Fighting back against synthetic identity fraud

Digging deep into the data trails people leave behind can help banks detect whether their customers are real or not and stem losses from this fast-growing financial crime.

Bryan Richardson and Derek Waldron

JANUARY 2019 • RISK PRACTICE

Banks have become much more effective at preventing many types of fraud thanks to their investments in technology, but criminality has evolved in response. Rather than using a stolen credit card or identity (ID), many fraudsters now use fictitious, synthetic IDs to draw credit. Indeed, by our estimates, synthetic ID fraud is the fastest-growing type of financial crime in the United States, accounting for 10 to 15 percent of charge-offs in a typical unsecured lending portfolio.[1] Instances of synthetic ID fraud have also recently been reported in other geographies.[2] More worrying still, much bigger losses are building up behind these IDs like hidden time bombs.

That risk is because of the way the fraudsters typically operate. Over months, if not years, they build up a good credit record with synthetic IDs. Only when the credit lines are maximized do repayments cease—or, in the jargon of the business, do the synthetic IDs "bust out." Fraud rings sometimes establish thousands of synthetic IDs, all waiting to default. The largest synthetic ID ring detected to date racked up losses for banks of $200 million from 7,000 synthetic IDs and 25,000 credit cards.[3]

To date, there has been no efficient way of uncovering synthetic ID fraud. To crack down on it, every customer seeking credit would have to undergo even more rigorous ID checks than they do already. This article proposes a new approach that, with the help of machine learning, digs deep into vast amounts of third-party data to gauge whether the basic information given by an applicant matches that of a real person, thereby weeding out the small proportion of those likely to be using a synthetic ID. It is on this group that banks, or indeed any organization wanting to stop synthetic ID fraud, can focus their ID checks without inconveniencing other customers.

The scam

Synthetic IDs are created by applying for credit using a combination of real and fake, or sometimes entirely fake, information. The application is typically rejected because the credit bureau cannot match the name in its records. However, the act of applying for credit automatically creates a credit file at the bureau in the name of the synthetic ID, so the fraudster can now set up accounts in this name and begin to build credit. The fact that the credit file looks identical to those of many real people who are just starting to build their credit record—that is, there is limited or no credit history—makes the scam nearly impossible to detect.

The question that springs to mind is, Why do financial institutions fail to conduct additional, more rigorous screening to identify synthetic IDs when onboarding new customers? In the United States, a large part of the problem is that there is no efficient government process to confirm whether a Social Security number, date of birth, or name is real. And although the government is developing a service to address this, the release date and precise capabilities remain unclear.[4]

The sophisticated technology that has helped detect other types of fraud is not of much assistance. Machine-learning techniques such as deep neural networks that find patterns associated with fraud are of little use, because so few cases of synthetic ID fraud have been uncovered on which to train models. Unsupervised machine-learning techniques that look for anomalies in data also struggle, because there are few, if any, differences between real and synthetic IDs at the time of application.

This leaves financial institutions having to conduct their own additional—and sometimes intrusive—checks, slowing an already complex onboarding process. The danger becomes that banks deter not

only the fraudsters but also the very customers they wish to attract, who may well turn to competitors instead.

How extra data helps

An approach to identifying synthetic IDs that entails leveraging third-party data can be a powerful tool. It is grounded in the fact that real people have real histories, evidence of which they scatter behind them in dozens of different data systems, physical and digital. These trails are hard to fake. They have depth—that is, large amounts of data that stretch back years. For example, a real teacher might have a student loan taken out ten years ago, a social-media account, a cell-phone record, a couple of past employers, several previous addresses, an email account set up years ago, and property records. The trails of real people are also consistent: the same address, email account, and phone number crop up in various databases. Synthetic IDs tend to be inconsistent, because although the applicant may give some real details (perhaps a name that reoccurs in various data systems), others are fabricated, so they will not reoccur. In cases in which the synthetic ID is entirely fabricated, the ID may be too consistent—that is, there are no changes at all to the address, email account, and other data over several years.

A rich demonstration

By evaluating the depth and consistency of information available about applicants in third-party data systems, institutions can determine whether the applicants are real or not. McKinsey undertook research to demonstrate the efficacy of this approach. While adhering to all applicable privacy regulations, we used a sample of 15,000 profiles gathered from a consumer-marketing database (exhibit):

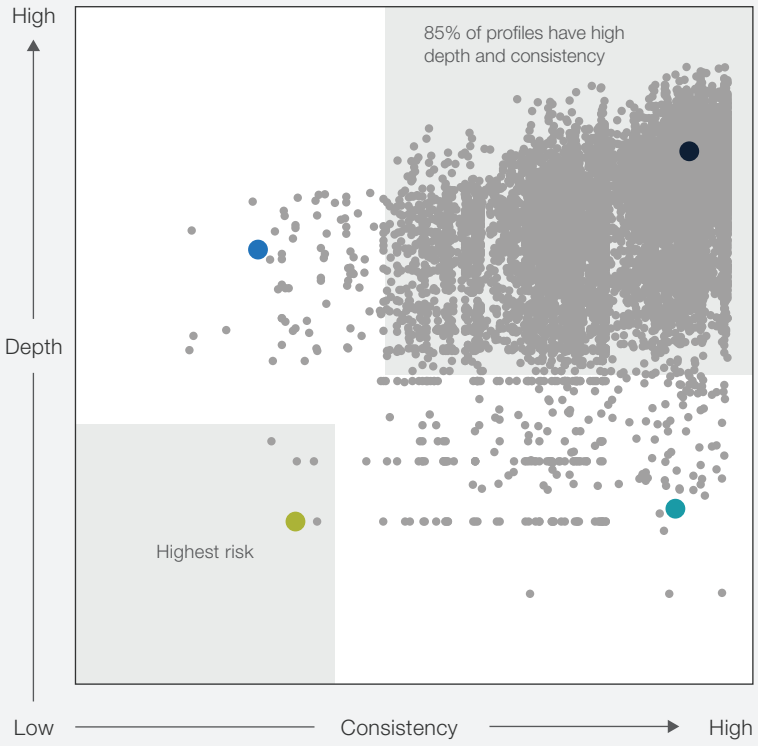
- We used nine external data sources to check and augment the data in each profile, looking at social-media accounts, email addresses, mobile-phone and landline numbers, financial behavior, property records, and other information. The nine sources chosen were those with the most digital and nondigital information that matched our sample group. The sources yielded more than 22,000 unique fields of information.
- We then identified some 150 features that served as measures of a profile's depth and consistency that could be applied to all 15,000 people. (The fact that there were so many suitable measures illustrates the wealth of relevant external data available.) The features related to depth included the age of a first loan, age of the oldest recorded nondigital event (a vehicle registration, for example), and age of an email address. Features related to consistency included matches of unique names with the same data in many sources, as well as reverse matches of particular data points (such as addresses and phone numbers) leading back to the same name.
- An overall depth and consistency score was then calculated for each ID. The lower the score, the higher the risk of a synthetic ID.

By evaluating the depth and consistency of information available about applicants in third-party data systems, institutions can determine whether the applicants are real or not.

Exhibit

From nine sources of external data, McKinsey researchers determined the likely authenticity of identities based on data depth and consistency.

Matrix for scoring profile depth and consistency



Arturo
High consistency, high depth

Cheryl
Low consistency, high depth

Maria
High consistency, low depth

John
Low consistency, low depth

Consistency				
Unique names	1	2	1	2
Unique numbers	1	1	1	2
Suspicious indicators?	No	No	No	Yes
Depth				
Age of nondigital data	25 years	8 years	15 years	<1 year
Nondigital data	12 records	5 records	1 record	0 records
Email age	9 years	9 months	<1 year	<3 months

Note: Under consistency and depth, only top three features are listed.

- For some identities, low depth or consistency scores clearly did not indicate high-risk profiles. Someone fresh out of school may well have a new email address, for example. A suite of machine-learning models was used to take account of such anomalies and adjust overall scores accordingly.

The final results of our demonstration showed that 85 percent of the profiles we examined had high depth and consistency, and a further 10 percent fell just outside the normal range. The remaining 5 percent, as depicted in the lower left-hand quadrant of the exhibit, were profiles that would raise suspicions. "John," for example, has two different names linked to the same phone number, his email is fewer than three months old, and the age of his oldest nondigital record is less than a year.

If armed with similar scoring systems, banks could ascertain whether an applicant's profile looked real. They could then instantly extend credit, perhaps limited, to those applicants with high depth and consistency scores. They could even offer higher initial credit limits than would normally be the case for first loans, since low-risk applicants could be distinguished from high-risk ones.

Very limited credit, or none, would be extended to high-risk applicants while their IDs were reviewed more thoroughly with the help of a range of processes, such as in-person verification of documents and third-party income verification, as well as increasingly sophisticated tools. These tools include biometric screening that matches a face to a photo on a driver's license or passport, voice identification that assigns the unique voiceprint of a customer to a Social Security number, and geospatial technology that confirms whether a customer's application was made from the stated address. Some checks are less obtrusive than others, and it may be wise to conduct these first. That said, many customers understand and appreciate banks' efforts to reduce fraud.

Importantly, banks could also review existing accounts to avoid any further buildup of debt through synthetic IDs. High-risk accounts would require extra ID checks; in the meantime, additional credit would be denied or limited.

Next steps

Chances are, if your onboarding processes for customers applying for credit do not include in-person verification of documents or biometric screening, you are exposed to synthetic ID fraud. The extent of that exposure is harder to gauge, as even the most sophisticated banks struggle to know whether an unpaid debt is a result of synthetic ID fraud, another type of fraud, or simply a customer who cannot pay. One approach is to look for charge-offs that resemble synthetic ID fraud—for example, those that occurred fewer than two years after the account was opened, had minimal account activity, and for which there was no customer contact once credit limits were reached. The results are likely to spur you to further action.

If so, assemble a team of data scientists, compliance experts, and fraud experts to gather third-party data and develop a synthetic ID risk model. A good one will be built from external data sources that have a good match rate. For example, an online bank will likely find plenty of additional information on applicants in social-media data. Banks whose customers have an older demographic will find information on property ownership helpful. The model will also have good-quality data, and all data will adhere to privacy regulations. So test multiple external data providers. Remember, too, that while machine learning can help sort through the data and formulate models, risk-model managers need to validate them. If the models and data introduce bias or incorrect information, they can be riskier than the fraud that companies seek to mitigate.

Finally, when it comes to deployment, test any changes you choose to make to the customer-onboarding process as a result of the model's findings on a sample of customers. You may find, for example, that the extra time it adds to the application process is unacceptably long, so you would have to rethink the design.

Fraud will continue to evolve to evade detection. However, by mining the growing number of third-party data sources available, banks can deepen their understanding of their customers. This knowledge can help banks enhance risk controls and stem losses associated with synthetic ID fraud—all without burdening the vast majority of honest customers with ever-more intrusive and time-consuming ID checks. ■

[1] AnnaMaria Andriotis and Peter Rudegeair, "The new ID theft: Millions of credit applicants who don't exist," *Wall Street Journal*, March 6, 2018, wsj.com.

[2] "'Synthetic' identity fraud costs Canada $1B a year," CBC/Radio-Canada, October 11, 2017, cbc.ca.

[3] "Eighteen people charged in international $200 million credit card fraud scam," US Department of Justice, February 5, 2013, justice.gov.

[4] The US government is building an application that will verify Social Security numbers, names, and dates of birth as part of the Economic Growth, Regulatory Relief, and Consumer Protection Act (S.2155).

Bryan Richardson is a senior knowledge expert in McKinsey's Vancouver office, and **Derek Waldron** is a partner in the New York office.

The authors wish to thank DemystData, a comprehensive data-access company, for helping provide the data used in this article. The authors also wish to thank Kevin Buehler, Mark Hookey, Ivan Pyzow, and Shoan Joshi for their contributions to this article.