XI Konferencja Naukowa
Bezpieczeństwo w Internecie. Analityka danych
Warszawa 6-7.06.2019

Anonimizacja danych do analizy *(dane syntetyczne)*

dr inż. Maciej Kiedrowicz, Wydział Cybernetyki WAT

Urząd Ochrony Danych Osobowych

Naukowe Centrum Prawno-informatyczne

UKE

Przestępstwa gospodarcze i finansowe – skala i waga problemu

Według analiz *Association of Certified Fraud Examiners* firmy ponoszą straty rzędu 5% swoich dochodów spowodowane rozmaitymi oszustwami i przestępstwami gospodarczymi lub finansowymi. Co więcej, okazuje się, że bezpośrednio i już na wstępnym etapie rozpoznawania tego typu przestępstw pracownicy wysokiego szczebla nie posiadają adekwatnych narzędzi informatycznych, aby w porę wykryć i/lub zapobiec przestępstwom.

Jest to szczególnie widoczne w dzisiejszej epoce tzw. *Big Data*, gdzie ogromna masa danych opisujących scenariusze (domniemanych) przestępstw jest szansą dla (technologicznie) przygotowanych, a barierą dla nieprzygotowanych w monitorowaniu i prewencji.

Stąd rosnąca rola i skuteczność informatycznych metod zaawansowanej i wydajnej analizy danych.

Przestępstwa gospodarcze i finansowe – skala i waga problemu

Metody te pozwalają w bardzo krótkim czasie zebrać dane pochodzące z wielu rozproszonych źródeł i w rozmaitych formatach, zoptymalizować ich postać względem konkretnego rodzaju analizy – która może być wielowątkowa, wielowymiarowa i dotyczyć wielu obszarów tematycznych (tzw. analiza hybrydowa) – a następnie udostępnić wyniki analizy zainteresowanym stronom.

Jednym z przykładów mogą być dane (osobowe) pochodzące z mediów społecznościowych, billingów rozmów telefonicznych, transakcji z użyciem kart kredytowych, polis ubezpieczeniowych, a ostatnio również handlem z użyciem kryptowalut (np. bitcoin).

Dane te mogą przybierać rozmaite formy: oprócz dobrze zdefiniowanych danych ustrukturyzowanych (*structured data*), typowych dla baz danych, coraz większy udział zyskują dane w formie „luźnego” tekstu (*unstructured data*), gdzie zamiast predefiniowanych atrybutów występuje rozproszony, często nieformalny opis słowny.

Przestępstwa gospodarcze i finansowe – skala i waga problemu

Stosowanie tradycyjnych (informatycznych bądź nie) metod audytu wewnętrznego jest nie tylko żmudne i czasochłonne, lecz coraz częściej mało skuteczne, ze względu na liczne powiązania i wysoką częstotliwość procesów natury gospodarczej i finansowej.

Audytorzy na ogół wciąż nie posiadają adekwatnej wiedzy i doświadczenia pozwalających sprostać aktualnym wymaganiom w świecie, gdzie lawinowo powstają coraz to nowe metody oszustw i nadużyć a scenariusze tychże rozgrywają się nierzadko globalnie (a nie jak dawniej w jednym tylko państwie, czy nawet przedsiębiorstwie), błyskawicznie i angażują jednocześnie wiele osób oraz podmiotów.

Stąd potrzeba automatyzacji i wyspecjalizowanej tematycznie „maszynowej inteligencji". Szczególne znaczenie mają metody pozwalające na monitorowanie i rozpoznawanie scenariuszy przestępczych w czasie (prawie) rzeczywistym (*near real-time*) na podstawie dużej ilości danych oraz informacji z różnych (na ogół rozproszonych) źródeł.

Przestępstwa gospodarcze i finansowe – skala i waga problemu

Big data nie tylko umożliwiają stosowanie zaawansowanych metod analitycznych rodem z *machine learning* (m.in. drzewa decyzyjne, sieci neuronowe, klasyfikacja, klastrowanie, reguły asocjacyjne) i na ich podstawie z dużą skutecznością rozpoznawanie lub oszacowywanie prawdopodobieństwa wystąpienia nadużycia, lecz również są pomocne w podejmowaniu decyzji opartych o lepiej weryfikowalne przesłanki.

Pozwalają też wykrywać i definiować nowo powstające trendy i wzorce przestępstw, jednocześnie wzbogacając profesjonalną wiedzę audytorów i śledczych.

W tym kontekście nie dziwi pojawienie się nowej specjalności zawodowej pod nazwą *informatyka śledcza*, której przedmiotem działania jest analiza ryzyka związanego z pojawieniem się różnych form komunikacji elektronicznej i transakcji finansowych online.

Poufność danych i konieczność stosowania danych syntetycznych

Praktyka pokazuje, że dane rzeczywiste związane z przestępstwami gospodarczymi i finansowymi nie są publicznie dostępne w celach badawczych (ograniczenia prawne, obawy związane z utratą reputacji instytucji bądź korporacji). Jednym z remediów na tego typu ograniczenia jest wygenerowanie danych *syntetycznych* – czyli danych sztucznych.

Tak jak w naukach ścisłych, badacz (przestępstw gospodarczych) może wygenerować dane syntetyczne o identycznych lub podobnych właściwościach, jak dane oryginalne (rzeczywiste). Im wyższy poziom złożoności danych (syntetycznych), tym trudniejsze zastąpienie danych rzeczywistych danymi syntetycznymi. Dla jednej lub kilku charakterystyk wygenerowanie danych jest na ogół proste, jednak dla większej ich liczby zadanie się komplikuje. Istnieją dwie metody generowania danych:

- produkcja danych zgodnie z rozkładami statystycznymi,
- poprzez zdefiniowania i zbudowania modelu fizycznego wyjaśniającego obserwowany proces lub zachowanie w celu zastosowania tego modelu przy generowaniu danych.

Poufność danych i konieczność stosowania danych syntetycznych

Istotną cechą danych syntetycznych jest to, że mogą one reprezentować sytuacje dotychczas niespotykane w obserwowanej rzeczywistości. Na przykład, aplikacje do wykrywania włamań do systemów informatycznych (*intrusion detection*) testowane są przy użyciu danych syntetycznych.

Właściwie wygenerowane i dostatecznie wiernie „naśladujące” rzeczywistość dane syntetyczne umożliwiają badanie, czy i jak dana aplikacja (w trakcie testu) wykrywa sztucznie zainicjowane zagrożenie (np. włamanie do sieci).

W przypadku użycia danych rzeczywistych (niesyntetycznych) testowana aplikacja reagowałaby tylko na sytuacje zdefiniowane w trakcie uczenia (na danych z etykietami przypisanymi przez ekspertów).

Generowanie danych syntetycznych

Celem projektu było stworzenie procedur służących wygenerowaniu dużej ilości danych syntetycznych umożliwiających ocenę wybranych metod analitycznych w dziedzinie przestępczości gospodarczej i finansowej.

Oczywistym i kluczowym aspektem stosowania danych syntetycznych jest zapewnienie poufności i prywatności osób fizycznych i prawnych (ogólniej: jakichkolwiek zdefiniowanych bytów). Dane syntetyczne nie zawierają (rzeczywistych) informacji o tych podmiotach pozwalających identyfikować je w sposób jednoznaczny.

Teoretycznie rzecz ujmując, koncepcja danych syntetycznych jest szczególnym przypadkiem danych anonimizowanych, których istotą jest gwarancja zapewnienia poufności.

Z praktycznego punktu widzenia dane syntetyczne można stosować do badań tam, gdzie istnieją surowe ograniczenia związane z poufnością informacji.

Generowanie danych syntetycznych

Na zakres projektu miały wpływ reguły biznesowe i przepisy prawne potencjalnych użytkowników i beneficjentów (gestorów). Narzucone zostały ograniczenia dotyczące dostępu do rzeczywistych danych reprezentujących konkretne osoby, transakcje, miejsca, przedmioty i rozmaite relacje między nimi.

Ponieważ (niejako z definicji) każde narzędzie informatyczne polega na przetwarzaniu danych (wejściowych), ich brak implikuje konieczność wygenerowania danych syntetycznych. Oczywiste jest, że im lepiej dane wygenerowane syntetycznie odwzorowują rzeczywistość, tym solidniejsze są podstawy do rzetelnej symulacji konkretnego stanu, procesu bądź scenariusza.

Dlatego symulacja wybranych scenariuszy poprzedzona była dokładnym rozeznaniem rzeczywistych charakterystyk natury demograficznej, prawnej, i specyfiką poszczególnych obszarów zastosowań (m.in. ubezpieczenia, rejestracja pojazdów, opłaty celne).

Generowanie danych syntetycznych

Rzeczywiste dane opisujące konkretne scenariusze biznesowe chronione były klauzulą poufności; nie mogły być udostępnione do tworzenia i testowania narzędzi analitycznych.

Dlatego doraźnym celem było *jednorazowe* stworzenie wirtualnych scenariuszy, gdzie możliwie wiernie odwzorowane byłyby rzeczywiste sytuacje biznesowe – w sensie demograficznym, społecznym i komercyjnym – w których mogło dojść do naruszenia prawa (tzn. przestępstwa gospodarczego lub finansowego).

Wirtualna przestrzeń osób, relacji między nimi oraz zdarzeń opisanych adekwatnymi atrybutami (parametrami) posłużyć miała jako materiał do przetestowania skuteczności wybranych metod informatycznych, mających pomóc w rozpoznawaniu i kategoryzowaniu predefiniowanych przestępstw.

Daje to możliwość przeprowadzenia wielokrotnych testów dla różnych liczebności adekwatnych populacji oraz tworzenia scenariuszy przestępstw trudno identyfikowalnych („nietypowych”).

Generowanie danych syntetycznych

Istotnymi założeniami w ramach tworzenia scenariuszy przestępstw były:

- stosowanie znanych *charakterystyk* (parametrów) istotnych ze względu na możliwie realne oddanie specyfiki scenariuszy, w szczególności struktury wiekowej populacji, zróżnicowania demograficznego (zaludnienie w rozpatrywanych obszarach), stopnia powiązań uczestników, dynamiki zdarzeń oraz ich związków przyczynowo-skutkowych,
- ograniczanie *złożoności* w celu umożliwienia szybkiej i dającej się ocenić implementacji,
- *rozmaitość kategorii* (scenariuszy) przestępstw, ażeby otrzymać dostatecznie bogaty obraz symulowanej rzeczywistości,
- tworzenie scenariuszy o kategoriach *mieszanych* i/lub *nietypowych*, co pozwoliłoby surowiej ocenić *precyzję* stosowanych metod analizy i wizualizacji,
- zapewnienie *adekwatnej wielkości* generowanych populacji, aby zminimalizować obecność przypadków trywialnych i umożliwić zastosowanie szerszego przedziału wartości stosowanych parametrów.

Charakterystyki, parametry i słowniki

Generowanie danych syntetycznych było poprzedzone rozeznaniem demograficznych charakterystyk ludności w Polsce, takich jak rozkład wg płci, wieku, liczby dzieci, migracji wewnętrznej, średniej długości życia, częstotliwości występowania imion i nazwisk itp.

Na podstawie określonych wyżej charakterystyk wygenerowana została wirtualna populacja miliona osób, gdzie każdej osobie przypisano fikcyjny numer PESEL oraz:

- płeć,
- datę urodzenia,
- imię, nazwisko, nazwisko rodowe (w przypadku kobiet),
- rodziców,
- adres zamieszkania,
- status małżeński,
- a także – dla wybranej części osób – niektóre dodatkowe informacje.

Charakterystyki, parametry i słowniki

Generując dane o osobach fizycznych do każdego losowo wygenerowanego numeru PESEL poczyniono następujące założenia:

- miejsca urodzenia były przypisywane losowo, jednak z prawdopodobieństwem faworyzującym miejscowości o dużej liczbie mieszkańców (zgodnie z danymi demograficznymi dla miejscowości w Polsce), jednocześnie nie wykluczając małych miejscowości;
- imiona i nazwiska były nadawane wg dostępnych informacji o ich częstotliwości występowania (wykorzystano wspomniane słowniki imion i nazwisk);
- wszystkie wygenerowane relacje rodzicielskie miały charakter biologiczny (w tej fazie nie generowano np. przypadków adopcji i konkubinatów);
- kobiety i dzieci przyjmowały nazwiska swoich – odpowiednio – małżonków i ojców.

Charakterystyki, parametry i słowniki

Dla pewnego podzbioru M < N wirtualnych osób fizycznych przyporządkowano co najmniej jeden z dwu identyfikatorów: NIP lub REGON, związanych z zarejestrowaną działalnością gospodarczą. Osoby te – zachowując charakter osób fizycznych – jednocześnie mogły być rozpatrywane jako osoby prawne w sytuacjach, kiedy reprezentowały firmy będące ich własnością (bądź współwłasnością).

W celu eliminacji oczywistych nieprawidłowości algorytm generowania zbioru PESEL posiadał zabezpieczenia walidacyjne – reguły wykluczające niektóre (predefiniowane) sytuacje, w szczególności: (i) nie można być dzieckiem i rodzicem tej samej osoby (realistyczne odstępy dat urodzenia), (ii) jeśli np. X, Y są kuzynami o wspólnych dziadkach – wówczas między X i Y nie może zachodzić relacja rodzicielska, (iii) związek małżeński między rodzeństwem, osobami tej samej płci, rodzica ze swoim dzieckiem – to tylko niektóre przykłady zastosowanych reguł.

Charakterystyki, parametry i słowniki

Zbiór KRS (podmioty gospodarcze) posiada identyfikatory: NIP, REGON, KRS, NrEDG, PESEL i został wygenerowany z użyciem predefiniowanych parametrów: liczba podmiotów gospodarczych – osób prawnych, maks. liczba udziałowców, stopień różnorodności zakresu aktywności biznesowej, maks. kapitał pieniężny, ułamek firm sprzedających udziały, ułamek udziałowców, min. wiek dla sprawowania funkcji, maks. wiek dla sprawowania funkcji, ułamek osób fizycznych prowadzących działalność gospodarczą oraz słowników – Polskie Kody Działalności (PKD), miejscowości, sprawowane funkcje pracownicze (m.in. prezes zarządu, prezes rady nadzorczej, dyrektor generalny, główny księgowy).

Funkcje dla tych osób prawnych (np. dyrektor, prezes, księgowy) obsadzane były spośród osób fizycznych, z uprzednio wygenerowanej populacji PESEL. Takie podejście spełniało założenie, że wszyscy potencjalni uczestnicy scenariuszy biznesowych (przestępczych bądź nie) pochodzą z traktowanej jako „zupełna” populacji PESEL.

Dane wygenerowane: zakres i struktura

W celu budowania scenariuszy przestępstw z udziałem wirtualnych osób fizycznych i prawnych utworzone zostały podzbiory populacji PESEL i KRS.

Podmioty te zostały wyselekcjonowane w dwojaki sposób: (i) przypadkowo i (ii) wg predefiniowanych kryteriów.

Na bazie tych podzbiorów wygenerowane zostały relacyjne bazy danych syntetycznych:

- ZUS (Zakład Ubezpieczeń Społecznych),
- CEPIK (Centralna Ewidencja Pojazdów i Kierowców),
- Systemy celne,
- Księgi wieczyste,
- System Bilingowy (operatorów telefonii komórkowej),
- Systemy GIIF (banki).

Dane wygenerowane: zakres i struktura

Zdefiniowane zostały (nieformalnie) następujące „kategorie” przestępstw, opisane w postaci konkretnych (uproszczonych) wzorcowych scenariuszy:

- **Ruchy finansowe:** wzajemne wysoce niesymetryczne przelewy między firmami będącymi własnością dwojga małżonków.
- **Nepotyzm:** dwie (lub więcej) niespokrewnione osoby wzajemnie zatrudniają w swoich firmach członków rodzin osoby drugiej.
- **Zależności cykliczne pomiędzy własnościami firm:** firma A ma udziały w firmie B, B w C, C w D, a firma D w pierwszej firmie A; kupno i sprzedaż towarów bądź usług pomiędzy tymi firmami mogą być tłem dla trudnych do wykrycia przestępstw gospodarczych.

Dane wygenerowane: zakres i struktura

- **Wyłudzenia ubezpieczeń:** (i) częste i fikcyjne wypadki komunikacyjne z udziałem osób powiązanych (lecz niespokrewnionych) i luksusowych samochodów w celu wyłudzenia wysokich odszkodowań, (ii) analogicznie jak w (i) lecz zamiast wypadków (fikcyjne) kradzieże.
- **Sprzedaż nieruchomości:** zakup w krótkim okresie wielu nieruchomości (odnotowanych w KW), po którym następują szybkie odsprzedaże podmiotowi gospodarczemu bądź osobie fizycznej po cenie znacznie przekraczającej cenę zakupu.
- **Paliwa:** znajomy lub krewny właściciela stacji paliw regularnie tankuje pojazdy ze swojej firmy transportowej wydając kwoty niebudzące podejrzeń (typowe dla samochodów swego przeznaczenia). Niezależnie od tego okresowo dochodzi do wielu transakcji w krótkim okresie czasu opiewających na bardzo wysokie (nietypowe) kwoty.

Dane wygenerowane: zakres i struktura

Dodanie do istniejącej już bazy danych nowych osób (tzn. numerów PESEL) na ogół skutkuje dodaniem nowych relacji – włączając ew. relacje z osobami w „starej” populacji.

Dla utrzymania logicznej spójności należy dokonać aktualizacji (wszystkich bądź niektórych) relacji, stanów rzeczy i charakterystyk osób, rzeczy i zdarzeń. Z kolei poczynione zmiany mogą „łańcuchowo” wymuszać kolejne, np. dodając do populacji O nową osobę identyfikowaną jako PESEL1 możemy uczynić ją dzieckiem osób identyfikowanych jako, odpowiednio, PESEL1a i PESEL1b, implikując relacje rodzicielskie (z których mogą wynikać *implicite* inne relacje pokrewieństwa).

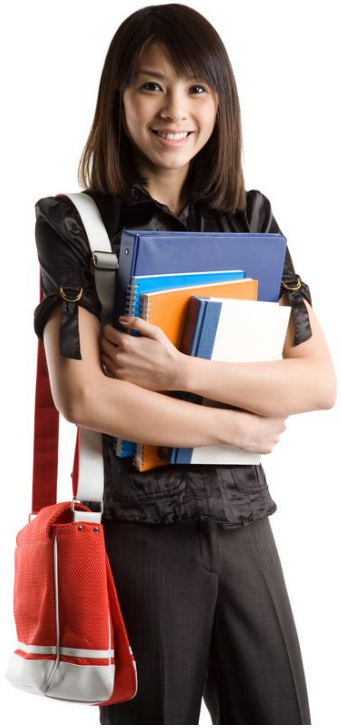
Analogiczne artefakty pojawiają się w sytuacjach, gdzie zdarzenia (np. przelew bankowy) pociąga za sobą zmianę wartości adekwatnej charakterystyki (w tym przypadku stan konta), lub tam gdzie zakup udziałów jakiejś firmy nie może spowodować przekroczenia wartości wszystkich akcji danej firmy.

Dane wygenerowane: zakres i struktura

Wygenerowana jak w powyższym opisie populacja PESEL ma charakter „statyczny” i jest przygotowana do ew. etapu realizacji dynamicznych procesów, które odzwierciedlałyby ruch naturalny ludności. Tak jak filarami etapu statycznego były dwa rodzaje powiązań: małżeńskie i rodzicielskie, które konstytuowały rodziny i pokolenia, etap dynamiczny opierałby się na czterech kategoriach zdarzeń: małżeństwa, rozwody, urodzenia i zgony, które realizowałyby ruch naturalny ludności.

O ile w fazie „statycznej” założono pewien realistyczny rozkład wg wieku, płci, małżeństw, liczby dzieci, to faza „dynamiczna” powinna być realizowana według rzeczywistych parametrów opisujących ruch naturalny ludności. Mówiąc nieformalnie, wygenerowane stany i zdarzenia będą (stochastycznie lub według predefiniowanych reguł) przypisywane wirtualnym osobom oraz powiązanym z nimi bytom (np. rozwód X i Y powodować może zmianę właściciela nieruchomości, w przypadku wspólności majątkowej).


GIF
EDG
RZ
KW
PESEL
REGON
CEPiK
ZUS
KRS
NFZ
KCIK
KSIP
EGIB
CRU
OEWIUDO
CEWIUP
KRK




§

Zgodnie z art. 4 pkt 1) RODO:

„**dane osobowe**” oznaczają informacje o **zidentyfikowanej** lub **możliwej do zidentyfikowania** osobie fizycznej („osobie, której dane dotyczą”);

możliwa do zidentyfikowania osoba fizyczna to osoba, którą można bezpośrednio lub pośrednio zidentyfikować, w szczególności na podstawie identyfikatora takiego jak imię i nazwisko, numer identyfikacyjny, dane o lokalizacji, identyfikator internetowy lub jeden bądź kilka szczególnych czynników określających fizyczną, fizjologiczną, genetyczną, psychiczną, ekonomiczną, kulturową lub społeczną tożsamość osoby fizycznej;

Dziękuję za uwagę !

dr inż. Maciej Kiedrowicz, Wydział Cybernetyki WAT