Fighting Future Fraud

A Strategy for Using Big Data, Machine Learning, and Data Lakes to Fight Mobile Communications Fraud

Authored by:
Dr. Ian Howells
Dr. Volkmar Scharf-Katz
Padraig Stapleton

ARGYLE DATA

TABLE OF CONTENTS

Introduction – Future Fraud and Big Data

We have had the privilege of working with global leaders and visionaries on their strategies for future fraud, big data, and machine learning. What consistently comes up is that best-in-class carriers know the fraud types and fraud methods that they have been attacked with in the past, and they know the scale of fraud today. However, what keeps them up at night can be captured in a famous phrase by Donald Rumsfeld:

"There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know."

This translates to:

- **New Fraud** - What new types and methods of fraud are criminals using that we don't know about and that we aren't detecting?
- **New Platforms and Technology** - How are we going to use new big data platforms and machine learning to detect both old and new types of fraud in real time?
- **New Sources of Data** - How are we going to protect subscribers from fraud in the new world of connected cars, connected home, mobile payments, IoT in utilities, and IoT in health and fitness?

What is critical to understand is that a) criminals are continually innovating; b) each subscriber will have many devices, many channels, and many potential attack points; and c) we need a better way to detect new fraud and protect customers and carriers in this new world – today in 2015, not in 2020.

It is commonly agreed that you can no longer effectively protect a subscriber with silos or disconnected, disparate systems. What is required, to protect subscribers effectively in the new world, is a "big data/data lake" strategy that encompasses batch billing data, real-time call, VoIP, SMS and data packets, and business data. To truly get the value from an application accessing a data lake means discovering what you don't already know. This makes asking the right questions much harder. What is needed to get the right answer, at scale, is real-time machine learning and anomaly detection.

Criminals make their money in mobile fraud by behaving in an anomalous way, exploiting loopholes or arbitrage opportunities in the system. If they continue to behave in this way, automated anomaly detection systems can now identify them.

Modern machine learning can discover anomalous behavior in real time, uncover new and old types of fraud, and treat a loyal customer who has paid their bills on time for five years differently than a brand new customer. It is critical to no longer have separate applications and databases for each type of fraud. What is required is to provide a unified strategy to detect both traditional and new attack vectors, with an integrated infrastructure for applications such as:

- Domestic and roaming fraud
- Mobile payment fraud
- Never pay fraud
- Negative margin fraud
- Arbitrage fraud
- SS7 security
- IoT abuse and fraud – connected car, connected home, and mHealth

Modern machine learning and analytics systems can visualize fraud by highlighting anomalies that likely point to fraudulent behavior. When you see fraud visualized in this way, it shows beautifully obvious anomalous behavior that is very difficult to hide. What struck us when looking at these visualizations is that there is a common pattern to fraud and combinations of fraud. Visualizations bring fraud to life and make it beautifully obvious to a human. This is what machine learning, when combined with big data, does at scale.

Big data is commonly described in terms of the ability to handle “volume, velocity, and variety”. In this book we take these big data concepts and show how to apply them to beat mobile fraud. In the following pages, we examine a common volume and velocity attack pattern across a variety of channels that is equally applicable to many combinations of fraud types and methods, showing different views through graphs, analytics, and anomaly detection visualizations.

Part 1:

CURRENT FRAUD LANDSCAPE

2014 will be remembered as the year that the fraud and security dam broke with fraud moving from being a back-office subject to front-page news. Existing fraud management systems are losing the innovation battle against sophisticated cyber criminal gangs. What struck us when we researched the market is that “Fraud” is a dirty word that people don’t like to talk about. There is a lack of knowledge about the scale and impact of fraud at senior management and board level. On average, a company loses 5% of revenue to fraud. One of our favorite articles, *When Will CFO’s Put the “F Word” in Their Annual Reports?*, discussed this and provided the following questions that every analyst and shareholder should be asking a company now:

- What is the bottom line impact of losses due to fraud in your company?
- What is the cost of customer service and churn due to fraud in your company?
- What is the impact of fraud on your earnings per share and stock price?
- How do you compare to the average performer in your industry in preventing fraud?
- What are you doing to protect your brand from the reputational damage fraud causes when it becomes public?

We have been privileged to work with some global leaders that understand the impact of the “F Word” and believe that every enterprise and carrier has an obligation to protect its subscribers from fraud – and can differentiate itself by doing so.

In order to broaden the conversation about fraud, and make it possible for all relevant players to participate in the conversation about fraud prevention, everyone must first have a basic understanding of the current fraud landscape.

FINANCIAL IMPACT OF FRAUD

When we researched the market, one thing that struck us was that criminal adversaries are out-innovating enterprises – and are doing so because the rewards are so large. It was common to see enterprises using the same approaches they were using 3 to 5 years ago, but the world of fraud and the level of sophistication has moved on dramatically since then. This is literally costing communications service providers (CSPs) billions of dollars per year.

- The Association of Certified Fraud Examiners (ACFE) reported that the typical organization loses 5% of revenue each year to fraud – a global loss of $3.7 trillion.
- The Communications Fraud Control Association (CFCA) reported mobile and fixed line carriers lose $46 billion per year to fraud.

The CFCA also details fraud’s impact at a more granular level. Following are their reported fraud costs by fraud type, method, and region.

Costs by fraud type:

- Roaming Fraud
  - $6.11 billion Globally
  - $1.75 billion North America
  - $1.95 billion Western Europe
- Premium Rate Service Fraud
  - $4.73 billion Globally
  - $1.35 billion North America
  - $1.51 billion Western Europe
- IMEI Reprogramming Fraud
  - $2.60 billion Globally
  - $0.74 billion North America
  - $0.82 billion Western Europe
- Interconnect Bypass Fraud
  - $2.00 billion Globally
  - $0.56 billion North America
  - $0.63 billion Western Europe
- International Revenue Share Fraud (IRSF)
  - $1.80 billion Globally
  - $0.51 billion North America
  - $0.57 billion Western Europe

Costs by fraud method:

- Subscription Fraud
  - $5.22 billion Globally
  - $1.49 billion North America
  - $1.66 billion Western Europe
- PBX Hacking
  - $4.42 billion Globally
  - $1.26 billion North America
  - $1.41 billion Western Europe
- Abuse of Service
  - $2.70 billion Globally
  - $0.77 billion North America
  - $0.85 billion Western Europe
- Wangiri Fraud
  - $2.00 billion Globally
  - $0.57 billion North America
  - $0.64 billion Western Europe

- Phishing
  - $1.70 billion Globally
  - $0.50 billion North America
  - $0.55 billion Western Europe
- SMS Faking or Spoofing
  - $1.60 billion Globally
  - $0.46 billion North America
  - $0.51 billion Western Europe
- Signal Manipulation
  - $0.90 billion Globally
  - $0.27 billion North America
  - $0.30 billion Western Europe
- SIM Cloning
  - $0.50 billion Globally
  - $0.15 billion North America
  - $0.17 billion Western Europe

HOW FRAUD OCCURS

The distinction between fraud types and fraud methods is an important one as it distinguishes how fraud is perpetrated. A fraud type is a way to monetize fraud. For example, premium rate service fraud involves the use of premium rate numbers where callers have to pay a fee for calling the number. But if no one calls that number, no profit is made. That is where the fraud method comes in. A fraud method is a way to drive a large amount of traffic to a fraud type. A common fraud method to drive traffic to a premium rate number is known as Wangiri fraud, where a robo-dialer calls thousands of numbers and hangs up after just one ring. People see a missed call, and a surprising number of them (on average 20%) call back without realizing they are calling a premium rate number. Wangiri creates the demand and premium rate service fraud monetizes it.

Note that there can be many different combinations of fraud types and methods. For example, SMS phishing is another common method to drive traffic to a premium rate number. A very successful real-world fraud campaign used SMS phishing to send a message that said “please call this number, we are trying to deliver flowers to your wife” and listed a premium rate number for call back.

Other key distinctions are that fraud may occur:

- When a subscriber is roaming abroad
- When a subscriber is in their home country

Fraud may also occur with:

- Voice
- Data
- Text
- And combinations of the above

FRAUD DETECTION AND ANALYTICS SYSTEMS

Criminals typically use new variations or different combinations of fraud, which are very difficult to detect using traditional methods that attempt to detect known, previous patterns of fraud. Criminals are innovating rapidly while many carriers continue to try to defend themselves with old techniques and technology based on:

- Silos of data across multiple systems
- Batch approaches using ETL/EDW
- Rules based on old, known types of fraud
- Business intelligence

Existing systems that that utilize outdated technologies simply can’t catch modern fraudsters because they:

- Fail (don’t discover fraud)
- Overwhelm (bombard users with false positives)
- Operate in batch (discovers fraud after the criminal has gone)
- Use dated rules (discover last year’s fraud, not today’s fraud)

To compete and out-innovate modern cyber criminal gangs, leading CSP’s are utilizing big data, Hadoop, and machine learning to provide real-time access to huge amounts of data stored in a “data lake”. There is a shift from:

- Batch to real-time
- Thresholds to anomaly detection
- Rules to machine learning
- SQL to SQL and graph analysis
- Silos of data to data lakes
- Scale-up hardware to commodity Hadoop architectures

These modern systems defend against and proactively attack fraud by:

- Detecting fraud in real time
- Detecting both new (unknown) and old (known) types of fraud

- Minimizing false positives
- Identifying crime rings and not just individuals
- Detecting test attacks
- Identifying local accomplices

More information about a modern approach to fraud detection that utilizes new technology can be found in part 3 of this book.

ABOUT ARGYLE DATA

Argyle Data is the leader in real-time fraud analytics at network speed and Hadoop scale, offering solutions for the largest data-driven mobile communications companies. Argyle Data's application is built from the ground up on Hadoop using the latest big data, machine learning, and anomaly detection technology proven at Facebook and Google. It is able to detect fraud not detected by existing systems, discover fraud in minutes vs. days, discover both new and old fraud attack techniques, and dramatically reduce false positives.

ARGYLE DATA, INC.
2755 Campus Drive, Suite 165
San Mateo, CA 94403
USA

Tel: 800.695.6021
Email: info@argyledata.com
Web: www.argyledata.com

ARGYLE
DATA