Operational Simulation and Optimization of Fraud Detection Systems

GCX Advanced Analytics LLC

In this paper, GCX shows how our fraud operations simulation model provides insights into the operation of a detection system, and identifies optimization opportunities. The model is 'trace-driven,' meaning that is uses the actual historical event sequences generated by the detection system to drive the model. Visualization of the detection system state over time includes dollars-at-risk, fraud investigator utilization, and alert backlog. Conformance with service or operating level agreements can be assured using the simulation model.

Fraud operations managers need to be able to predict and optimize their staffing levels and alert score threshold. This is usually a trial-and-error process based on reports that are weeks or even months out of date. Online and mobile banking particularly make this problem complicated, since the system is on 24x7. The GXC Operations Simulator provides a rapid and effective method for determining answers to the following business questions:

- What is the right number of detection bankers or fraud investigators? By day of week, hour of the day?
- How is the economic value of the fraud detection system maximized?
- What is the best score threshold for review, for each alert type?
- How should alerts be prioritized? By score? By dollars-at-risk?
- How long do alerts stay in the queue before a detection banker starts the review process?
- Are service-level or operating-level agreements being met for reviewing alerts?
- How many alerts should there be in the queue at any time?
- How much risk score is in the queue?
- How many dollars are at risk in the queue?
- What will the actual performance of the detection system be in terms of dollar detection rate, false positives, and latency?

OVERVIEW OF THE GCX OPERATIONS SIMULATOR

The GCX Simulator models the operational system shown in Figure 1 below. This model captures the characteristics of the system necessary to answer the business questions posed at the beginning. Alerts are generated by the function of *the fraud risk scoring engine* (or a policy comprised of rules in some cases) as inputs are provided to it by either events (messages), batch delivery of data, or both. In this paper, we are dealing with the example of a fraud detection system for online and mobile banking, and as such, these types of systems are always on (24x7), and consequently generate alerts at all hours.

Alerts enter into a queue for assignment to fraud investigators (or detection bankers), and eventually receive a disposition as either *true positive* (the alert correctly identified fraud), or *false positive* (the alert was a false alarm). The life cycle of the fraud alert is shown in Figure 2 below. While the actual workflow states of an alert may be more complex than what is shown in the figure, this life cycle captures the salient characteristics required to answer the business questions posed at the beginning of this paper.

Effectively, this is a type of queuing system[1].

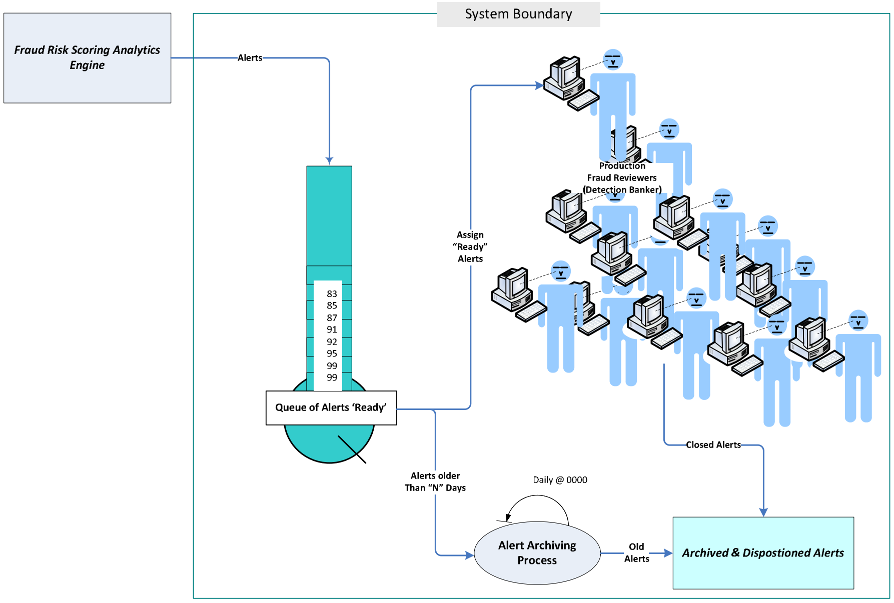


Figure 1. Operational Model of the Detection System

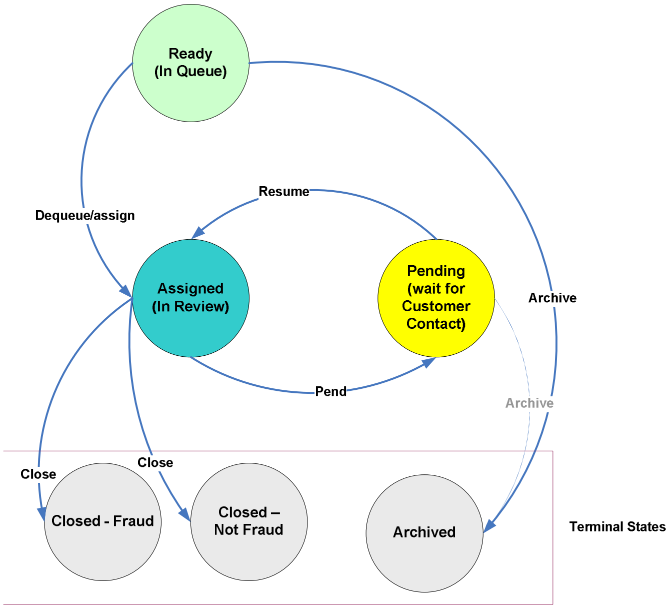


Figure 2. Alert Life Cycle

[1] For the technical reader, this is a Single Queue/Multi-server non-Stationary arrival rate/non-stationary service rate with multiple entity types system, a.k.a. Ga(t)/Gs(t)/1; no closed-form solutions exist for this type of system.

TRACE-DRIVEN SIMULATION

The GCX Simulator is *trace-driven*, i.e. it processes a record of the actual alert generation process of the operational system. This is different from typical simulators that use random number generators to produce the arrival of events into the system. The alert timestamps are something that we know with certainty. Here, reality drives the model. Table 1 below shows a small sample of an alert trace file. The trace file is essentially a log of the arrival time of an alert to the system, and its various attributes. The 'outcome' is optional, but useful for *ROC optimization* (detailed later).

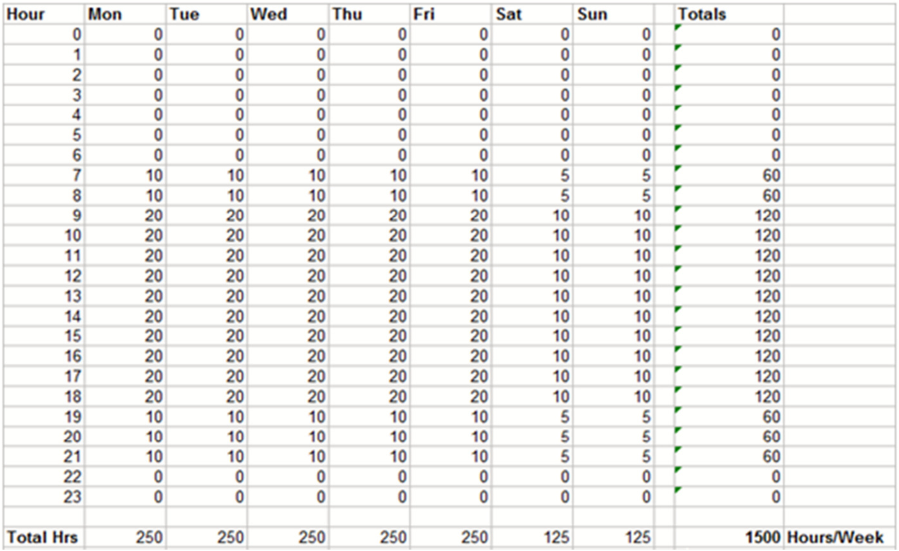
Table 1. Example Trace File

Timestamp	Alert_Type/ID	Score	Alert_Dollars	Outcome
22-NOV-2010 11:20:01	MOBILE_00012011	850	$0.00	FP
22-NOV-2010 11:20:32	LOGON_12120001	900	$0.00	FP
22-NOV-2010 11:22:00	ACH_01012233	950	$973.57	TP
22-NOV-2010 11:22:05	BILLPAY_00023211	850	$2,522.34	FP
22-NOV-2010 11:22:15	WIRE_12120002	1000	$0.00	FP
22-NOV-2010 11:25:02	ACH_01012234	950	$2,775.24	TP
22-NOV-2010 11:25:55	MOBILE_00012012	890	$2,891.47	TP
22-NOV-2010 11:27:27	ACH_01012235	700	$927.54	FP
22-NOV-2010 11:28:12	BILLPAY_00023212	950	$3,676.67	TP

Alert review times, on the other hand, are usually difficult to know precisely, since investigators are usually performing many tasks at once. The open-to-close time of the alert is usually not a very good estimate of the amount of time actually spent on the alert by the investigator. However, for purposes of analysis and simulation, the open-to-close time is a fair operational measure regarding SLA or OLA compliance. Alert review times in the GCX Simulator are provided by random deviates from the normal, exponential, Erlang, and Weibull distributions.

THE STAFFING PLAN

The operational staffing plan is a key element in the simulation, since it provides the number of available investigators for each hour of the day and day of the week. All of the various constraints on operations are reflected in the plan, i.e. holiday, core working hours, night shfits, weekend, minimum staff level policies (usually 2), and so on. The staffing plan is typically modified iteratively during the simulation and analysis process, as we examine various 'what if' scenarios. Figure 1 below shows an example of a staffing plan.



Hour	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Totals	
0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	0	
6	0	0	0	0	0	0	0	0	
7	10	10	10	10	10	5	5	60	
8	10	10	10	10	10	5	5	60	
9	20	20	20	20	20	10	10	120	
10	20	20	20	20	20	10	10	120	
11	20	20	20	20	20	10	10	120	
12	20	20	20	20	20	10	10	120	
13	20	20	20	20	20	10	10	120	
14	20	20	20	20	20	10	10	120	
15	20	20	20	20	20	10	10	120	
16	20	20	20	20	20	10	10	120	
17	20	20	20	20	20	10	10	120	
18	20	20	20	20	20	10	10	120	
19	10	10	10	10	10	5	5	60	
20	10	10	10	10	10	5	5	60	
21	10	10	10	10	10	5	5	60	
22	0	0	0	0	0	0	0	0	
23	0	0	0	0	0	0	0	0	
Total Hrs	250	250	250	250	250	125	125	1500	Hours/Week

Figure 1. Example Fraud Investigation Staffing Plan

STATISTICS PRODUCED BY THE SIMULATOR

The GCX Simulator generates a wide variety of statistics and measurements about the operational characteristics and detection effectiveness[2] of the system. The measures fall into the following categories:

- Staff Metrics – utilization of investigators, associated costs, alert backlog
- Detection Metrics – the contingency tables by score threshold of the system
- Service Metrics – how long alerts are queued up, how many are in process

These are detailed further below.

Table 2. Staff Metrics

Actual Review Cost – total time spent reviewing alerts × the labor rate for detection bankers
Staff_Cost – total time clocked by detection bankers × the labor rate for detection bankers
Staff_Hours – the total time the staffing plan provided
FTE_Level – the Staff_Hours represented as FTE
No_Alerts_Archived – number of alerts not reviewed and finally archived
Utilization of detection bankers against staffed hours (busy time/Staff_Hours)
Utilization of detection bankers against staffed hours plus unstaffed periods (e.g. 10pm to 6am).
Number of bankers
Number of busy bankers
Run rate of labor (number of bankers x labor rate)

[2] See GCX's white paper, "Performance Evaluation and Optimization of Fraud Detection Systems," for complete details, www.gcxanalytics.com.

Table 3. Detection Metrics

For each type of alert {Bill Pay, ACH, Wire, Transfers, Login, Mobile, etc) counts of:
Number of alerts above review threshold
Number of alerts below review threshold
Dollars above review threshold
Dollars below review threshold
For true positive alerts
Number of frauds above review threshold
Number of frauds below review threshold
Number of false positives
Fraud Dollars above review threshold
Fraud Dollars below review threshold
For each type of alert {Bill Pay, ICT, Login, EDP, Mobile):
True Positive Rate
False Positive Rate
Receiver Operating Characteristic Points (iterated)
Contingency Table by Units
Contingency Table by Dollars
Review Cost incurred
Loss Dollars
Fraud Dollars detected
Net Economic Benefit
'Better than Doing Nothing' dollars

For each alert type: The Average, Max, Min, Standard Deviation, Variance, and Count as show below.

Table 1. Service Metrics

Time in System – time of alert arriving into ACM to when it is closed
Time in Queue – time in the 'Ready' state until assigned to a detection banker
Service Time – time spent by detection banker dispositioning the alert
Number in System – the number of alerts in the queue and under review by bankers
Number of alerts in Queue

All of these statistics are further sub-grouped as follow:

- Hours of the day (00-23)
- Day of Week (Mon-Sun)
- Business Days (M-F)
- Weekends (Sat-Sun)

VISUALIZING THE SYSTEM STATE

Visualization is an effective way to understand the characteristics of complex systems, such as fraud detectors. Instead of poring over tables of numbers, we can generate various graphs and charts that readily reveal what the system is doing over time. The charts and graphs in this section are generated from the GCX Simulator's output file.

Fraud detection is performed either continuously (a.k.a. 'real time[3]' or 'event driven'), in batch processes, or a combination of the two. In the case of online and mobile banking systems, customers are able to perform banking operations continuously. But, customers tend to use the system more during business hours and less at other times. A typical pattern of usage of online banking is shown in Figure 2 below. This load function varies with the day of the week, within the month, and the time of year.

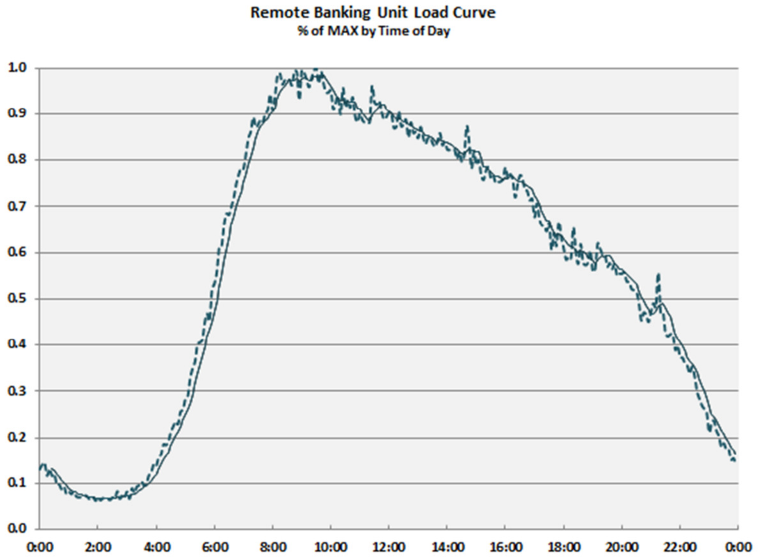


Figure 2. Remote Banking Unit Load Curve

WORKLOAD OVER TIME

The *alerting rate* of a system is fraction of transactions or events for which an alert is generated. Most of these are false positives, and some are true positives. Since a large fraction of alerts are false positives, the arrival rate of alerts to the system tends to trace the load curve shown above, just scaled to a different level. This is shown in Figure 3 below.

[3] This is not real time processing in the engineering sense of the term, as in hard deadline avionics control computers; the term *event-driven* is more appropriate.

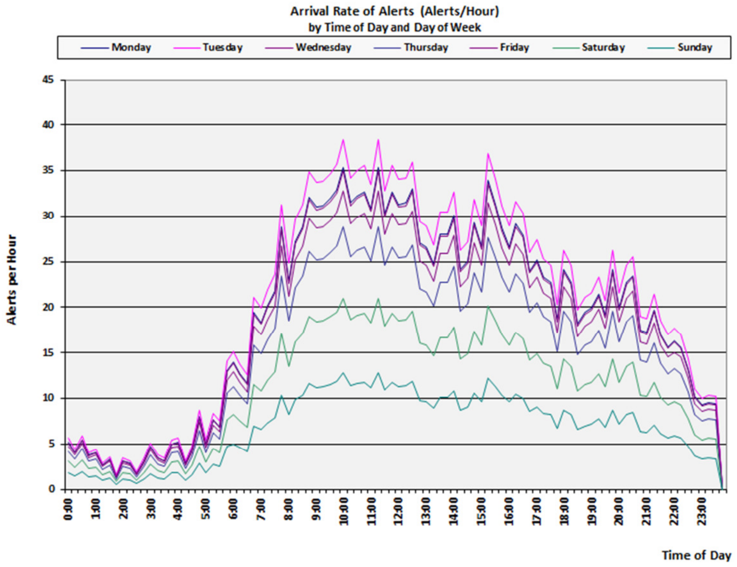


Figure 3. Arrival Rate of Alerts by Time of Day and Day of Week

SYSTEM STATE OVER TIME

The number of open alerts and whether they are in the process of investigation or still queued up waiting for an investigator is of great interest to the fraud operations manager. In the example shown below, there are no investigators on duty between 2200-and 0700 (this is the schedule in Figure 1 above). Alerts build up in the queue while no one is at work, and then at 0700 they begin to get assigned and closed, as the divergence between the 'No. in System' and 'No. in Queue' shows at 0700. The backlog is gradually worked off as the day goes by, so that at the end of the day (2200), there are only 5 alerts in the queue.

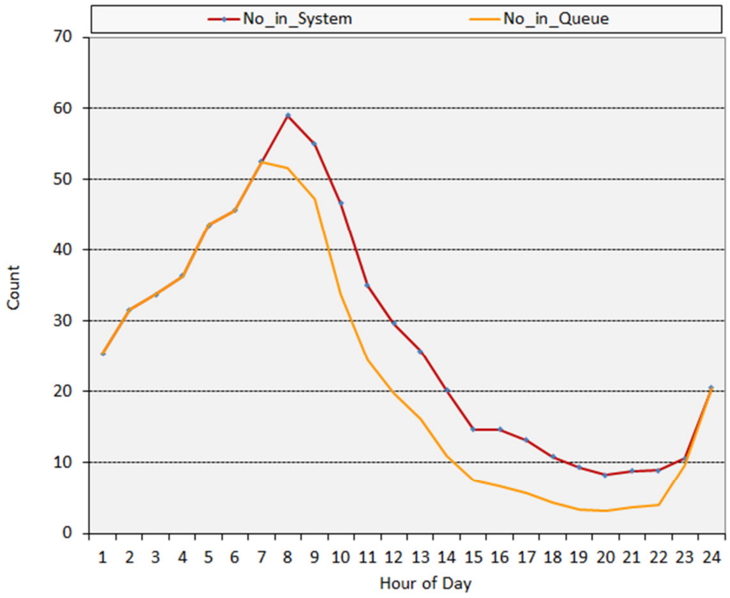


Figure 4. Number in System and in Queue by Time of Day

Investigator Utilization

Also of interest to the fraud operations manager is the utilization of the investigation staff. The available staff are shown in blue below, and the number who are actually busy somewhat less. For example, at 1400, there 20 investigators available, yet only 7 are busy. This helps identify optimal staff levels and opportunities for operational efficiencies.

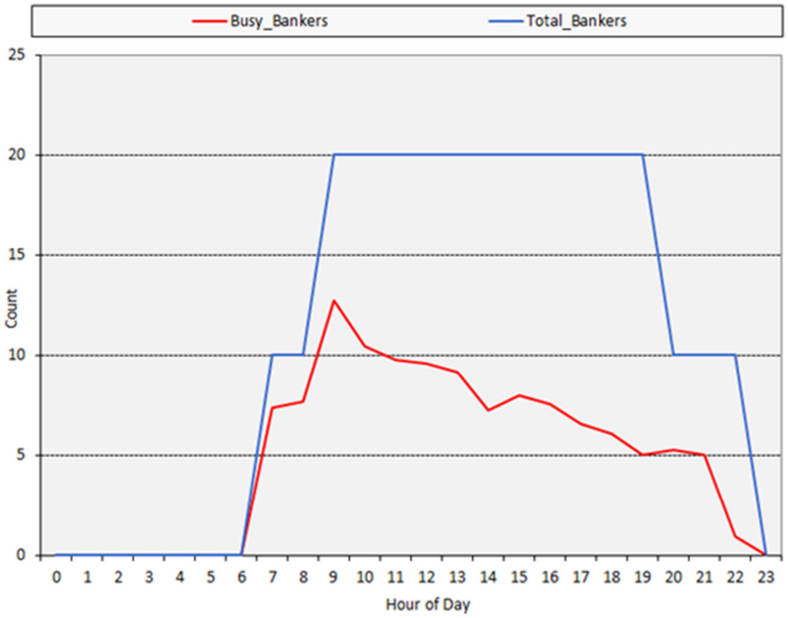


Figure 5. Fraud Investigation Banker Utilization by Time of Day

Service Levels

For some types of transactions, there is an opportunity to interdict them before money leaves the bank. Service levels (or internal operating levels) should be established to assure the bank the opportunity to interdict fraudulent transactions in a timely manner, or take other risk management measures (such as notifying the customer, resetting the password, and so on).

An example of an operating level goal is, "*90% of alerts shall be closed within 22 minutes of their arrival.*" When the time in queue is much greater than the total processing time, this is an indication of an operational issue.

In the graph below, we see the time in system rising around 0800 (the statistic is collected when an alert is closed), and then gradually decreasing over the day as the night's backlog is worked off. Along with the banker utilization chart, this helps identify when we need to have investigators available in order to satisfy the operating level goal.

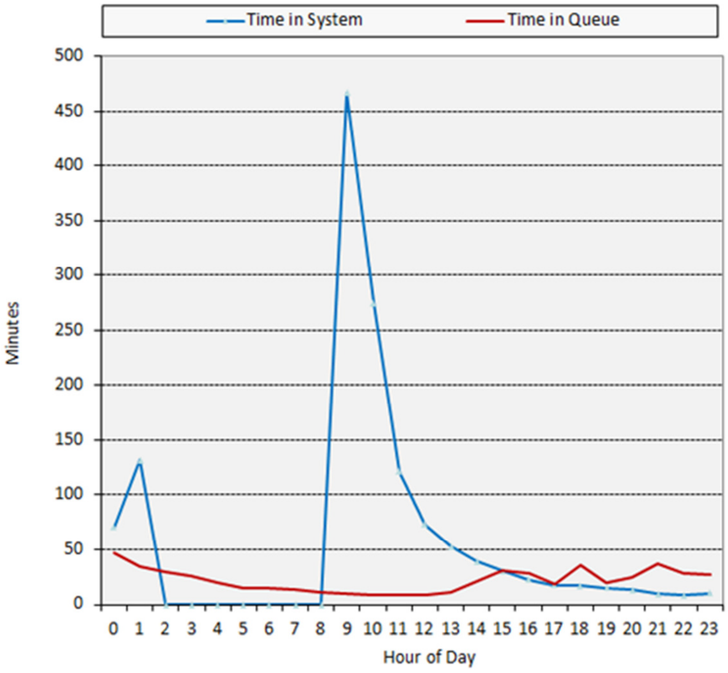


Figure 6. Alerts: Time in System and Time in Queue by Time of Day

Optimizing Review Thresholds

Another major factor in the value of a fraud detection system to the bank is the score threshold above which alerts are queued for review. In the context of credit authorizations, these are the thresholds for *decline* versus *accept* versus *accept & alert*. This analysis assumes that the detection analytics produce some sort of score, where the scores are coherently indicative of the fraud risk (i.e. we can generate a receiver operating characteristic, or ROC, curve for the detection model).

Identifying the optimal score threshold for alert generation is simple and straightforward when there is no constraint on the number of investigators available. This is covered separately in another GCX white paper [1] . However, when we have the constraint of the investigator staffing plan, this becomes what is known as '*hard'* problem[4].

GCX has developed a proprietary algorithm that leverages the simulator's capabilities for evaluation of economic benefits. This enables us to determine the optimal staff schedules and alert review thresholds that maximize the system's benefits to the bank.

Conclusion

Given an alert trace, GCX can optimize the operational parameters of your fraud detection system. This maximizes the contribution of the fraud operations to the bottom line of the bank. The entire process of trace acquisition, simulation, optimization, and reporting out to the bank normally takes less than one month.

[4] Combinatorial optimization (i.e. integer programming) of a constrained non-stationary stochastic system is NP-complete. There are no closed-form solutions for these types of problems (essentially a scheduling problem), so simulation methods must be used.

REFERENCES AND FURTHER READING

[1] GCX Advanced Analytics LLC, "Performance Evaluation and Optimization of Fraud Detection Systems," www.gcxanalytics.com, 2008.

[2] George S. Fishman *, Principles of Discrete Event Simulation*, John Wiley & Sons, Inc. New York, NY, USA ,1978. http://www.amazon.com/Discrete-Event-Simulation-George-S-Fishman/dp/0387951601/ref=sr_1_2?ie=UTF8&s=books&qid=1226080383&sr=1-2

[3] Jerry Banks, John Carson, Barry L. Nelson, David Nicol*, Discrete-Event System Simulation,* 4th Edition, Prentice-Hall International Series in Industrial and Systems, 2004. http://www.amazon.com/Discrete-Event-Simulation-Prentice-Hall-International-Industrial/dp/0131446797/ref=sr_1_1?ie=UTF8&s=books&qid=1226080383&sr=1-1

[4] Leonard Kleinrock*, Queueing Systems. Volume 1: Theory*, and *Volume 2: Computer Applications*, Wiley Inter-Science, 1975. http://www.amazon.com/Queueing-Systems-Theory-Leonard-Kleinrock/dp/0471491101/ref=sr_1_2?ie=UTF8&s=books&qid=1226080601&sr=1-2

[5] Jorge Nocedal and Stephen Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, 2006. http://www.amazon.com/Numerical-Optimization-Operations-Financial-Engineering/dp/0387303030/ref=sr_1_2?ie=UTF8&s=books&qid=1226080756&sr=1-2

[6] R. Fletcher, *Practical Methods of Optimization*, Wiley, 2000. http://www.amazon.com/Practical-Methods-Optimization-R-Fletcher/dp/0471494631/ref=sr_1_3?ie=UTF8&s=books&qid=1226080756&sr=1-3

[7] Tom Fawcett, ROC *Graphs: Notes and Practical Considerations for Researchers*, HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto, CA 94304

[8] Tapas Kanungo and Robert M. Haralick , *Receiver Operating Characteristic Curves And Optimal Bayesian Operating Points*, Intelligent Systems Laboratory, Department of Electrical Engineering, University of Washington, Seattle, WA 98195

[9] Mithat Gonen, Analyzing Receiver Operating Characteristic Curves with SAS, SAS Institute, 2007