Computer Location Determination Through Geoparsing and Geocoding of Extracted Features

Chad M.S. Steel, *Member, IEEE*

***Abstract*—This paper compares the extracted feature data from a sample set of hard drive images in an effort to relate the features to the physical location of the drive. A list of probable zip codes, phone numbers, place names, and IP addresses are extracted from raw drive images and compared to manually identified geolocation data. The results of the individual extractions are then analyzed to determine the feasibility in using automated extraction and analysis techniques for geolocating hard drives.**

***Index Terms*— Hard Disk, Forensics, Geocoding**

I. Introduction

This paper compares extracted feature information to manually identified physical drive location from a series of hard drive images to evaluate the ability of different features to predict the physical location of that drive. The determination of the geographic locations of interest on a hard drive can be used to track the travel of a drive, identify locations associated with the drive's primary users, and find locations of interest to the users of the drive. Through the extraction of key location features in an automated fashion from hard drive images, we are able to provide a probable primary location for the computer in which the drive was located with varying degrees of accuracy.

Initially, each drive image is manually reviewed to identify its primary location, followed by an automated analysis of each drive. The first automated step is the use of feature extraction to pull out information of interest, followed by the geocoding of that information. The geocoded information is then analyzed for patterns that would uniquely identify drive location. A comparison is then made between the extracted features to determine the feasibility of using each feature type for geolocation.

II. Related Work

Forensic feature extraction was used by Garfinkel to extract large amounts of string data from drives using regular expressions. Garfinkel's work focused on privacy-based data and financial information, but the concept of reduction through string pre-processing is useful and not specific to the featured he extracted[1].

Exploitation of IP address (and hostname information) was done successfully by Buyukokkten and McCurley on a local level. They utilized *whois* lookups to build a database of IP address location information which they then applied to a set of web pages[2, 3]

Relation of the geodata through contextual parsing was shown as effective by Li, who successfully used context data to perform disambiguation, e.g. Springfield MO v. Springfield, NJ[4].

Mapping of key place names has been successfully done using the geolocation lists from the extractions of US Census Data in the 2000 Gazetteer as performed by US Government[5]. Place name, area code, zip code, and latitude/longitude have been correlated in the GeoLite database[6].

IP address information has been successfully mapped by IP2Location[7] as well as GeoLite[6]. Additionally, information on network address translation was presented in the original proposal for removal of non-public IP addressing[8].

III. DataSet

A set of thirty six hard drive images was used as the initial dataset for the research. The drives were all purchased on eBay and contain varying amounts of user data which is used for geographic feature extraction. Each drive has been manually verified to have at least one partition with data present to eliminate "wiped" drives. The drive partitions each have at least one FAT or NTFS partition.

The drives images used range in size from 300MB to 40GB. The drives are converted into raw disk images using dd, and stored as image files on a drive array. Searching the drives is done at a physical level (as opposed to logical) using command-line tools in a Windows XP environment. All of the drives were parallel ATA (PATA) technology, 3.5" drives. The majority appeared to have come from home computers, though a few were clearly used for business storage.

IV. Methodology

A. Overview

Each of the drive images obtained was imaged and then a series of feature extractions and validations was performed in an automated fashion. Simultaneously, a manual review of each image file was performed to provide a check value for the extracted geographic information. The overall methodology is shown in Figure 1.

For the automated analysis, an initial feature extraction first extracts raw strings then uses grep-based regular expressions to parse out values of interest. Then a validation routine is run on each extracted feature to remove unwanted artifacts and compare the data with known-valid geographic values. Finally, the individual values for each image file are examined to find patterns indicative of geographic location.

For the manual review, each image file is loaded into a forensic tool and reviewed manually for indicators of its original physical location. Physical location names, email origination points, and IP addresses are used to identify a likely origin for comparison with the automated results.

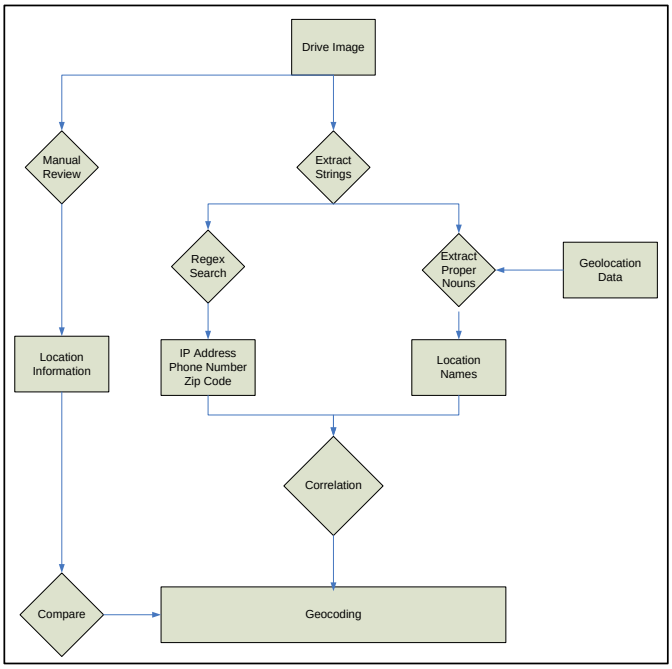

Fig 1. Methodology for drive image geolocation extraction

B. Initial Feature Extraction

For each drive in the experimental corpus, a feature extractor is run. The feature extraction uses the same approach as that used by Garfinkel[1], but with a different set of extraction expressions:

1. Initially, text strings of size four or greater are extracted and stored as intermediary text files to speed the actual processing. The average ratio of image size to extracted text was 9.24 to 1, allowing for an almost tenfold increase in followup query speed.

2. A series of regular expressions are used to extract text which matches the following feature profiles:
   A. **Zip Codes**. Any 5 digit number which is not embedded in a longer string of numbers and/or letters is extracted.
   B. Phone Numbers. Numbers fitting the format (xxx)xxx-xxxx, with or without parentheses and dashes are extracted.
   C. IP Addresses. Any series of numbers w,x,y,z between 0 and 255 that fits in the pattern w.x.y.z is extracted.

3. A set of proper nouns is extracted from the text files for geographic lookups. These are extracted by finding strings which begin with a capital letter and contain at least four characters (to reduce the noise created by smaller, randomly occurring strings.) Strings which start a sentence are then removed. This has the potential to remove actual place names, as in the sentence “Springfield is the greatest town on earth,” but their removal greatly reduces the number of false positives (proper nouns that don’t relate to place names.)

C. Feature Validation

Following the initial feature extraction, secondary validation on the remaining values is performed and the validated values are loaded into a database. The following individual validations were performed:

1. **Zip Codes.** No feature validation was performed on zip codes. The zip codes were linked to specific location codes from [6].
2. **Phone Numbers.** The individual area codes associated with the phone numbers were extracted for geographic region information. These area codes were compared to valid area codes from [9] and those were linked to specific location codes from [6].
3. **IP Addresses.** Each IP address was validated to remove any quads with leading zeros (e.g. 02.03.04.05) and any reserved use addresses[10] were discarded. IP address geolocation was obtained from [6] to find location codes for each IP address.
4. **Proper Nouns.** All of the proper nouns were processed for stopword removal [11] and any very long words (greater than twenty characters) were removed. The remaining words were compared to [5] to obtain location information.

After validating each of the features, histograms of each feature on a per-image basis were made and a cross-image analysis of each was performed to identify commonalities (which would likely be unsuitable for drive location identification if used.)

D. Manual Location Identification

Simultaneous with the feature extraction, each of the images was loaded into the AccessData Forensic Toolkit (FTK), an analysis tool used in digital investigations. For each of the drive images, the files present were indexed and the following information used to make a likely determination of the computer location using a manual analysis:

1. Time zone/clock settings.
2. IP address settings (if it did not use Network Address Translation)
3. Email message origination locations.
4. Locations mentioned in resumes, address books, and other logical documents.

The most likely value from the above analysis is stored and compared to the automated analyses to determine a distance deviation.

V. Results

The initial results appear promising for area code and IP address extraction, but zip code and proper noun extraction show excessive noise. Additional techniques for improving each of the four extraction types are presented in the Future Work section which follows.

The raw results and analysis for the four data types used are detailed below.

A. Manual Review

A manual review of the drives was performed using AccessData's Forensic Toolkit. The imaged drives were originally analyzed for time zone/clock settings, IP address settings, email message origins, and addresses listed in text.

Of the drives analyzed manually, 36%, or 13 drives, could not be accurately geolocated with a simple manual analysis. Of these drives, the following were determed to be the reasons for not being able to accurately geolocate the drives manually:

- 2 drives were found to be storage drives for business data on internal servers (and were not "personal" drives).
- 3 drives had no "fresh" installations of an OS and had their drives wiped clean. The "fresh" installations were of older operating systems with no location data provided.
- 5 of the drives had no local network or Internet connectivity and contained no personal correspondence.
- 3 drives had multiple phone books but no discernable patterns in them to identify location.

The origin of purchase for the drives was originally thought to be a good indicator of geolocation, but further review showed inconsistent correlation between the purchase location and the actual location of use for the drives that could be identified.

As an outcome of the automated analysis, one of the best determinants for manual drive location identification turned out to be the phone number settings for dial-up service providers (like America Online), which are set to local numbers for cost reasons. In addition to Windows dial-up settings, error logs and dialing logs were good sources for this data. Time zone settings were too vague to be of direct use.

IP address settings were useful in drives that did not use private IP addresses, but due to the age of the drives (the average age was 5 years old) many of them came from systems that pre-dated the home networking explosion that arrived with inexpensive broadband. The same lack of connectivity effected the manual identification using email message originations (when no email messages were present).

Finally, the use of locations mentioned in logical documents turned out to be a double-edged sword. The proliferation of large address lists obscured the ability to identify unique locations in three cases, and the presence of computer-generated phone books altered the results for some of the automated techniques below. The results of the manual review are shown in Figure 2.

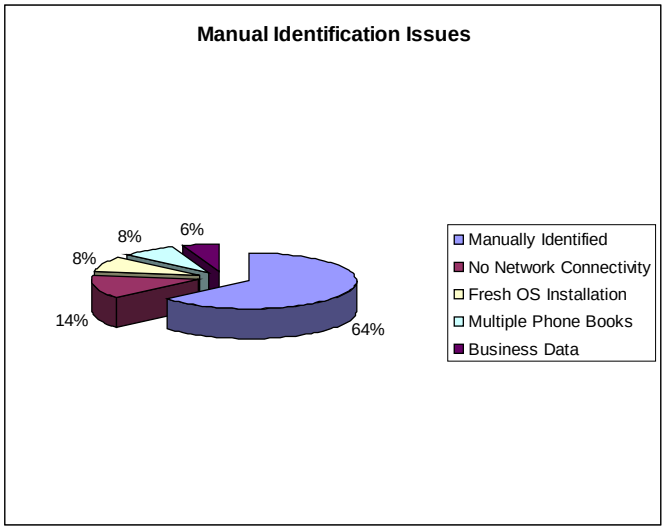

Fig 2. Manual review results

B. Zip Code

The zip code extraction was performed by searching for strings of numbers as noted above. Each of these numbers was initially assumed to be a zip code for analysis purposes. Assuming evenly distributed, random data (and an ASCII character set), approximately four percent of characters would be numeric. Given that, a five digit string of numbers such as a zip code should appear approximately 91 times per gigabyte. Constrain the same string by the rules used above - the preceding character is a space, comma or period and the following character is a space, period, or dash – and the rate goes down to approximately one occurrence per 100

gigabytes.

The drive data showed a significantly greater rate of occurrence for zip code-like strings than random data. Specifically, a mean occurrence rate of 14,515 per gigabyte with a standard deviation of 10,864 was found. This rate appeared to be promising for the extraction of data, but significant signal to noise ratio problems were identified. The normalized frequencies for the top fifty occurring zip codes are shown in Figure 3. As seen, there is an exponential decay in the rates of occurrence. To eliminate the influence of the highest ranked value, new frequencies are calculated for the remaining values and an exponential decay is still evident as seen in Figure 4.

The top occurring strings meeting the zip code criteria are likely not zip codes. Table 1 shows the top 10 zip code values and their number of occurrences. As seen, the highest ranked zip code is the number 00000, which does not map to an actual map address. Similarly, the remainder of the top ten zip code matches contain other false positives. The number 65537 is a frequent stop number used by programmers ($2^{16} + 1$), and the remaining numbers are all modem frequency pre-sets.

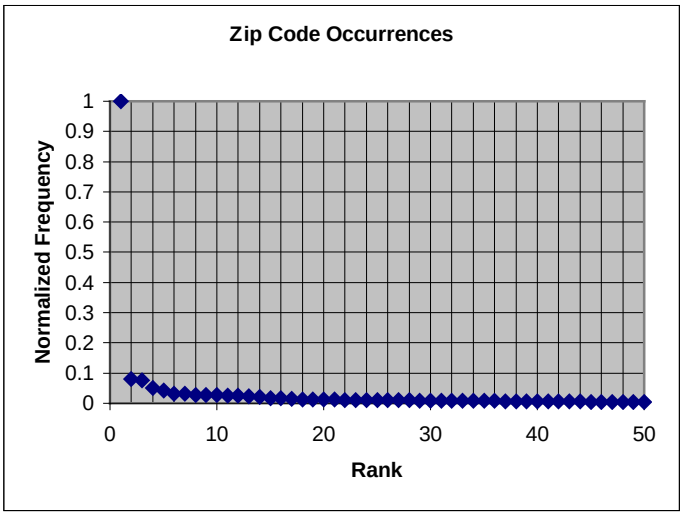

Fig 3. Zip code occurrence frequency by rank

Even if the top n zip codes are removed, the remaining zip code data is filled with noise. For the sample set of drives used, over thirty three thousand distinct zip code number matches were identified with no clustering in the distribution – without discernable clusters even after noise reduction their value for geolocation is poor.

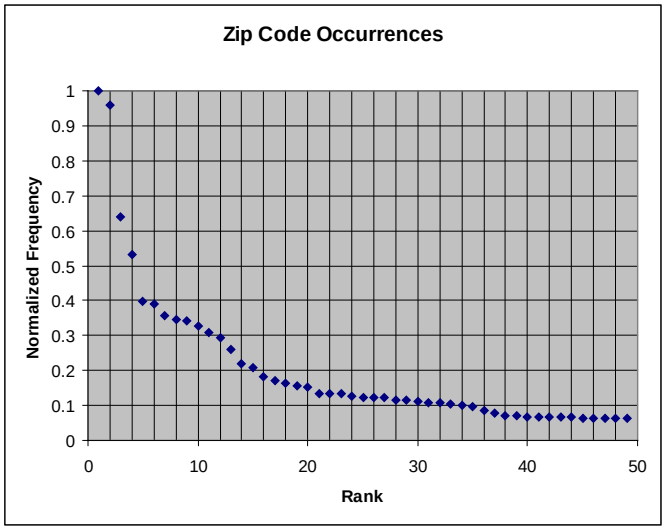

Fig 4. Zip code occurrence frequency by rank with largest removed

Zip Code Match	Number
00000	209941
65537	16806
14400	16144
28800	10751
99999	8911
19200	6697
12000	6575
33600	5972
21600	5804
16800	5726

Table 1. Common Zip Code matches

C. Phone Numbers

The use of phone number formatted strings is highly unlikely for non-phone number use, and the odds of random occurrence are negligible given the size of the drives in the dataset. Because this, the string match for a phone number is more precise than that of a zip code. As such, phone numbers provide a more likely candidate for geolocation of hard drives.

The occurrence rate of phone numbers on the drive images in the test set was 534 per gigabyte. This provides a large enough sample set for geolocation. Additionally, the non-valid area codes are easy to eliminate – a simple listing of valid area codes can be used. Doing area code validation removed 32 percent of the initial area code values identified, a significantly lower percentage than found with zip codes. Additionally, most of the removed area codes appeared to be part of sample phone numbers, with (000)000-0000 and similar numbers appearing frequently in the removed numbers.

An analysis of area code distribution showed the potential for easy post-processing. Specifically, area code 800 (which has no geolocation value) appeared in all of the drives examined, and area code 206, which is the area code for Seattle, Washington appeared in all of the drives examined with Windows installed. By removing the non-location area

code phone numbers such as 800, 888, and 877, and removing those with no discrimination ability, which are only area code 206 numbers in this case – all others have document frequencies below .75 – the remaining numbers can be used to geolocate the drives.

After the area code cleansing above, 61% of the drives were able to be identified by the primary area code extracted – confirmed as those that were directly related to the area code determined by manual analysis (or a same-location geographic overlay area code). The percentage identified is much greater than random, and with enhancement provides a good candidate for geolocation of hard drives. The drives which could not be manually geolocated were not included in the above, but are shown in Figure 5 below.

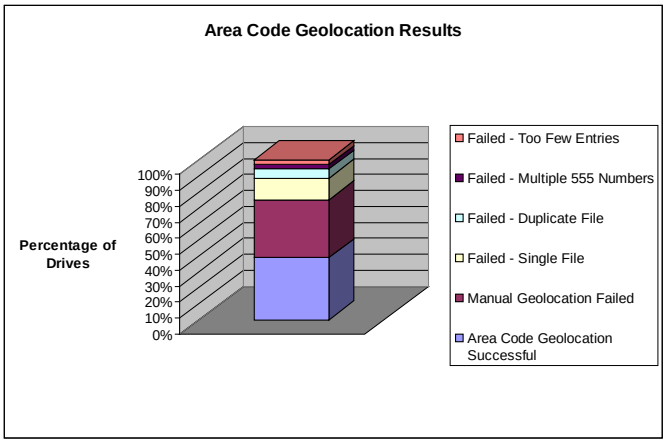

Fig 5. Results of Area Code Automatic Geolocation

For the drives which were not identifiable directly by area code, there were no "near miss" numbers. A near miss would be defined as an adjacent area code or one with a nearby geographic proximity, an example would be 212 and 973 – a New York City area code and a northern New Jersey area code. All of the drives which were misidentified were due to non-geographic reasons.

The most common reason for drive misidentification was the presence of a large number of phone numbers in a single file. The files included dial-up number lists for large Internet Service Providers in two of the cases (one for MSN and one for IOHK), two electronic phone books, and one file which contained a large amount of sample data that had a different area code. These problems could be eliminated by application of a weighting function based on area codes identified and the files in which they are present. A simple alteration in the calculations to apply a weight based on the total number of phone numbers in the file would have eliminated the non-sample data. The sample file could have been eliminated through the removal of positive hash hits using the NIST National Software Reference Library (NSRL) hashes or a similar hash set[12].

Other failures in area code-based geolocation were identified that can easily be handled in future work. Two images had a single file duplicated multiple times (and in both cases it was a known-hash file as above). One drive had a file that contained a large number of sample phone numbers in the XXX-555-XXXX format. Removing invalid "555" numbers would eliminate this situation. Finally, one drive had too few phone numbers to form an accurate identification profile (twenty total phone numbers were found).

D. IP Addresses

IP addresses were identified using a two-step process. The initial regular expression used identified four strings of one, two, or three numbers separated by periods. A post-processing step using PERL confirmed the values in the strings were valid for IP addresses and further sorted the IP's into public and private addresses. The two-step process was implemented to avoid using an overly-complex regular expression. Less than 2% of the numbers identified using the simple regular expression were found to be invalid IP addresses, and those that were identified tended to be version numbers for application, for example 1.0.0.601.

Using IP addresses to geolocate a drive relies on two factors – users accessing local IP addresses more frequently (DNS servers, dial-up addresses, routing information) than distant IP's, and users having access to public IP addresses. The separation in the initial identification process above concluded that 48% of the valid IP addresses extracted were private and/or reserved IP addresses and thus not suited for geolocation. With the large number of systems still using dial-up from the sample set, this implies many dial-up providers made use of a private address space for their modem pools.

Of the images examined, only 9% of those drives manually identified were able to be matched via IP address. While greater than random, the identification percentages were still low. An analysis of the misses showed no geographic correlation.

The miss analysis identified 28% of all IP address matches resolved to Microsoft, skewing the results for Washington-located area codes. Removing the Microsoft IP addresses, the number of matches rises to 28%. Of those that did not match, the most common reason appears to be drives with little or no connectivity. Without a network connection, the only IP addresses present are those hardcoded into the operating system. A listing of the IP addresses found on more than 75% of the drives is shown in Table 2 below.

The second most common reason for failure was a large number of IP address hits on the same address space not being grouped (as they were different IP's) – a more effective algorithm would find the most common area code, weighted by the number of occurring IP's on a per-block basis. A listing of the uncorrected reasons for failure is shown in Figure 6 below.

The results of IP address geolocation are less promising than that of the area code analysis, but further work with a more broadband-centric sample set may yield more useful results.

IP Address	Occurrence Frequency
102.54.94.97	1.00
38.25.63.10	1.00
102.54.94.102	0.97
102.54.94.123	0.97
11.11.12.13	0.97
102.54.94.117	0.97
101.2.1.1	0.90
157.54.23.41	0.87
198.105.232.1	0.84
198.105.232.6	0.77

Table 2. IP Addresses Appearing in Multiple Drives

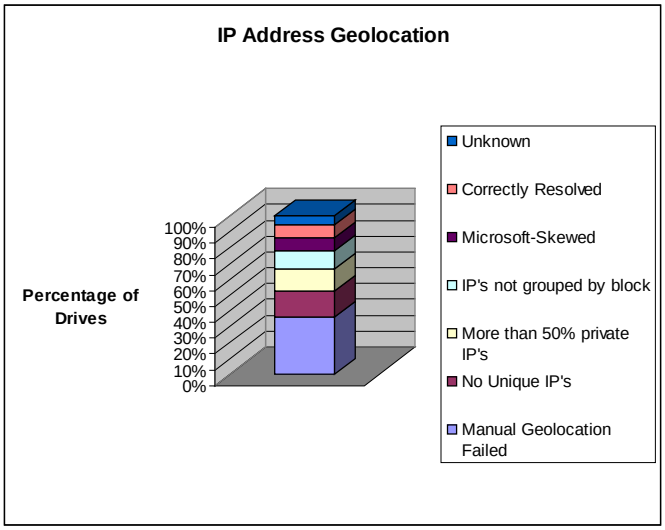

Figure 6. IP Address Geolocation Analysis

E. Proper Nouns

The use of a commercial geotagger like MetaCarta's product[13] to extract geographic place names was not feasible given the large size of the dataset (and would be computationally infeasible even on a modest array of a few terabytes.) Instead, a simpler extraction of proper nouns was used to determine the feasibility of a more advanced approach.

The secondary extraction of proper nouns from the string data yielded too many results. To reduce the resultant data, any proper nouns found at the beginning of sentences, which were not likely to be geographic place names, were eliminated. The remaining proper nouns were extracted and further reduction performed.

A tertiary reduction was performed on the proper nouns extracted to further reduce the size of the data. Smaller words, those that were three characters or less, were removed. Additionally, words larger than twenty characters were removed. Twenty six percent of the proper nouns extracted were under four characters in size, but a negligible amount (less than a tenth of a percent) was over twenty characters. After reduction, approximately eighty one million words remained.

Of the proper nouns identified, approximately two million unique words were found. There were a large number of common words identified – as shown in Table 3, the most common words appear to be programming related. The distribution of proper nouns appears to be Zipfian as shown in Figure 7. An additional complication with proper nouns is their discrimination power – over nine hundred individual proper nouns appeared in every disk image.

Proper Noun Match	Number
Responses	1680807
Name	1224924
File	1185960
Microsoft	959698
Stub	805560
String	798901
Type	585145
Object	577144
Windows	563196
System	520506

Table 3. Common Proper Noun matches

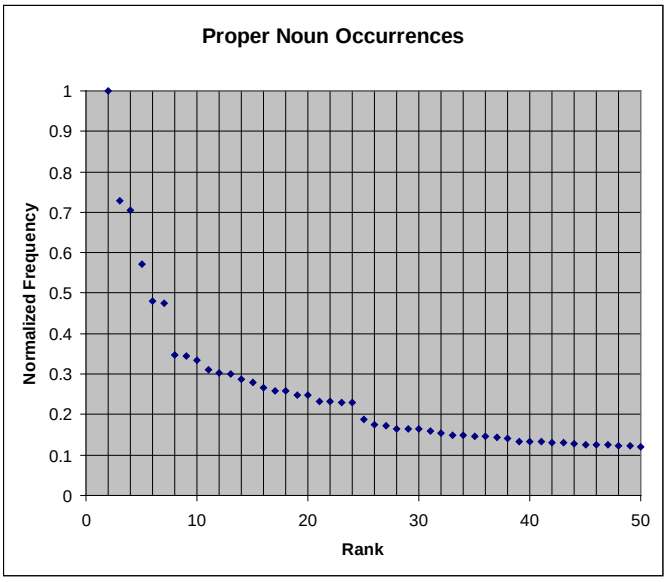

Figure 7. Proper Noun Occurrences

Of the proper nouns initially identified, the majority did not appear to have geographic place information. Once the non-geographic names were removed through comparison to a geographic dictionary, the remaining terms were evaluated for their geolocation value. As can be seen in Table 4 below, none of the top ten matches yield significant place information and are false positives. The majority of the names correlate with coding terms too strongly to be accurate place names, and none match the manual analysis place name findings.

Given the large number of occurrences of false-positive place name nouns, the expected signal-to-noise ratio would be too low to be useful. If the number of place nouns were of the same magnitude as that found in phone numbers, the expected ratio would be below .01, and direct mapping would not be

possible. In addition to the false positives based on unusual place names, there were many false positives based on creative naming within Windows. Names like Verdana and Vista exist as operating system names as well as place names, and other names like Redmond appear too frequently based on their inclusion in file comments.

Location Noun Match	Number
Media	174724
Port	117221
Main	52709
Post	47683
Dial	34107
Normal	24203
Front	23845
White	20730
Trust	19790
City	19627

Table 4. Location noun matches

VI. Discussion

The use of geoparsing and geocoding to obtain the geographic location of a hard drive is potentially feasible for both area codes and IP addresses with some refinement. The use of IP addresses, while showing poorer performance on the older dataset, has a higher potential for systems with more frequent Internet connections. The use of zip codes and proper nouns with geographic significance were unfruitful and unlikely to yield positive results, even with substantial refinement to the algorithms.

As a side finding, the results indicate the value of using forensic-specific stopwords when indexing hard drives. In addition to common stopwords (the, he, and, it, etc.) computer specific stopwords like name, file, Microsoft, and string are so prevalent that returning files containing these results are unlikely to be fruitful.

In addition to the proper noun stopwords, similar stopword-like data can be gleamed from the other data types. Removing common phone numbers (like those included in DLL's) and common IP addresses (those hardcoded into sample files and private/reserved IP addresses) may reduce the amount of data a forensic examiner needs to analyze by a significant amount.

One assumption made in the analysis was the computers would have a single geographic location associated with their use. Because the initial dataset used desktop hard drive images this assumption is more valid than it would be if laptop drive images were used, but even desktop drives can be used in multiple locations. People move to other cities, bring desktops to college, and sell machines to others in different locations, which can significantly confuse the analysis.

Although the intent was to identify drives in an automated fashion, the automated extraction data provides a feedback mechanism that can be used in a manual analysis as well. By identifying "interesting" phone numbers and IP addresses in an automated fashion, the forensic analyst can be provided with additional search terms for manual review.

VII. Future Work

The initial examinations used drives acquired in (and assumed to be used in) the United States. Generalizing the regular expressions to be international in nature would be needed for global use (though the IP address space used was global), as would the use of a global dataset that geolocated international phone numbers.

The drives analyzed in this paper were parsed on a drive-level (as opposed to a file-level) of abstraction. The parsing of individual files using their logical file structure would allow the targeting of specific file types, elimination of duplicate files, and a general improvement in data quality traded off for more complexity in the parsing algorithms. The use of a file-based approach would additionally solve some of the problems associated with common sample files. Known file filters like the Forensic Toolkit KFF and NIST NSRL hash sets could significantly reduce the number of false positive results.

Another limitation of the initial analysis was the dataset – all of the drives used were desktop-size drives. The use of laptop drives may present further difficulties associated with the multiple use locations expected. The primary use location would still be expected to dominate, but unusual usage patterns may be present in certain circumstances like long distance commuting.

The use of more advanced algorithms to identify patterns and cross-analysis of phone and IP address information may be useful as well. Though beyond the scope of this paper, looking at each drive individually and applying outlier detection techniques may yield better overall results.

The application of term frequency-inverse document frequency techniques to assist manual review could be useful as well. Instead of treating individual files as documents, each individual drive image could be treated as a document and new evidence drives added to the corpus. This would provide a benefit of identifying items of interest in the evidence drive that are not as prevalent in the corpus as a whole.

Finally, if the data could be culled to a smaller sample size through representative sampling or similar techniques then geotagging may yield more valuable results that are generated from place names present.

VIII. Conclusions

The goal of this research was to test the feasibility of different techniques in accurately geolocating a computer. The research was successful in identifying two techniques that would be appropriate for geolocation – phone number extraction and IP address extraction. Additionally, the

information gathered identified two other techniques as infeasible – the use of zip codes and the use of potential geographic place names.

As an additional outcome of the research, stopword lists that can be used for future information visualization efforts were generated. These will allow for more enhanced manual review efforts when applied to traditional techniques.

IX. References

Chad M.S. Steel (M'95) Chad holds Bachelors and Masters Degrees in Computer Engineering from Villanova University and is currently pursuing a PhD in Computer Science at Virginia Polytechnic Institute.

He has served as the Chief Security Officer and Managing Director of a Fortune 100 corporation, been the Head of IT Investigations at a Global 100 corporation, and taught Computer Forensics at Penn State Great Valley.

[1] S. Garfinkel, "Forensic feature extraction and cross-drive analysis," *Digital Investigation*, vol. 3S, pp. S71-S81, 2006.

[2] K. S. McCurley, "Geospatial mapping and navigation of the web," presented at Proceedings of the 10th international conference on World Wide Web, 2001.

[3] J. C. Orkut Buyukokkten, Hector Garcia-Molina, Luis Gravano, Narayanan Shivakumar, "Exploiting geographical location information of web pages," presented at Proceedings of Workshop on Web Databases (WebDB'99) held in conjunction with ACM SIGMOD'99, 1999.

[4] H. Li, R. K. Srihari, C. Niu, and W. Li, "Location normalization for information extraction" presented at Proceedings of the 19th international conference on Computational Linguistics, 2002.

[5] U. S. Government, "2000 Gazetteer," U.S. Census Bureau, 2000.

[6] MaxMind, "GeoLite IP-City Database," MaxMind Inc., 2006.

[7] IP2Location, "IP2Location™ IP-Country Database," IP2Location Inc., 2006.

[8] P. F. Tsuchiya, "Extending the IP internet through address reuse " *SIGCOMM Comput. Commun. Rev.* , pp. 16-33, 1993.

[9] L. Madison, "LincMad's Telephone Area Codes and Splits," 2006.

[10] IANA, "RFC 3330: Special-Use IPv4 Addresses," Network Working Group 2002.

[11] C. Fox, "A stop list for general text," *SIGIR FORUM*, vol. 24, pp. 19-21, 1989.

[12] NIST (19 November 2006), "National Software Reference Library," [Online], Available: http://www.nsrl.nist.gov/Downloads.htm

[13] MetaCarta (15 November 2006), "MetaCarta's Technology and Products : A Public Sector White Paper," [Online], Available: http://www.metacarta.com/docs/Public_Sector_White_Paper.pdf