**Finding Missed Phone Call Fraudsters**

**Team: Sri Kanajan (Zipfian Academy), Arshak Navruzyan (Argyle Data), Nataliya Nadtoka (Argyle Data)**

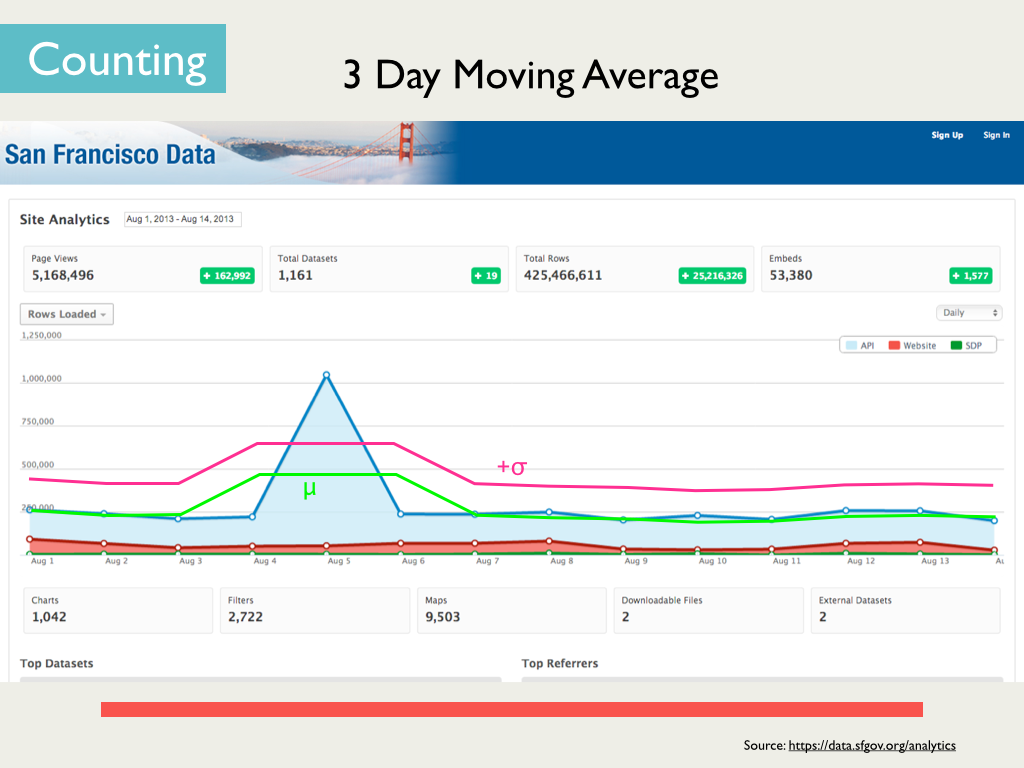
July 2nd 2014

$2 billion dollars a year is lost due to fraudulent cases of phone calls. For example, a fraudster calls a large number of phone numbers at random and leaves missed calls on these phones. Unsuspecting users may consider the missed call as a legitimate number and redial the fraudster's number. The fraudster's number however is a premium rate call or contains automated advertisements which is then automatically charged to the unsuspecting user while enriching the fraudster. This fraud is named the Wangiri fraud [1].

**The challenge is to identify the fraudulent calls while strictly minimizing the number of false positives[[1]](#footnote-2).**

**Current Solution**

The current solution is to use an ensemble of statistical moving averages around the time series data that represents the caller-callee sequence. This approach only relies on the aggregated count of phone calls between countries over time, and whether this count exceeds some threshold (e.g. two standard deviations above the 1-day, 3-day, EWMA moving average). Illustration illustrates this concept.

Illustration 1: Example of a Moving Average Anomaly Detection Methodology

**Limitations of Current Solution**

Smart fraudsters maybe able to fool statistical detection means by adapting their fraud call rates such that the rates are within the moving average windows. The other limitation is that this approach is prone to false failures[[2]](#footnote-3) which are very costly and creates a sense of disbelief on the fraud detection system in general.

**Proposal**

The opportunity for improvement is based upon the premise that there maybe more interesting features beyond the aggregate volume of calls between countries, that may further improve the fraud detection capability. Two specific approaches are currently planned:

1. Using the statistical approaches as the baseline, identify anomalous calls. Then using feature importance concepts in machine learning (E.g. using random forests), identify other features (e.g. phone region code, caller number type) that most differentiate between the anomalous calls and normal calls as identified by the statistical approach. The overall goal is to use a trained supervised machine learning classifier utilizing these features as an additional ensemble participant that may help refine the statistical fraud detection approach.
2. Model the call sequence of caller's and callee's as a directed graph where the nodes are specific phone numbers and the directed edges represent the caller/callee order. This can be done within specific time windows. After generating the graph, derive graph metrics such as PageRank, Closeness, Betweeness and evaluate the outliers based on these metrics [2].

**Data Source**

Argyle Data [3] is the company sponsoring this project. The time series data of all calls going to/from the country (Ireland is the specific country of focus) is contained in a key-value pair database Accumulo which has a SQL interface using Presto. An example of the exported data is in Illustration.



Illustration 2: Example of time series data representing caller-callee data

**Technologies**

For the data extraction, SQL using **Presto** will be utilized. For the EDA (exploratory data analysis) and statistical methods, **Python pandas** will be used. For feature importance analysis, either **BigML** or **SKLearn** in Python will be utilized. Graphs will be generated using Python's **Matplotlib**.

For the directed graph portion, Python's **GraphLab** will be utilized and for visualization, **Gephi** is the likely candidate tool.

**Deliverables:**

1. A presentation slide summarizing the project and its findings.
2. iPython notebook containing a well described flow of the code and visualizations
3. SQL, Python Code
4. Any visualization models (e.g. Gephi)

**Project Plan (Tentative based on Zipfian approval)**

*7-2 – 7-9 : Perform statistical based anomaly detection on isup time series data*

*7-9 – 7-12: Derive most informative features through machine learning approaches*

*7-14: Report intermediate findings to Arshak and Natalia*

*7-14 – 7-18: Transform data into network graph*

*7-19 – Perform graph metric analysis and derive outliers*

*7-21 – 7-25: Clean up, prepare presentation and report*

*7-28: Present project findings to Argyle and Zipfian*

**Notes:**

1. Twice a week status updates to be sent to Arshak and Natalia.

2. Project commitment will be 50%-80% based on classes and ongoing job search activities

3. Project data[[3]](#footnote-4), code, results and visualizations maybe presented to the external parties as well as on a public repository such as github.

**References**

1. Wangiri Fraud – <http://en.wikipedia.org/wiki/Wangiri>
2. Graph metrics – <http://reference.wolfram.com/mathematica/guide/GraphMeasures.html>
3. Argyle Data – [www.argyledata.com](http://www.argyledata.com/)

1. A small number of undetected fraud cases are more acceptable than a large number of false positive cases that may cause significant wasted effort and loss of confidence around the fraud detection system. [↑](#footnote-ref-2)
2. For example, sudden increase in call volume maybe genuine and attributed to external factors such as holidays, major events or personal situations between family or friends from different countries. [↑](#footnote-ref-3)
3. Any data will go through Argyle approved sanitization prior to publishing. [↑](#footnote-ref-4)