A Real Time Network State Analysis Framework using Machine Learning Tools

Charles Meyers

Applied Mathematics Department

New York City College of Technology

**Back Ground Information**

*Signal Processing*

To even begin how to model a signal at a given point of space and time, we have to start with the model of the signal itself. All signals can be modelled as the sum of a series of cosines and sines. In the simplest case, for a continuous variable *E(t)*, we can model field strength as



Where A is the amplitude at the transmitter, c is the center frequency and phi is the phase shift. As we can see from the above equation, that only three parameters are within our control: amplitude, frequency, and phase. With these three parameters, there are many modulation schemes. The simplest case is show below in Figure 1.



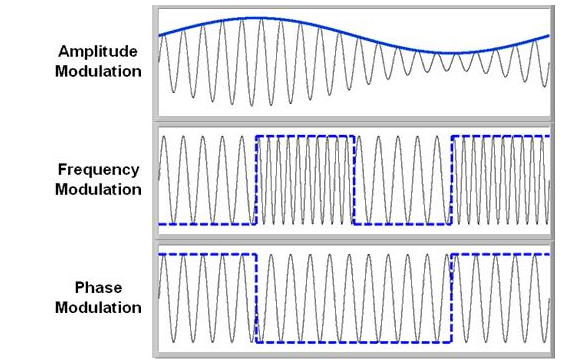
where *a* is the average amplitude of the signal strength at the receiver, given by the root-mean square estimation.

Figure 1

Modern wi-fi systems (since 802.11n) have deployed MIMO antenna arrays that allow for a particular type of modulation that exploits the phase and amplitude characteristics of signal, called Quadrature Amplitude Modulation (QAM). This is where we must recall Euler's Identity



If we decompose this according to a basic trigonometric identity and let x = 2\*pi\*f, we see that



Without loss of generality, the above identity holds true for any scaling factor $E$. If we rewrite this in terms of phase and quadrature components we find that



where Q is the phase of the signal and I is its amplitude.

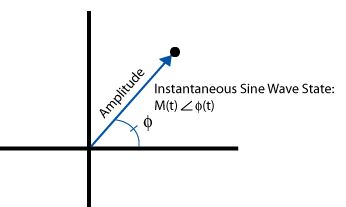
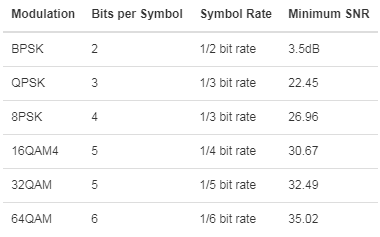


Figure 2

Furthermore, since phase is a continuous variable, our sample space (as a function of $\theta$ can become arbitrarily small. However, there is a trade-off here. As we decrease our $\theta$ sampling window, we increase our effective noise floor, as each pizza-shaped slice of our measurement window contains the same information in less geometric space. The effect of this is to transmit for bits per symbol. Note this only applies to receiving systems with a single antenna. [1] [2]

Table 1



**Design Constraints**

I want to be able to build this path loss model using a large set of numerical data. In order to do that, we must limit ourselves to measuring things that are available on common radio circuits. For model-building purposes, I set up tcpdump, iw, iperf, and ping to gather network state data from the transmitter and receiver ends. I have developed an android application using the Automate tool that gathers sensor data from all three on-board radios (bluetooth, wifi, and cdma) as well as the the 9 degree of freedom orientation chip that quantifies location (in 3 dimensions), orientation (along each of 3 axes), and magnetic field (in each of 3 orthogonal directions). In addition to that, we get other common wifi packet data. In addition, the wi-fi packet frames provide more information:

* a frame control segment that indicates whether the frame is a control, management, or data type
* address segments that include the MAC addresses of transmitter, receiver, and the final destination at either end
* sequence control data that helps to reorder frames that arrive at phase-delayed times
* the actual body of the frame (data)
* a frame sequence check (a 32 bit checksum for error correction) [3]

At the receiver end, I can verify the state of this data relative to the transmit data, verify timing assumptions (to an imprecise degree), gather RSSI levels and the bit error rate for a given modulation. At the receiver end, I log the data continuosly for the sensor data. Every time the Android OS has the ability (due to processor congestion and timing), it measures the raw voltage levels on the 9-DoF chip. The network state data, because it is user-space software, takes time to make its measurements. Since this problem is outside of the scope of this project, training data will only be taken from timestamped data that has no missing features. It is possible that a time-series average of the sensor data can replace the instantaneous measurement in the model. However, more investigation is needed.

In addition to wifi data, I can collect the same information from a bluetooth radio. However, instead of operating in the 2.4 or 5 Ghz ranges familiar to wifi, it operates in the sub Ghz range around 900Mhz. Because of the difference in these wavelenghts, the two fields operate somewhat differently at a human scale. Bluetooth is much less sensitive to fading and other non-line of sight measurements. The difference in estimated fading margins between wifi and bluetooth can be used to get a characteristic of the environment.

In addition to these services, most Android phones have LTE radios. The packets for which carry all of the above information, plus some location data provided by the network. Using multilateration, the CDMA frames can transmit data using one of 4 techniques:

* Cell ID: where a user is localized to their home node using a path loss fading model. It is highly inaccurate and the easiest to implement, requiring only information from the cell tower. In essence, this method places in the range of one tower over another but does a poor job of fine localization.
* ECID: position is localized to a finer level using path loss models. However, this only gets us a position within 50-100 meters of the true location. For our purposes this is much too rough an estimate.
* OTDOA: This method uses the timing of downlink signals from multiple CDMA towers to establish a location via multilateration. Measurements are made on the base station end, reducing the privacy of the user.
* Uplink time of difference arrival: This method, newer than the others, uses timing differences between transmit time and receive time to calculate distance. Because it exploits already existing network control packets, it doesn't require any additional bandwidth resources. This method however, is only accurate when the base stations are on the order of a kilometer away from the transmitter. If the distance is significantly less than that, a speed-of-light signal wave will cover the distances in less time than standard computer clocks can measure--ie the difference of timing will be in milliseconds while the smallest timing resolution is only in milliseconds. Despite the limitations of these technologies, a hierarchical regression algorithm can exploit the confidence of the CDMA measurements and the high number of bluetooth and wifi access points to get a fine resolution localization using the law of large numbers and a nearest-neighbor weighting scheme. However, in order to determine the location of other access points, we must measure from many directions, use appropriate path-loss models to measure distance, and average these measurements in a way that consistently gives more weight to less noisy measurements.

**Previous Models [4]**

*Friis Equation*

In the simplest scenario, we can model the free space path loss. This function of received power in terms of distance is known as Friis' Law:



Where *P* is the power, *G* is the gain, and *A* is the area of the receiving antenna, *d* is distance between the transmitters and lambda is the frequency. This equation only applies for systems that are separated by at least one Rayleigh distance defined by



This is known as the far field. When dealing with link budgets, it is best to use a logarithmic scale because signal levels will vary across many orders of magnitude.



if and only if the powers and gains are consistently in dB or dBm.

**Kirchoff Theory**

In the same way that Rayleigh distance defines the breaking point of Friis system, Rayleigh roughness can be thought of as the order of scale that causes interference due to scattering on a surface.



Likewise, the spatial correlation function for a Gaussian signal and Gaussian height distribution is defined as



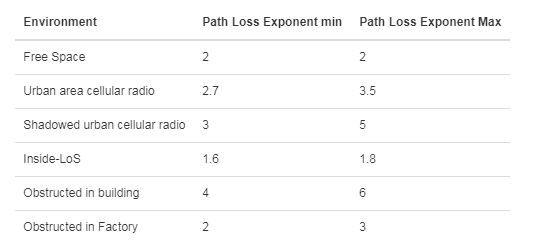
where W is defined above, E is the signal function, and h and is the height distribution. These calculations are particularly useful in outdoor urban environments as the topological height has nothing to do with the sight lines to the horizon.

*Log-Distance Path Loss Model*

The next model, log-normal shadowing can be thought of as an extension of the Friis model with the added inclusion of a random variable. If the receiver is in the far field of the receiver (where d > d\_r), PL(d0) is the path loss measured at a distance d0 from the transmitter, then the path loss when moving from distance d0 to d is given by the equation



where *n* is the path loss exponent, given by the table below and chi is a zero mean random normal distribution. [5]

Table 2

**Noise**

In most circumstances it is too cumbersome to describe all the sources of noise. If it wasn't, then we'd solve Maxwell's Equations within the system and be done with it. However there are many sources of interference. Multipath propagation is a big limitation to predicting the shape of a wireless network in an urban space as different rays from the transmitter bounce around the room and reach the receiver at different times. Because receivers cannot distinguish the 'true' signal from the multi-path signal, the receiver just adds the components of those multipaths up, creating interference. This interference can be constructive or destructive. Small-scale fading occurs when a user is moving--when the user moves further away, they increase the phase and possibly the measured voltage level. Additionally, the transmitter and receiver each produce noise. The total power due to noise can be described as



Where xI and xQ are linearly independent vectors of amplitude, *I*, and phase, *Q*. The signal to noise ratio can be expressed as



where P is the average power measured at the equivalent points in a system and within the same system bandwidth. If we include interference, the SNR becomes SINR or the signal to interference and noise ratio.



Interference can also be created by objects in the environment. Buildings and other obstacles to propagation can block regular transmission, causing any signal to make it behind the building to be greatly attenuated. This is called shadowing. Note that this not only happens to the line-of-sight components of the wave front, but any multi path ray! For this reason, buildings and other obstacles give rise to large-scale fading by creating a diffraction pattern of the signal wave in the shadow of the building. [4]

These components can undergo many types of transformations. Reflection is when a field reaches an object with very large dimensions compared to the wavelength of the field. A wi-fi signal is reflected off of buildings and the ground, creating patterns of de/coherence. Diffraction occurs when a transmitter and receiver are obstructed by a surface, 'bending' the wave around the obstacle. When a field travels across a surface with dimensions that are small compared to its wavelenght and where the number of obstacles per volume is high. Refraction is a result of the field traversing multiple mediums, much like how a pencil appears to 'bend' when placed in a glass of water. For much higher frequencies (starting at 15Ghz), atmospheric absorption is a problem, but for indoor distances and low frequencies, the effects are negligible.

[4]

*Log-Normal Shadowing Propagation Loss Model*

The next model builds on the Log-Normal shadowing model by including a random variable, chi used in the calculation of shadowing effects. This law can be expressed as:



Where chi is a zero-mean Gaussian distributed random variable. This variable is only used when there is a shadowing effect. Equivalently, $\chi\_2=0$ when no shadowing effect is present.

*Two Ray Ground Propagation Loss Model*

Thus far, we have not needed to think of our signal and other multipath components as being complex vectors. When we start trying to predict the way these multiple paths interact analytically, phase shifting due to refraction, fading, and other multipath effects can be modelled by using a two path tracing technique. The figure below illustrates this model.

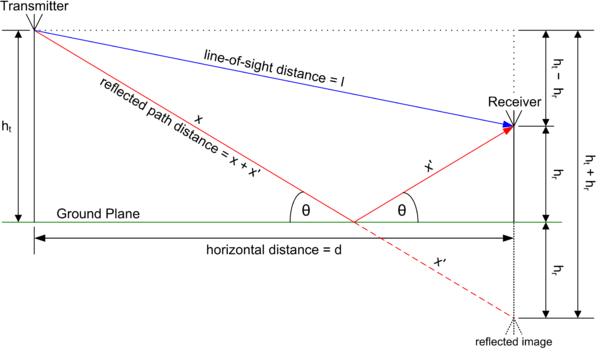
****

Figure Source: Wikimedia Commons

If the path length between Tx and Rx is d, then the two math model can be expressed as



*ITU Propagation Loss Model*

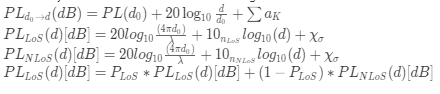
The next model is a semi-empirical model, based on the same Friis equatoin, with added parameters for noise due to fading between floors in a building.



Where Pf(n) is a power loss function due to the number of floors through which the field propagates. In addition, the $\chi$ value of previous models has been set to -28 in the ITU model, taken from standardized urban measurements. [4]

*Motley-Keenan Model*

The Motley-Keenan Model follows the same logic.

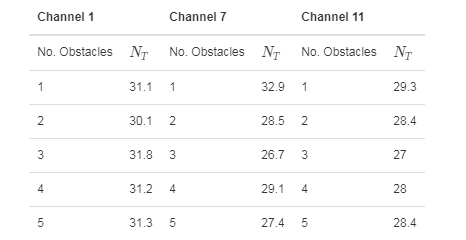


Except in this case, a path loss model is drawn for each dominant LoS and NLoS path.

*Tata Modified ITU Model*

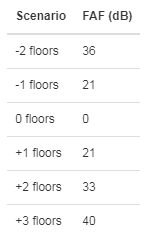
A modified version of the ITU model has been used to great effect in India. [6]





It uses an estimate of the number of obstacles to determine a site-specific path loss parameter. In addition, it adds a floor attenuation factor for each floor between the user and the transmitter. Those FAF numbers are given below (Table 3).

Table 3



The Longley-Rice model is the most comprehensive model for propagation across complicated urban spaces. This model requires a very large amount of parameters

* Frequency
* Distance
* Antenna Heights
* Polarization (vertical or horizontal)
* Terrain irregularity ($\Delta h$)
* Electrical Ground constants
* Refractivity
* Climate
* Siting (how well the technician chose the broadcast site)
* Reliability and confidence level (of the other parameters)

Suffice it to say, that while some of this data can be estimated from a map database, the depth of this model is outside the scope our design constraints (as well as this 15 page paper).

https://www.ntia.doc.gov/files/ntia/publications/ntia\_82-100\_20121129145031\_555510.pdf

**Purely Statistical Models**

There are some purely statistical ways to describe the properties of a wireless network. In the case of transmission/reception among primarily non-line-of-site paths, the cumulative distribution of the received signal strength indicator will follow a Rayleigh distribution where



In case the primary mode transmission is across the line-of-sight path, then the RSSI distribution follows a Rice distribution such that



By performing a \chi-square test on our measured data at a given location and either distribution, we can map a given location as either predominantly LoS or NLoS relative to a given receiver.

In addition to noise levels, deemed sufficiently normal (ie follows a Gaussian distribution for a given space), we can calculate the bit error rate for a given modulation scheme as



where v is the velocity and and T is the period at a given time and frequency respectively. Both of these are easily reported by a nine degree of freedom chip. Experimental and theoretical estimations reduce the complexity of our calculations, BER can be calculated as



where K is a proportionality constant, S is rms delay spread on the channel, and T is the period. These calculations can help us parse a measured loss of signal into noise and geometric components. The soft error rate is the rate at which errors occur. On a physical level, this is calculated with the same methods as BER. However, it comes in the form as a Poisson distributed event that leads to the average rate of soft errors in a channel. Parsing this from BER will be difficult, especially with autoencoding techniques and

**Analysis**

Standard wireless models are great at modelling unbostructed line-of-sight connections. However, they fail to accurately model the network states at peak noise, in complicated urban environments both indoors and outdoors. In addition, these models live entirely in the proprietary world of the IEEE website. Figure 4 (from the Tata model paper) shows how these models compare to real world data. As we can see, real data is most closely modelled by the Motley-Keenan method, and simplistic model standards can lead to results that are of by as much as 10 or 20 dBm ! Implementing these models in Python would go a long way to modelling the bigger problem. This is the first step. Since some quantities required for these models are unknown, a gridsearch algorithm will be used to optimize these algorithms and compare results.

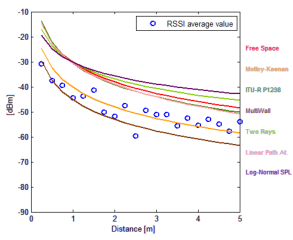


Figure 4

**Conclusion**

Since these models require different sets of parameters, what features our dataset needs is an open question. The first method will be using a k-random-forest algorithm to see what features are relevant to the signal strength between two nodes.

The next method will take the most useful features and use them to build a logistic regressor that predicts the distance between the two nodes in the network according to the most appropriate model. Then, these distance vector calculations can be cross-referenced to the location data of the LTE packets as well as large online databases of SSIDs (like wigle.net) to build an on-device location estimator. The chief advantages of this system are that location services will not require transmitting private information to any external databases, and since GPS requires LoS to a satellite, this system would be far more appropriate in dense urban environments. As a point of reference, off-the-shelf systems already exist to use these techniques as a location estimator. However, they use site fingerprinting to develop a classifier model instead of using distance data to determine a precise location.

As far as I can tell, this method has not been implemented yet, and while outside the scope of an undergraduate term paper, I plan on continuing to build this system while pursuing my Master's degree in Data Science.

References

1. What is I/Q Data? (n.d.). Retrieved from <http://www.ni.com/tutorial/4805/en/>
2. X. (n.d.). Performance analysis of adaptive M-QAM over a flat-fading Nakagami-m channel.
   1. Retrieved from <http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S0038>
   2. 23532011000100012
3. Molisch, A. F. (2011). Wireless communications. Chichester, West Sussex, U.K.: Wiley
4. Fund, F. (2017, April 26). Understanding the 802.11 Wireless LAN MAC frame format.
   1. Retrieved from <https://witestlab.poly.edu/blog/802-11-wireless-lan-2/>
5. H. K. Rath, S. Verma, A. Simha and A. Karandikar, "Path Loss model for Indian terrain

empirical approach," 2016 Twenty Second National Conference on Communication

(NCC), Guwahati, 2016, pp. 1-6.doi: 10.1109/NCC.2016.7561183