WHITE PAPER | APRIL 2017

Using Models for 3-D Secure Authentication Across a Real-Time Network

Models that identify and act on fraud patterns in real time leveraging both card and device data across a network of participating card issuers.

Paul Dulany, VP of Data Science
Hongrui Gong, Master Data Scientist
CA Technologies, Advanced Analytics and Data Science

ca technologies

Table of Contents

Challenge

The battle against global debit and credit card fraud may sometimes feel like an uphill battle. Card issuers continue to find themselves victims of ongoing attacks perpetrated by fraudsters looking to rapidly exploit any opening they can find in e-commerce transactions. At the same time, fraudsters understand that issuers are intent on reducing both fraud and time to detection. That said, a fraudster's goal is to steal card information as fast as possible and either use it himself or sell it.

The underground market—or the so-called darknet—has become the location of choice from which many of these illicit transactions originate. Here, fraudsters often sell stolen card data to different buyers, who in turn automate their attacks to get the most value from their purchase in the shortest amount of time before the card is blocked by the bank. In the process, fraudsters could use either the same (or multiple) cards to make transactions from only a few devices. Traditional analytical models tend to focus on historical card information only, rather than also assessing related transaction information via other dimensions as they occur. Authorization models are also unable to capture the information coming from a device level. This limits the ability to appropriately and accurately score and flag risky transactions like in the example given.

At the same time, fraudsters often exploit cards from multiple issuers, which makes analyzing the behavior on devices a key aspect. While a card or BIN-level analysis of behaviors cannot address this situation, device-level analyses can. Creating a model that uses real-time device information from multiple issuers allows these issuers to work together to enable more accurate fraud detection with fewer instances of false positives, while keeping the raw transactional information confidential.

Opportunity

The 3-D Secure protocol is currently undergoing advancements (e.g., EMVCo launches EMV 3-D Secure 2.0) that offer issuers the opportunity to leverage more advanced analytics capabilities for authentication. This includes mobile capabilities that theoretically will improve the volume of transactions coming from multiple devices, such as tablets and smartphones.

Because of these advancements, support for our clients extends to:

- The analysis of app-based purchases on mobile and other consumer devices, as well as traditional browser-based devices.
- Intelligent, risk-based decisioning that encourages frictionless consumer authentication.
- Up-to-the-millisecond fraud detection and prevention across global banking networks.

These capabilities are embedded within the CA Risk Analytics Network. The CA Risk Analytics Network is a SaaS, data-driven solution that acts on fraudulent transactions in true real time by leveraging transactional data from a consortium of participating network members. Instead of using transaction data from just one bank, the real-time network allows the sharing of fraud data across participating member banks. This data sharing among the network community effectively prevents both known and suspected fraud, faster.

As one of its key capabilities, the CA Risk Analytics Network employs a self-learning model capable of analyzing and comparing multiple dimensions of large-scale data, both recent and historical, across banks and geographies. By doing so, it helps 3-D Secure issuers detect anomalous behaviors for the cardholder, the device or the combination of the two. For instance, by employing a model that can conduct both a card- and device-based analysis and share data across participating network members, issuers can immediately identify and act upon a fraudster attempting to make purchases using a batch of cards from multiple issuers on the same set of devices. With the newfound capability of making even more intelligent decisions on e-commerce transactions, card issuers can reduce more e-commerce fraud losses, false positives and transaction abandonment.

Benefits

Issuers can provide greater protection against fraud and an improved customer experience for their cardholders by using real-time neural network models that give a broad view of the connections among card and device activities as they are happening. Thanks to the SaaS environment—and the sophisticated variables created by CA for use in the models—this real-time update of device behavior, viewed in the context of the cardholder behavior, provides greater clarity regarding legitimate and fraudulent transactions.

For example, there are times when more than one fraud ring may have bought the same batch of cards and may be attacking them separately. When card issuers use models that view transactions in multiple dimensions and uncover the suspicious behavior of one ring, they may be able to stop the other ring more quickly. Consequently, fraudsters will find little to no success in committing fraud as they jump from card to card using the same device, and cardholders who normally use multiple cards on their devices will not trigger as many false positives; genuine customers will be able to complete transactions with minimal friction and disruption.

This enhanced customer experience can help boost top-line revenue, while improved fraud detection reduces potential losses for better bottom-line results. And benefits like these can extend to all network participants thanks to a real-time model that enables clients to rapidly share behavioral information through the model.

The Need for Real-Time Risk Analytics

To date, predictive models for risk analytics based on 3-D Secure 1.x have focused on the use of distilled behaviors keyed upon the cardholder transaction data as the means to help issuers reduce fraud. For instance, issuers already can view the merchants with whom a cardholder does business, the frequency of a cardholder's purchases, and the devices and IP addresses associated with that specific card. While this allows the model to understand and assess certain patterns of behavior for individual cardholders, it restricts analysis to the issuer's card data and does not provide visibility across cards within a bank or across multiple banks.

At the same time, most models today do not account for the fact that both legitimate and fraudulent users are increasingly making purchases in-app or through smartphones—anywhere and at any time. These advancements in payment technology have added another layer of complexity to risk assessments. That's why issuers need to be able to fully understand what kind of patterns and behaviors are normal for today's consumers due to the changes in cardholder device usage patterns and behaviors.

This includes grasping the importance of real-time updates given the global surge in automated attacks and resulting fraud. In the event that there is even a one-minute delay in updating device information, critical time and transactions can be lost. Updating in real time ensures that each transaction understands the up-to-the-millisecond information on the device, as well as the card. This is something that can't be achieved by other fraud network offerings that batch up summaries on a nightly basis, nor by on-premises authorization systems that inherently restrict this kind of cooperation.

To address these concerns, the real-time network model augments a traditional model by enabling it to track behavior patterns on both cards and devices in real time, examining the device, regardless of the card issuer, for a given transaction. This approach, which incorporates self-learning techniques, allows for more accurate detection of fraudulent and legitimate patterns across a network of banks and geographies. Given the global nature of fraud, as can be shown in historical data, this adds a significant dimension to an issuer's defensive measures.

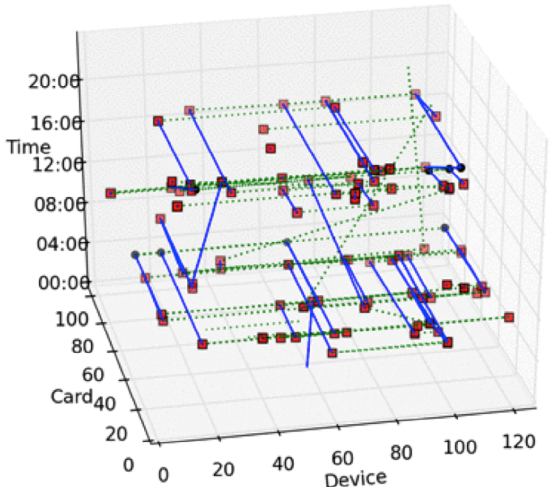
Let's take a closer look at how this approach can help issuers take full advantage of the data available to them. This approach requires a SaaS environment and provides significant new techniques to ensure proper handling of the difficulties introduced by using real-time, cross-issuer device distillates, as well as cooperation among issuers through their trusted third party: CA Technologies.

Making Use of Connected Information

When associated cardholder transaction data and device data are disconnected in analytical processes, the potential to more accurately detect and flag instances of fraud remains untapped. And as the connectivity of information in the real world continues to increase, so will related risks from fraudsters if models continue to rely solely on one-dimensional perspectives.

For this reason, it's important to extend 3-D Secure model capabilities to include orthogonal views of information, such as device and card dimensions. In this way, models can see all the connections among a device and the card or multiple cards that are transacting on it. Likewise, they can view all the different devices that a single card interacts with. Typically, this data reveals what can be referred to as a many-to-many relationship, which provides a better scope of fraud across bank or network members.

Figure 1: Connecting card and device dimensions to facilitate fraud detection.



The CA Risk Analytics Network also allows issuers to share connected information across multiple institutions, so that device behaviors on Bank B can inform the predictive scores for Bank C and Bank D, enhancing the accuracy of detecting risk factors.

One can extend these views to look at other high-cardinality, cross-issuer entities, like IP address, merchant or device country. It's useful to complement device information with an IP address because of situations where information used to identify a device can be changed or is unavailable—leading to a new device ID—whereas an IP address may be stable and consistent over that time frame. In other words, by gathering anonymous behavioral information at both the device and IP address levels, the model has some redundancy when one key or another is being disrupted for any reason. Note that like device behavioral information, real-time updating of IP address behavioral information is also possible.

In addition, one may find information from entities like merchants and/or device countries to be valuable for model usage and the identification of risky transactions. Given the potentially high transaction volumes of an individual merchant and device country, real-time updating of the information on these entities may not be necessary. Instead, processing on a near-real-time, hourly or daily basis may be good enough to provide the needed behaviors for the model while reducing the burden on the real-time scoring system in production.

Learning from connections

It's important to note that we are not suggesting that information from entities (like device, IP and merchant) be taken advantage of in a rules-based approach where only simple counts—such as device velocity—can be implemented. However, an augmented 3-D Secure model that gathers and uses behavioral information to determine which transactions are risky and which aren't can fully leverage this information in a secure, PII-compliant way.

This means that the model will be trained to analyze devices that are using multiple cards from multiple issuers, some of which may just be power users of payment cards, including those who have debit and credit accounts at the same bank or even at different banks. The model learns to distinguish normal patterns from fraudulent patterns by combining all the real-time information available into an optimized view of fraud and non-fraud in the high-dimensional space of the transaction and the histories of the given card and device. All the while, the model uses that data in a secure area where the raw information is never shared or visible across clients.

Addressing Real-Time Scoring, Model Development and Other Challenges

There are major technical challenges in performing updates in real time, both in production and in the training environment.

Real-time scoring

In real-time scoring during production, there are significantly more database tables that store historical information on multiple entities, and they must be updated for each and every transaction. This process involves table-fetching; locking, updating and unlocking—with each step impacting system performance. Also, with more fraud patterns to look up and compare from static tables, and more complicated model variables to be computed, measures must be taken to ensure acceptable response times during real-time scoring.

Model development

In model training environments, a prime consideration involves how to properly partition data into train, validation and holdout datasets when crossing the device and the card, which can often lead to making the whole dataset irreducible. Data scientists developing the model must deal with this carefully to avoid leaking of information, as improper partitioning can lead to inaccurate estimates of the model's ability to generalize to new datasets.

Also, as opposed to the use of one-dimensional historical card data, the need to enforce causality along multiple dimensions creates some difficulties in the parallelization process for feature creation on the data set as a whole. Standard techniques for parallelizing within a given data set no longer hold because the data cannot be partitioned cleanly on orthogonal dimensions. The irreducibility of this data means that it is no longer possible to subset the data into groups of transactions that have no dependency on any other subset. Without this ability to subset, standard sampling techniques and parallelization techniques become untenable. What's more, I/O limitations make using a standard database unattractive for processing a few years of data in days. However, today there are other larger-memory configuration options available to address this, thanks to larger disks and modern, distributed computing.

So, while it's valuable to make use of connected information as discussed earlier, doing so actually creates difficulties in model development. This is because each time the model connects with a new dimension, such as an IP address, it creates more connections within the entire data set.

Data scientists can address this complexity by grouping data in an organized fashion, and splitting the data in time for the creation of training and holdout datasets. But before doing so, it's essential to have a long-enough history of data with which to work; without large datasets for training, validation and holdout, the problem could be isolated in time such that the model doesn't generalize over the natural ebb and flow of a year. For example, an issuer could use 2015 as training data, 2016 1Q as evaluation/validation and 2016 2Q as holdout data. In this way, transactions in the training data from 2015 are not seen in the transactions from 2016 Q1.

When a data scientist chooses to develop a model by using training, validation and holdout data whose transactions are separated by time, it can be instructive to swap the ordering of those data around to see what a particular model build reveals in terms of variables sensitive to a specific time ordering. For instance, if the training and validation data are from a pre-EMV time frame, and the holdout is from a post-EMV go-live, the model may not generalize properly. Performing a time-delimited cross validation diminishes this risk, allowing the data scientist to remove those time-sensitive variables and avoid damage to the model in the future.

However, when a data scientist opts to create and run model builds, it's important to have truly isolated holdout data. Otherwise, there's no way to understand how well the model was built from data it knows to data it's never seen before.

Onboarding new network members

A key aspect for the production system is that the subset of transactions seen on a device may change significantly over time based on the inclusion or exclusion of issuers in the production system. This is not caused by any cardholder or fraudster behavior, and therefore it is important that the model be invariant under the inclusion of a new issuer.

Also, be aware that model variables going into the final model to produce the risk score are usually based on either card history, device history or both. The variables based on the card history will not be affected by new issuers joining the network because card numbers are unique irrespective of issuer. But model variables based on the device or the combined device and card histories could be sensitive to the volume changes on devices due to the onboarding of new clients.

For example, imagine that a data scientist is creating a simple daily velocity on a device, seeing approximately 25 percent of all transactions in a market. Now, sign on several new clients that go live the same week, such that the representation increases to 50 percent of all transactions in the market. By a simple daily count, there may be devices for which the count doubles overnight. This clearly isn't due to a change in behavior, but rather a change in the coverage.

Ultimately, there should be no instability in volumes or other negative impact resulting from the dependence that arises because the network is now going across multiple issuers. Instead, the issuers involved should benefit in terms of reduced false positive and increased fraud detection.

3-D Secure 2.0.

As a leading provider of 3-D Secure solutions, CA Technologies is the ideal partner to help issuers migrate to 3-D Secure 2.0. With the current 3-D Secure payment security solutions from CA Technologies, issuers can achieve a powerful CNP payments foundation, create a more seamless customer experience and confidently prepare themselves for 3-D Secure 2.0. CA will continue to support both the current and new 3-D Secure protocols so that issuers can easily accept transactions from all merchants, regardless of which version of 3-D Secure they employ. Here are some recommendations that organizations should consider when getting ready for 3-D Secure 2.0:

- Prepare for a consistent customer journey.
- Adopt an analytics-driven approach to risk-based authentication.
- Adopt user-friendly methods for stronger authentication, such as push notification and one-time passwords.

Solutions

As mentioned previously, the onboarding of new issuers to the network may change the volume of transactions on the same device. This situation may cause some model variables—based in part or in whole on device history—to deviate from previous behavior, which may lead to an undesired risk score. To be effective, it's critical for data scientists to ensure that model variables be invariant to the sudden volume changes resulting from normal client onboarding, which should not be viewed as a change in normal behavior. There are two approaches that data scientists can take to deal with this issue.

The first approach is to create and use model variables that are not sensitive to a sudden change of transaction volumes on the same devices. During model development, scientists should carefully create the model variables and verify them after creation by analyzing the value distribution of variables over the data, with new organizations added midway through the model development data time span. This approach is easy to implement both in model development and in production, and model performance will be appreciably improved over a model only based on card history. However, because the type of model variables that can be used on device history is limited, the model may not have the full benefit of all device-history information.

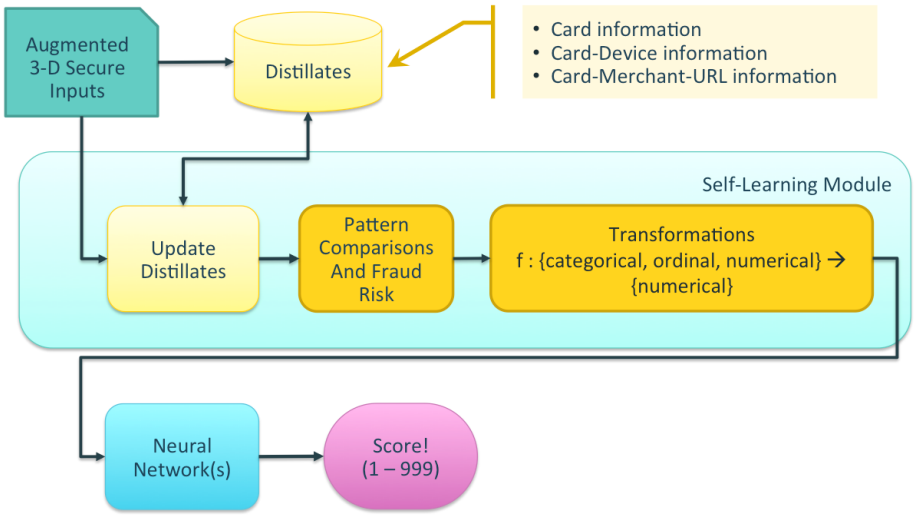
The other approach, which allows data scientists to achieve the full potential of the model when including the device history, is to create all necessary model variables when making use of the device-history information, disregarding the limitation on the type of model variables.

To make the model variables invariant to the sudden change of the volumes due to the onboarding of a new client, data scientists need to implement a bias-correction mechanism both in model development and similarly in model scoring in production. This mechanism involves either a near-real-time or batch process to continuously track all required information on an organizational level or multiple levels. The model can use this information to perform the bias correction on model variables as needed. The reason for the use of batching stems from the fact that onboarding a new organization happens infrequently (typically every few days, weeks or even months).

The following steps describe the batching process in production (the process in model development should be flowed in a similar way):

1. The scoring platform, within the SaaS model, initiates a batch process in the scoring system every few hours to accumulate all the transactions and model variable values based on device history for each transaction (which has been updated in real time), storing them either in memory cache or database tables. If data scientists use memory cache for this purpose, they must ensure it will be fault tolerant in the event of sudden outages or other disruptions.
2. For each organization, the scoring system calculates the volume and the age of the organization, as well as model variable distributions (e.g., mean, standard deviation, mode, skewness).
3. At the end of the day, the scoring system summarizes the volume, age and model variable statistics for that day from the statistics on all the batches during the day—for each organization and all populations. The scoring system stores the daily statistics in its database.
4. The scoring system retains at least the last 30 days of daily statistics information—pushing the newest day in and popping the oldest day out. The length may be configurable based on the variables in the model.
5. The model compares the current day's statistics with the daily statistics information to identify if there are new organizations onboarding that are changing the overall statistics or distributions of the model variables.
6. If the comparison determines that the onboarding of new organizations is responsible for changing the statistics of the model variables, then a bias-correction factor for each model variable based on the device will be computed based on the daily statistics information, and be stored for the model-scoring process to pick up.
7. The model-scoring process checks the availability of the new bias-correction factors and normalizes the model variables using them. The scoring system then calculates the model score from the normalized model variables.

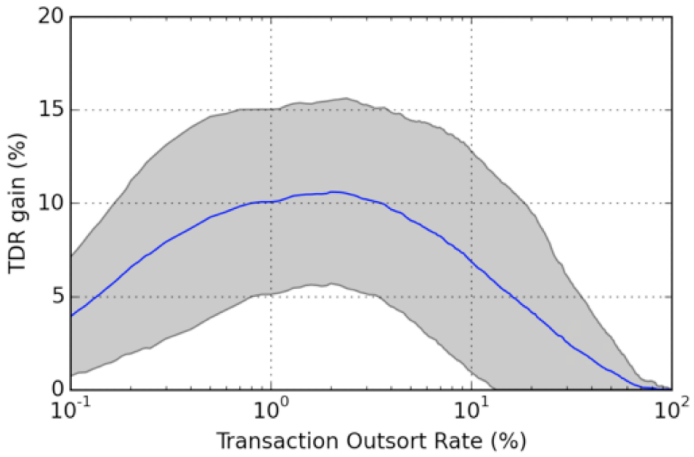
Figure 2: Model scoring under the new 3-D Secure 2.0 protocol helps unlock issuers' full potential to detect fraud.



Benefits

Including the device history, as opposed to just card history, in the model greatly improves model performance: by providing a better separation of fraud and nonfraud transactions, the model is better at both detecting fraudulent transactions and reducing fraud loss for clients. The following plot (see Figure 3) shows the typical fraud detection improvement of the real-time network model over the model based on card history only.

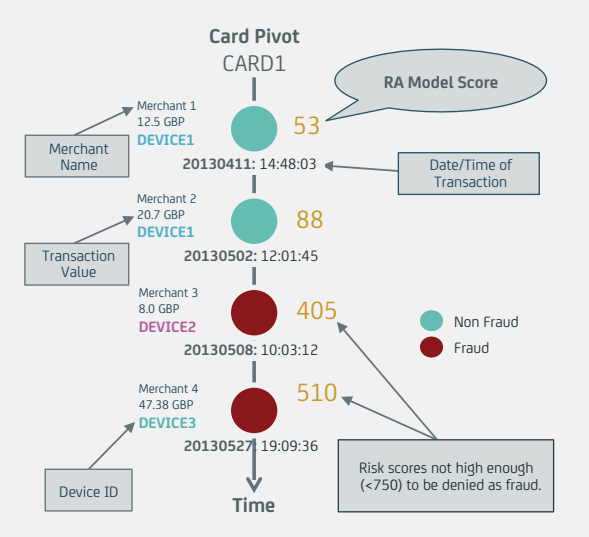
Figure 3: Net gain of real-time network model over model based on card only.



The real-time network model on average enables the fraud detection of an extra 10 percent of all fraud when the model is used to deny or challenge between 0.8 to 3 percent of all transactions, which is often the ideal operating range for many clients. And this extra 10 percent fraud detection can be quite significant for issuers in terms of reducing fraud and associated financial losses.

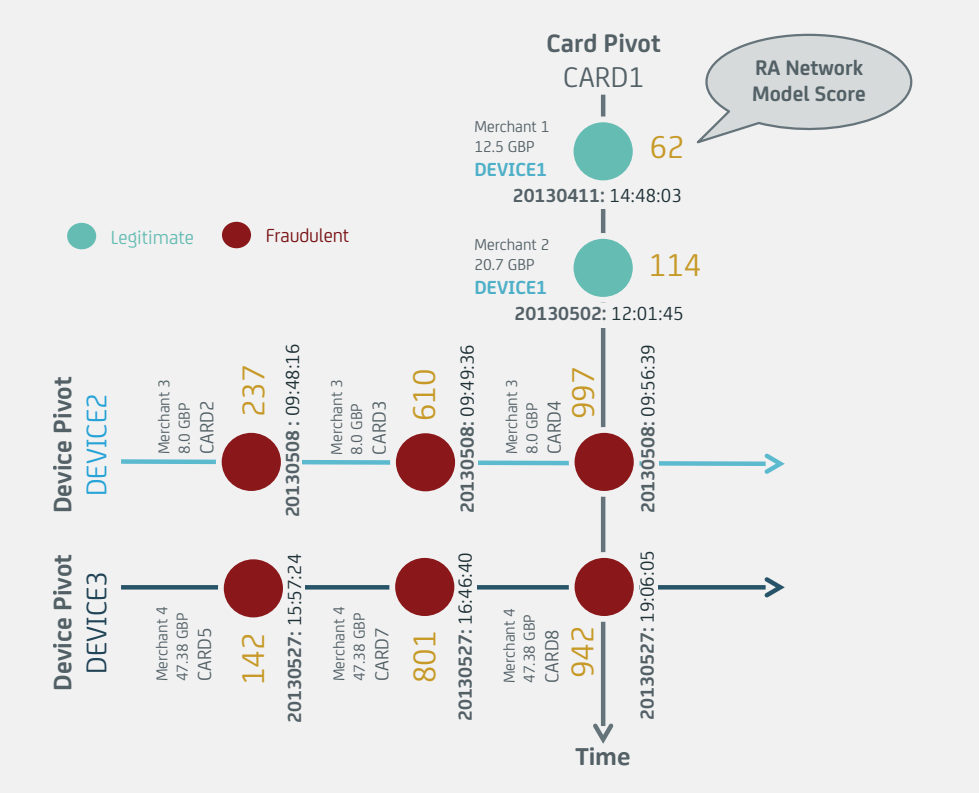
How is this achieved? Looking at an example from the data, Figure 4a shows a series of transactions on a single card as well as their risk score, using a model that leverages only data from the card pivot. The green circles represent legitimate transactions (i.e., those not later reported as fraud), and have such a low risk score that (depending on the issuer's fraud thresholds) they should pass through without any additional authentication. The red circles represent fraudulent transactions (i.e., those later reported as fraud), but would also typically go through without added authentication.

Figure 4a: Fraud detection using a card-only model.



Let us assume that, model scores from 500 to 749 might be challenged, and 750 and above might be denied. Using this as a rubric, the model used in Figure 4a might have only challenged the last transaction, allowing the first three to progress with no intervention. Even the fourth transaction, depending upon the challenge method and the sophistication of the fraudsters, might pass the challenge and result in a loss.

Figure 4b: Improved fraud detection via a risk-analytics-network model.



In contrast, Figure 4b uses the same data from Figure 4a, but puts it through the risk analytics real-time network. By looking at the horizontal arrows representing the device pivots, you begin to see that there have been other transactions coming from these devices that are coming from a number of different cards. And in fact, the real-time network model starts to score those high. So because we've added in the device-level assessment, transactions from Card1 suddenly jump to the 900s, which is typically identified as fraud. Therefore, it should be evident that this model denies these transactions outright without allowing the fraudster to even attempt to get past a challenge.

Top-line and bottom-line business improvements are also possible using the real-time network model. Depending on issuers, business requirements, they have the freedom to use the model to deny or challenge a smaller population and allow more transactions to go through in order to improve users' purchasing experiences. By doing so, the issuer can decrease fraud while increasing the approved transaction volume, which resulted from lower abandonment and failed authentication rates due to the lower challenge rate from the model.

Subsequently, issuers can benefit from earning more interchange fees and interest (from revolving balances) on more approved transactions. Lower deny/abandonment/failed authentication rates also lead to lower operating costs by reducing the volume of inbound calls users make to resolve blocked purchase problems.

Conclusion

For fraudsters, successful attacks on debit and credit card transactions can deliver a major windfall. So, while they continually devise new ways to avoid detection, issuers need to act just as diligently to thwart their efforts. By leveraging solutions that employ real-time multidimensional capabilities, they can move beyond the analysis of historical cardholder data to an approach that also connects transaction data with devices in real time. Through this multidimensional analysis, issuers can reduce fraud faster and prevent related disruptions to the online customer experience.

As more issuers participate and contribute to the network, the model can readily address the resulting surge in transactions with bias correction mechanisms that help ensure proper risk scoring. Just as important, the solution is designed to offer ongoing protection against fraud as issuers add in different fields and subsets of transactions in both the card and network space. Through capabilities like these, issuers can significantly cut fraud-related losses and risks while arming themselves to stay ahead of the evolving fraud landscape.

About the Authors

Paul Dulany has been in the advanced analytics and data science areas for 18 years. He joined CA Technologies in 2013, and led the development of the analytical modeling infrastructure, as well as the first model produced by the CA data science team, and has been leading the team since late 2015. Prior to joining CA Technologies, he was at the SAS Institute for over eight years, where he was on the team that developed the first models for the SAS Enterprise Fraud Management solution, led the development of the first debit card models and developed many new techniques. Prior to SAS, Paul was at HNC and Fair Isaac for more than five years, as a scientist and later as the manager of the Fraud Predictor modeling team, developing a number of Falcon payment card models as well as working in other areas. Paul holds patents from his time at CA Technologies, HNC and SAS, and has a Ph.D. in theoretical physics.

Hongrui Gong has extensive experience in the areas of advanced analytics and data science. He joined CA Technologies in April 2013 and played a key role in the efforts to build a modeling infrastructure and in developing models for 3-D Secure products. Prior to joining CA Technologies, he worked for more than 15 years with prominent analytic companies (SAS, FICO and HNC) to develop models for products, such as payment card fraud detection, insurance fraud detection, tax under-filers identification for federal and state government, anti-money laundry, loan loss forecast, brokerage margin landing risk management, and credit risk rating for public and private companies. Hongrui has a Ph.D. in computational fluid dynamics and spent four years in Los Alamos National Laboratory focusing on the research of theoretical modeling and computer simulations of turbulent fluid flow. He holds a number of patents from his prior work.

Contact CA Technologies

We welcome your questions, comments and general feedback.

For more information, please visit **ca.com**.

Connect with CA Technologies at ca.com

CA Technologies (NASDAQ: CA) creates software that fuels transformation for companies and enables them to seize the opportunities of the application economy. Software is at the heart of every business, in every industry. From planning to development to management and security, CA is working with companies worldwide to change the way we live, transact and communicate—across mobile, private and public cloud, distributed and mainframe environments. Learn more at **ca.com**.

CS200-238009_0417