What Where Wi:
An analysis of information leaked by millions of wireless access points

Kipp Jones
kippster@cc.gatech.edu

Abstract

Wireless networks have sprung up by the millions throughout the world as consumers and businesses move towards a mobile, networked world. These access points (APs) are installed and managed by individuals and businesses and are unregulated, allowing anybody to install and operate one of these radio devices. This has allowed literally millions of these APs to become available and 'visible' to any interested party who happens to be within range of the radio waves that are emitted from the device.

There has been much written about the security risks [1][2][4][5][8] associated with wireless access points, and many of them operate in a 'unsecured' mode which does not require authentication to use. Research to date has focused on two primary topics: protecting access to these wireless networks and maintaining the privacy of the users of these networks. But there is another potential for information leakage that is accessible even for those not interested in directly accessing the network. These wireless access points emit a certain amount of information that can be combined with the physical location and potentially used for other purposes.

This proposed research will explore the possibilities for using this information to discover patterns, analyze behavior, explore naming and location information that can be discerned by gathering information about these access points over time. The goal is to explore the "What, Where, and Why" of the WiFi access points and their information.

Motivation

We have obtained the rights to analyze information regarding over 3.25 million 802.11 wireless access points. This data has been gathered over approximately one year and corresponds to the systematic scanning in some 75 cities throughout the United States as shown in Figure 1. This scanning process has produced the correlation of each access point information with its GPS location and AP signal strength information.

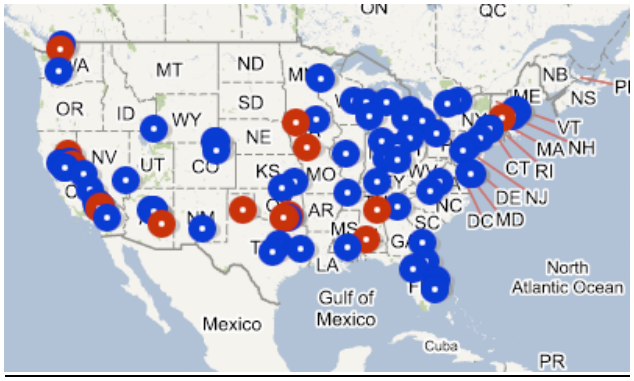

Figure 1. Skyhook Wireless coverage areas. Cities in red are in progress while blue cities have been completed.

The dataset (as detailed in Table 1) is comprised of GPS and WiFi access point logs obtained by Skyhook Wireless. Skyhook has granted us the ability to use this data to conduct research (with basic protections for their data being a requirement). We have access to both the raw data as well as the processed data indicating the location of these access points as calculated.

In addition, the data includes information that indicates the amount of motion that these APs experience over time, whether due to calculation error or due to physical movement of the access points. Information related to each access point includes the Service Set Identifier (SSID), the MAC address, the geographic location by longitude and latitude, and the dates the AP was first and last scanned.

Table Name	*Description*	*Number of Records*
CentralAP	Contains unique access points and their calculated geographical location	3,252,883
ChangeAP	Contains location adjustments to APs over time	2,957,034
RawScanningLogs	Contains the AP scan records from drivers	817,838,373
ScannerGpsLogs	Contains original GPS logs from drivers	760,369,932

Table 1. Description of data available for direct analysis.

Beyond the fact that this data is available for analysis, there are several other motivating factors that include:

- The company that generated the data, Skyhook Wireless, is interested in learning more about the value of the data and ways to use the information to improve their service;
- There are potentially interesting privacy and/or security implications, especially in the naming of access points that should be explored;

- The fact that these access points are associated with location information could be used to infer additional 3rd party details by means of data fusion;
- Companies such as FON [3] are reliant on the installation of their software on these access points and having sufficient coverage to provide their service, analyzing the manufacturer per region may yield good marketing, engineering requirements;

In short, the increase in number of access points is not going to stop. Nor is the gathering of information about these access points. This project intends to explore what value (good or bad) the information that is leaked by these devices can provide.

Approach and Project Plan

The general approach will be to identify items of interest and then create programs to analyze, calculate, and potentially map the results using an interactive process. In essence, this project will create a 'Mashup' [7] of wireless positioning data that will be mapped using the Google Maps API.

Project Outline

The following are the general steps that will be followed to accomplish this goal:

1. Obtain and validate data and rights to data
2. Install and load DB on a local system
3. Perform initial hand analysis to identify priority items (see below for candidates and initial order)
4. Create mapping Mashup visualization
5. Create interactive analysis interface
6. Evaluate the resulting information to discern items of value

Key elements of this outline will be detailed in the following section.

Project Schedule

Milestone	*Description*
March 1	Preparation
March 13	Pre-analysis
March 27	Interactive Interface
April 12	Initial Results
April 24	Final Project Delivery

Table 2. Key project milestones.

Preparation

Preparation consists of obtaining final rights for the data, installing and loading the database, installing and configuring a web server. These tasks are scheduled to be completed by the end of February.

Pre-analysis

During the pre-analysis phase, the data will be examined further and candidates for analysis will be prioritized. This step will require some manual processing of the data and will serve as the prototyping of the interactive system.

Some options that will be analyzed further to determine the potential value and possibilities for competition include:

Mine the data for interesting information:

- Manufacturer based on MAC address [9]
- Location information in SSID
- Naming characteristics of access points
- Default SSID naming practices (see Table 3 below)
- Movement of access points over time
- Speed of drivers during scanning
- Access point density ranking

Combine data with other sources:

- AP density correlated with household/regional income
- AP density correlated with zones or cities (see Figure 2 for an example city coverage map).

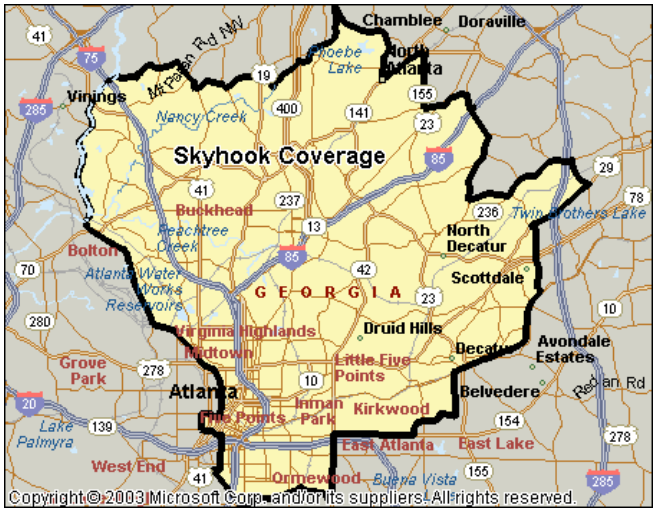

Figure 2. Skyhook Wireless Atlanta coverage map.

SSID	*Manufacturer*
Tsunami	Cisco
101	3Com
RoamAbout Default Network Name	Lucent/Cabletron
Compaq	Compaq
WLAN	Addtron
intel	Intel
linksys	Linksys
Wireless	Unknown

Table 3. Sample default SSID names by manufacturer.

Analysis & Visualization of Results

This phase of the project will extend the prototype to allow interactive querying of the data and the ability to create map and explore the results of the analysis and data mining. Several simple examples are given below using general access point location information. Note that some results may be simply statistical results such as those depicted in Table 4.

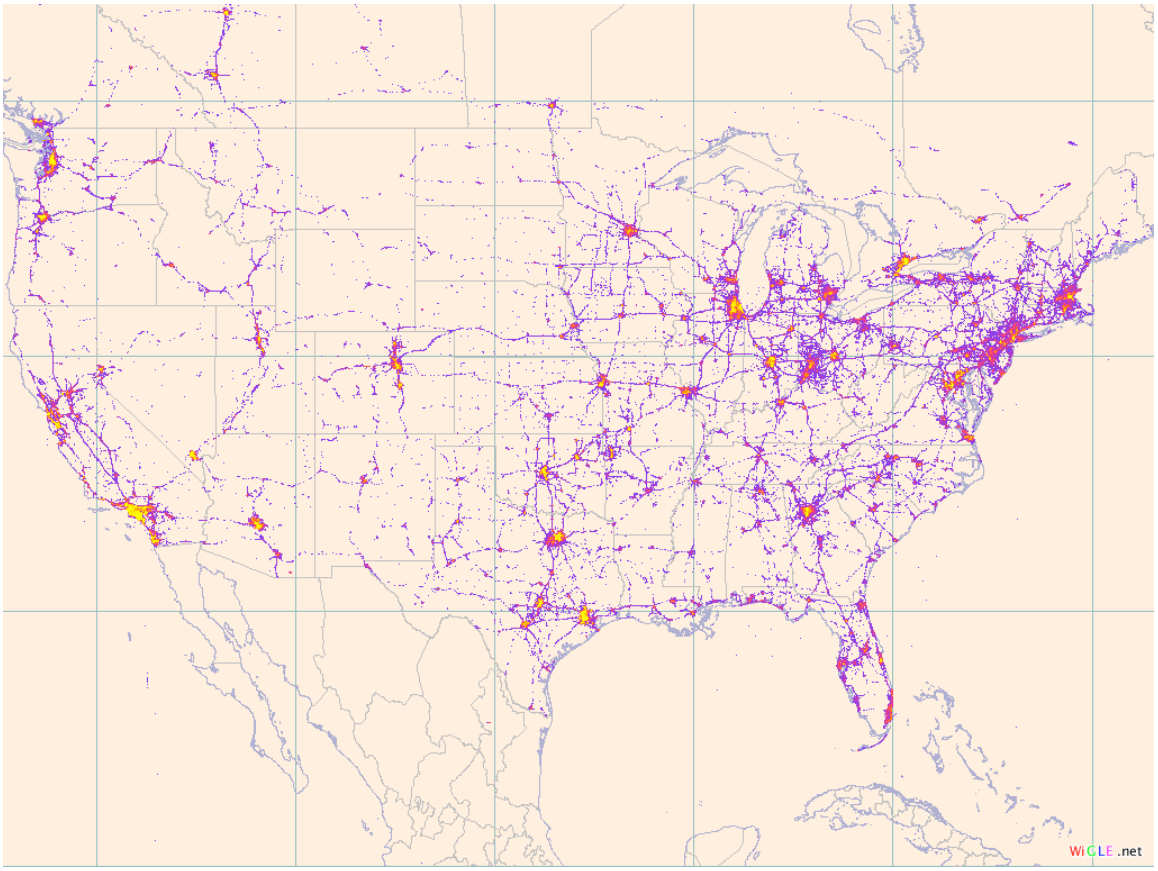

Figure 3. Wigle.net map of access points.

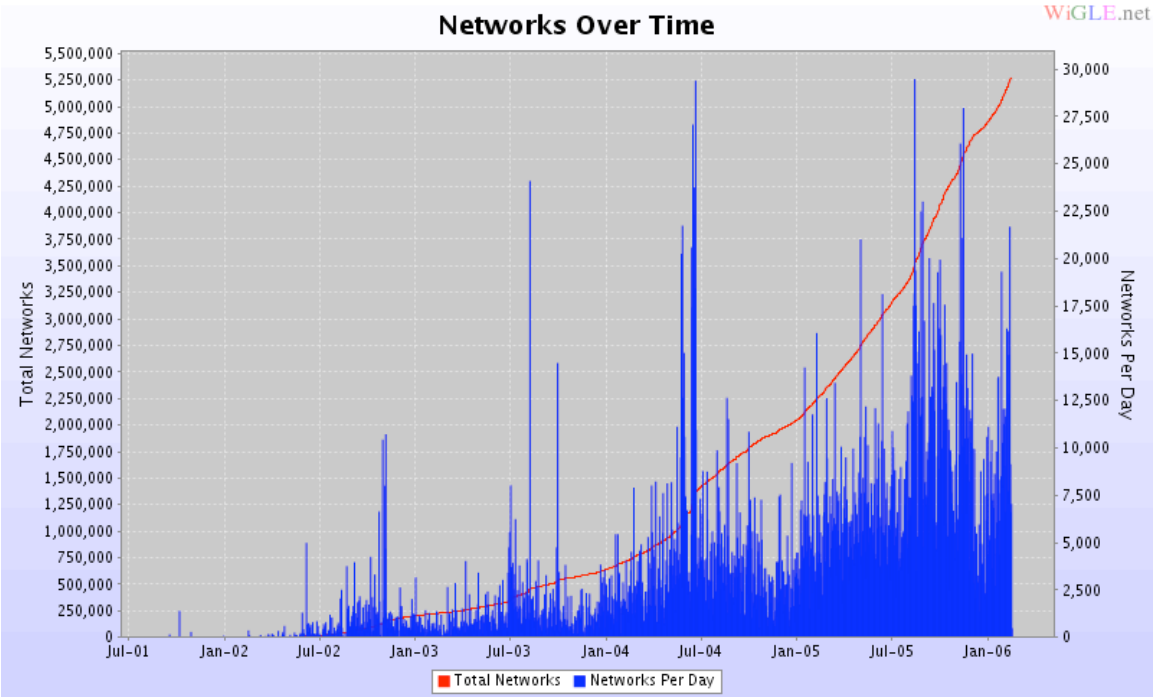

Figure 4. Wigle.net growth of mapped access points over time.

Description	*Count*
Unique networks in DB	5,542,267
Unique networks with location	5,259,353
Unique locations in DB	229,798,635
Networks with WEP	1,968,313 (35.5%)
Networks without WEP	2,590,332 (46.7%)
Networks WEP unknown	983,622 (17.7%)
Networks with default SSID	1,380,424 (24.9%)

Table 4. Statistics from Wigle.net.

Evaluation and Testing

The success of this project will depend on several key accomplishments. First, the creation of the data will create value in and of itself for further research. Second, the success of the analysis will depend on the 'interest' value of the findings where interest may be from a privacy perspective, a commercial perspective, or a purely intellectual perspective.

References

[1] Battiti R, Lo Cigno R, Sabel M, et al. Wireless LANs: From WarChalking to open access networks MOBILE NETWORKS & APPLICATIONS 10 (3): 275-287 JUN 2005

[2] Christopher W. Klaus (2002). Internet Security Systems Wireless LAN Security FAQ. http://www.iss.net/wireless/

[3] FON. http://en.fon.com

[4] Gruteser M, Grunwald D A methodological assessment of location privacy risks in wireless hotspot networks LECTURE NOTES IN COMPUTER SCIENCE 2802: 10-24 2004

[5] Intel's PlaceLab. http://www.placelab.com

[6] Mishra A, Petroni NL, Arbaugh WA, et al. Security issues in IEEE 802.11 wireless local area networks: a survey WIRELESS COMMUNICATIONS & MOBILE COMPUTING 4 (8): 821-833 DEC 2004

[7] Mashup definition. http://en.wikipedia.org/wiki/Mashup_%28web_application_hybrid%29

[8] Nikita Borisov, Ian Goldberg, and David Wagner. Intercepting mobile communications: The insecurity of 802.11. In Proceedings of MOBICOM 2001, 2001. http://citeseer.csail.mit.edu/article/borisov01intercepting.html

[9] Organizationally Unique Identifier (OUI) listing. http://standards.ieee.org/regauth/oui/oui.txt

[10] Skyhook Wireless. http://www.skyhookwireless.com

[11] Wireless Geographic Logging Engine – WiGLE. https://wigle.net/