maintain caller ID entries.

Lastly, records associated with unallocated MSISDN typically resolve as “Unavailable”. However, it should be noted that while “Unavailable” typically denotes an unallocated record, some providers do not store caller ID database entries. These providers may have active MSISDN that resolve as “Unavailable”. If a provider does define records, however, an “Unavailable” record can typically be assumed unallocated.

Schools of Fish

The same string of characters recurring across multiple MSISDN caller ID entries may be presumed to be tied to the same account. The advantage to this information is that various types of groups can be detected. For example, business organizations typically issue Mobile Equipment to management, executives, consultants, and individuals whose job description requires them to be on-call. These accounts can be detected by searching for the business name in caller ID records.

Depending on how the organization is structured, MSISDN may be issued by locale. Executives may be given ME with MSISDN that contain an NPANXX specific to the corporate office. For example, if the corporate office is in San Francisco, CA, executives may have MSISDN whose NPANXX are “818744”, while engineers working in a data center in New Jersey may be issued MSISDN whose NPANXX are “973997”. This gives researchers an advantage when attempting to classify MSISDN or when tailoring attacks to specific types of employees during a penetration test or a risk assessment.

6. Potential Threats

Organizations and individuals are both subject to the same general classes of threats. Interception of data, exposure of privacy, location tracking, and ME abuse are all pertinent to both classes of targets. These potential threats can be grouped into two separate classes: remote attacks and local attacks. Remote attacks do not require proximity to the intended target and are performed using a computer and resources available on the internet. Local attacks do require proximity and are performed using a sophisticated piece of hardware, called an IMSI catcher, possibly in combination with remote attacks.

Remote attacks can come in several forms. An attacker may leverage location data to watch a particular target traveling from city to city, or state to state, exposing privacy related data and potentially inviting more nefarious behavior. The ability to ascertain the provider for a particular subscriber allows the attacker to leverage provider specific attacks, decreasing the likelihood that an attack will be detected by even a vigilant and knowledgeable victim. This may lead to a loss of privacy or an exposure of PII or HBI data.

Associations may be made between a particular individual and other individuals in the same organization through the caller ID database, allowing an attacker to remotely track and monitor the behavior of all individuals of a particular group and distinguish when one individual has been separated from the rest. Remote attacks are primarily passive in nature, but may lead to more aggressive attacks depending on the length of time tracking information has been monitored and depending on the data exposed through provider specific attacks (Bailey, 2010).

Local attacks can be used in combination with remote attacks to fulfill more sinister objectives. IMSI catchers are capable of forcibly dropping encryption used by ME on GSM networks, exposing potentially sensitive conversations to eavesdropping and manipulation (Meyer, 2004). IMSI catchers can also be used to track a subscriber's location (Meyer, 2004) with a fair amount of granularity using signal strength and vectoring techniques. Since the signal of an IMSI catchers can span multiple kilometers, or approximately a mile (Strobel, 2007), remote attacks can be used to find an individual within a generally large locale and IMSI catchers can be used to pinpoint a target within that locale with much more accuracy.

Finally, manipulation and potential abuse of the ME itself can be performed (Grugq, 2010). IMSI catchers potentially control not only the data and voice channels of a mobile device, but the control channel as well. The control channel can be fuzzed or abused to implement device specific attacks against firmware or software operating the ME. While this area of research is still new, there is a potential for code execution attacks and other device specific attacks against the ME operating system. Abuse of these tactics may provide an attacker with access to not only the voice channel of a particular target, but all the information stored on the phone. If a rootkit can be deployed, a victim's exposure may be long term.

7. Recommendations

The ability to leverage location and identity centric data is key toward the subversion of security controls that protect both personally identifiable information and high business impact data. Identity centric data provides a window into the typically private lives of individuals and organizations by exposing resources that potentially store PII and HBI to unknown threats. Location centric data provides a window into the behavior of individuals, their typical routines, and deviations to those routines, exposing both the individual and associates to potential threats.

Organizations and individuals alike can defend from these threats, but it may depend on changes in their behavior and infrastructure. Organizations must determine whether these threats are significant to their business model and should incorporate these potential issues into the company threat model. Individuals should determine how relevant location data and caller ID data is to their daily lives. While some companies may not consider location tracking a threat, defense contractors, news agencies, and government branches may. Individuals that don't often travel out of their greater city area may not feel compelled to change their behavior or network provider. However, those individuals that frequently travel or work long distances from home may need to reassess their security and privacy requirements.

Both organizations and individuals can take steps to protect themselves. Research can be done to determine which providers reveal location data and which ones don't. While all providers whom have had discussions with the authors are vigilant and proactive in defending the security of their subscribers, some controls are not yet in place on certain networks. While this is likely to change in the near future, some individuals and organizations may determine that their security needs are more immediate.

Caller ID information changes from provider to provider. Some providers allow the restriction of the data and will mask caller ID entries, if requested. However, others will not. Some providers don't support caller ID for mobile networks and will not expose caller ID information at all. Individuals and organizations must determine their security requirements, perform the research, and choose a provider according to their needs.

Overall, the security of the mobile subscriber is placed solely within the subscriber's hand. The international telephone network is vast, complex, and of varying robustness. Security cannot be guaranteed on all parts of the network. Rather, security must be a choice made by the individual who is knowledgeable of the potential threats and mindful of the potential risks. Threats to the network must be monitored and mitigated by the provider. Threats to the individual or organization must be monitored and mitigated by that individual or organization.

8. Appendix A – Bibliography

Bailey, D. (2010). We Found Carmen Sandiego. *SOURCE Boston* .

Engel, T. (2008). Locating Mobile Phones using Signalling System #7. *CCC 25c3* .

Grugq, T. (2010). Base Jumping: Attacking GSM Base Station Systems and Mobile Phone Base Bands . *Blackhat Briefings, Las Vegas* .

Meyer, U. (2004). On the Impact of GSM Encryption and Man-In-The-Middle Attacks on the Security of Interoperating GSM/UMTS Networks. 5-6.

NANPA. (2010, 07 01). *Number Resources - NPA (Area Codes)*. Retrieved 07 01, 2010, from NANPA: http://www.nanpa.com/area_codes/

OpenCellID. (2010, 07 01). *www.opencellid.org*. Retrieved 07 01, 2010, from www.opencellid.org: http://www.opencellid.org/

Strobel, D. (2007). IMSI Catcher. 13-15.

White Pages. (2010, 07 01). *White Pages Web Developer Portal*. Retrieved 07 01, 2010, from developer.whitepages.com: http://developer.whitepages.com/