Recenzja rozprawy doktorskiej Tomasza Strumińskiego „Algorytmy ukrywania informacji”

Tomasz Jurdziński

24 lutego 2014

Rozprawa doktorska Tomasza Strumińskiego składa się z trzech części, w których porusza się różne zagadnienia ochrony informacji przed nieuprawnionym dostępem. W każdej części rozważane są precyzyjnie sformułowane problemy natury algorytmicznej i matematycznej, dobrze umotywowane praktyką informatyczną. W rozprawie dość szczegółowo omawia się rozważane modele i najważniejsze rezultaty, w każdej części prezentowane są też własne wyniki doktoranta.

Rezultaty rozprawy zostały zaprezentowane w czterech publikacjach doktoranta ze współautorami, w referowanych sprawozdaniach międzynarodowych konferencji z obszaru bezpieczeństwa informacji.

Omówienie rezultatów rozprawy

Poniżej omawiamy dokładniej rezultaty uzyskane w rozprawie.

Część I: usuwanie danych z dysków magnetycznych

Pierwsza część rozprawy dotyczy sposobów usuwania danych z dysków magnetycznych tak aby ochronić dane przed nieuprawnionym odczytem. Zakłada się, że konkretny bit na dysku jest nadpisywany tylko w sytuacji, gdy jego wartość się zmienia. Przyjęto dwa modele określające możliwości adwersarza próbującego odczytać dane z dysku: model z silnym adwersarzem (SA) i model z adwersarzem znającym czas zapisu (TA, od ang. time-aware adversary). W pierwszym modelu adwersarz ma możliwość ustalenia liczby zmian wartości każdego bitu na dysku. W drugim modelu wiedza adwersarza jest nawet większa: znana mu jest chronologiczna kolejność zmian wartości bitów. W rozprawie rozważa się scenariusz, w którym każdy zapis na dysku oznacza zastąpienie jego aktualnej zawartości ciągiem wylosowanym z rozkładem jednostajnym, przy zachowaniu zasady, że fizycznie nadpisywane są tylko wartości ulegające zmianie.

Zasadniczą kwestią badaną w rozprawie jest szacowanie prawdopodobieństwa, z jakim adwersarz może odtworzyć t-ty ciąg zapisany na dysku złożonym z n bitów przy założeniu, że na dysku dokonano m zapisów. Niezależność zmian wartości bitów przyjęta w rozważanym modelu implikuje, że kluczem do oszacowania wyżej określonego prawdopodobieństwa jest rozważenie przypadku dla “dysku” złożonego z pojedynczego bitu. W rozdziale 4 Pan Strumiński przedstawił szacowanie prawdopodobieństwa zdarzenia, że t-ta wartość bitu jest równa x pod warunkiem, że wykonano m zapisów w trakcie których doszło do d zmian wartości analizowanego bitu. Uzyskane szacowanie okazało się na tyle ścisłe, iż pozwoliło dowieść skuteczność ochrony przed nieuprawnionym odczytem polegającej na wielokrotnym zapisaniu losowych danych przed rozpoczęciem użytkowaniu dysku oraz po zakończeniu jego użytkowania. Dokładniej, pozwoliło ono na oszacowanie relacji między liczbą takich wstępnych i końcowych zapisów a uzyskanym poziomem bezpieczeństwa.

Rozdział 5 poświęcony jest modelowi TA. Kluczowe dla możliwości adwersarza jest tutaj założenie, że zmiany zawartości bitów w kolejnym ciągu wykonywane są “od lewej do prawej”: jeśli t-ty zapis wymaga nadpisania bitów x_i i x_j dla $i < j$, zapis na i-tej pozycji wykonywany

jest przed zapisem na j-tej pozycji. Pan Strumiński przedstawił dokładne i eleganckie obliczenia wykazujące, że przy zapisie na dysku ciągów losowych prawdopodobieństwo odczytania przez adwersarza wszystkich ciągów dąży do 1 wraz ze wzrostem wielkości dysku. Ponadto, przedstawiono dokładniejsze oszacowanie rozmiaru dysku wystarczającego do odtworzenia ciągów z prawdopodobieństwem co najmniej $1 - \epsilon$. Wobec niemożliwości obrony przed tak silnym adwersarzem, w rozprawie rozważa się możliwość „ochrony" poprzez przechowywanie na dysku przesunięcia cyklicznego danych (wartość przesunięcia jest sekretem przechowywanym poza dyskiem).

Część II: próbkowanie danych z zachowaniem anonimowości

Druga część rozprawy dotyczy zachowania prywatności danych przy udostępnianiu użytkownikom ich losowej próbki, uzyskanej przez włączenie każdej krotki z tym samym prawdopodobieństwem p. Rozważa się pojęcie (c, ϵ, δ)-prywatności pozwalające parametryzować szanse użytkownika na odróżnienie baz danych różniących się c krotkami na podstawie próbki losowej. A celem jest jak najlepsze oszacowanie prawdopodobieństwa próbkowania p gwarantującego (c, ϵ, δ)-prywatność. Problem ten badany był w wiodących ośrodkach, a wcześniejsze szacowania uzyskane przez K. Chaudhuri, N. Mishra prezentowane były na konferencji CRYPTO 2006. W rozprawie zaprezentowane zostało nowe szacowanie, które dla większości instancji problemu (tzn. wartości parametrów) znacznie poprawia najlepsze dotychczas znane szacowania prawdopodobieństw próbkowania zachowujących $(1, \epsilon, \delta)$-prywatność. Kluczowym elementem analizy jest rozważenie baz danych z dwoma typami krotek: czarnymi i białymi. Konkretną bazę danych można w tym przypadku scharakteryzować jako parę (B, W), gdzie B oznacza liczbę czarnych krotek a W liczbę białych krotek. Wynik uzyskany został poprzez eleganckie oszacowanie prawdopodobieństwa, że zarówno liczba wybranych białych krotek jak i liczba wybranych czarnych krotek dla bazy (B, W) mieszczą się w przedziałach nie pozwalających odróżnić bazy (B, W) od baz $(B - 1, W + 1)$ i $(B + 1, W - 1)$ z prawdopodobieństwem większym niż ϵ. Wynik ten został uogólniony na przypadek (c, ϵ, δ)-prywatności dla $c \geq 1$.

Część III: udostępnianie danych z zachowaniem anonimowości

Ostatnia część rozprawy dotyczy sposobów ukrywania informacji o dostępie do danych w sytuacji, gdy serwowanie danych zostało powierzone niezaufanemu *operatorowi bazy danych* (w skrócie *operatorowi*), z którym komunikuje się *bezpieczne urządzenie* (w skrócie *urządzenie*). Przyjmuje się tutaj, że baza danych składa się z różnych krotek $d_1, \ldots, d_n$. Użytkownicy końcowi kierują żądania dostępu do wybranych krotek. Bezpieczne urządzenie realizuje żądania użytkowników poprzez odpowiedni protokół komunikacji z operatorem bazy danych. Celem protokołu jest zagwarantowanie, że nie ma możliwości uzyskania informacji jakie były żądania użytkowników na podstawie historii komunikacji pomiędzy operatorem a bezpiecznym urządzeniem. Formalnie oznacza to, że każdy wzorzec dostępu do danych jest równie prawdopodny z punktu widzenia adwersarza znającego historię komunikacji między urządzeniem i operatorem. Jednocześnie należy zapewnić możliwie małą złożoność komunikacyjną protokołu, przez którą rozumiemy tutaj średnią liczbę bitów przesyłanych pomiędzy urządzeniem a operatorem przy realizacji jednego żądania. Przyjmujemy jednocześnie, że pamięć dostępna urządzeniu jest istotnie mniejsza od rozmiaru bazy danych (co uniemożliwia przechowywania bazy w pamięci urządzenia). W rozprawie opisano wprowadzony przez Yanga i in. schemat PIR (ang. private information retrieval) z pracy „An Effcient PIR Construction Using Trusted Hardware". W schemacie tym okresowo permutuje i (re)szyfruje się odczytane już fragmenty bazy danych. Jednocześnie, aby uniemożliwić adwersarzowi wiedzę o częstości dostępu do krotek, stosuje się sprytny trik polegający na odczytywaniu dwóch krotek z bazy przy każdym żądaniu użytkownika: jednej pobranej wcześniej i drugiej pobieranej po raz pierwszy. W rozprawie opisano dwa warianty schematu. W pierwszym urządzenie musi dysponować dużą pamięcią. W drugim

wymagania pamięciowe są znacznie mniejsze. W rozdziałach 11-12 wskazano błędy w analizie opisanego schematu przedstawionej przez Yanga i in. W przypadku schematu z dużą pamięcią autorzy argumentowali, że bezpieczeństwo schematu wynika z faktu, że każda odczytana krotka zajmie każdą pozycję w bazie z jednakowym prawdopodobieństwem. W rozprawie wskazano na naiwność wiary w siłę takiego argumentu, w którym niejako jednostajność rozkładu wielowymiarowej zmiennej losowej wnioskuje się w oparciu o własności rozkładu zmiennej definiującej pojedynczy wymiar. Pomimo wady w rozumowaniu Yanga i in. okazuje się, że zaproponowany protokół faktycznie jest bezpieczny. Dość prosty lecz elegancki dowód tego faktu został przedstawiony w rozprawie.

W odniesieniu do schematu bez pamięci wskazano w rozprawie na fakt, że w przypadku odwołań do tzw. białych krotek algorytm może odwoływać się (w jednej epoce) więcej niż raz do pewnych czarnych krotek. Oznacza to, że w niektórych etapach algorytmu prawdopodobieństwa odwołania do czarnej i białej krotki nie są równe z punktu widzenia adwersarza. W rozprawie skonkludowano tą obserwację wnioskiem, że schemat nie jest bezpieczny wg przyjętej definicji. Nieco zabrakło mi tutaj konkluzji. Ciekawe byłoby oszacowanie jaki wpływ na szanse adwersarza mogą mieć parametry systemu (n, k i in.). Inną ciekawą kwestią jest czy możliwe jest poprawienie podanego schematu tak, aby spełniał formalne wymogi bezpieczeństwa. Czy też konieczne jest rozluźnienie wymogów stawianych bezpiecznemu protokołowi?

Merytoryczna ocena rozprawy

Wszystkie problemy rozważane w rozprawie są dobrze umotywowane praktyką informatyczną i/lub wyzwaniami wynikłymi z powszechności komunikacji cyfrowej. Są one również precyzyjnie zdefiniowane i nietrywialne. Pan Strumiński uzyskał kilka ciekawych rezultatów dla omawianych problemów, wykazując dobrą znajomość warsztatu matematycznego i algorytmicznego, pomysłowość i znajomość badanej tematyki.

Część pierwsza rozprawy wprowadza bardzo ciekawy model i problemy. Uzyskane wyniki nie są zaskakujące, choć ich otrzymanie wymagało inwencji i dużej sprawności obliczeniowej. Przyjęte założenie o zapisie zawsze całego dysku, choć dobrze uzasadnione, nieco upraszcza problem. Wydaje się, że na bazie tego modelu warto studiować bardziej realistyczne i trudniejsze wzorce zapisu danych.

Dla odmiany w części drugiej autor zajął się modelem próbkowania analizowanym wcześniej w czołowych ośrodkach i, co warto podkreślenia, uzyskał lepsze oszacowania prawdopdobieństw próbkowania od prezentowanych wcześniej. Sądzę, że zarówno od strony technicznej jak i ze względu na konsekwencje uzyskanego rezultatu jest to jeden z najciekawszych wyników rozprawy. Również trzecia część rozprawy pokazuje umiejętność krytycznego i dogłębnego spojrzenia na wyniki i problemy prezentowane przez innych uczonych. W tym przypadku główny wkład polega na wskazaniu istotnych błędów i uproszczeń w analizie schematu PIR wykonanej przez Yanga i in.

Uwagi techniczne

Mocną stroną rozprawy Pana Strumińskiego jest sposób prezentacji wyników. Autor włożył dużo wysiłku w dobre uzasadnienie dla rozważanych problemów, precyzyjnie i przystępnie opisał rozważane modele i wcześniejsze wyniki. Przydatne były również liczne przykłady ilustrujące wprowadzane pojęcia i analizowane algorytmy. Wyniki własne prezentowane są szczegółowo i dokładnie, co nie pozostawia wątpliwości do poprawności rezultatów. Wyjątkiem od tej reguły jest jedynie dowód lematu 39, w którym sposób wyboru wartości γ nie został właściwie opisany; czytelnikowi pozostaje zajrzeć do publikacji z IWSEC 2010, gdzie znajduje się dokładniejszy dowód.

W rozprawie pojawiło się kilka drobnych błędów lub literówek, zaznaczyć jednak trzeba że ich ilość jest bardzo mała w zestawieniu z długością tekstu. Poza tym nie mają one wpływu na

poprawność postulowanych rezultatów.

Poniżej podaję kilka drobnych błędów, które mogą być kłopotliwe dla czytelnika:

- Uzasadnienie nierówności w przekształceniu na dole strony 52 jest niepoprawne (pomijając dodatnie składniki sumy uzyskujemy wartość mniejszą, nie większą); nie zmienia to wszakże wyniku tw. 23. Wydaje mi sie również, że sumowane wyrażenia to $P[Y_i = 0|L_i = k]$, a nie $P[Y_i = 0|F_i = k]$. Poza tym zmienna k występuje tu w dwóch znaczeniach.
- Kierunek ostatniej nierówności na stronie 53 powinien być przeciwny.
- Na str. 77 autor postuluje, że wartości h_w i h_b „oznaczono na rysunku 7.1". Wydaje się jednak, że czytelnikowi trudno je określić na podstawie rysunku. Nie jest również oczywiste oszacowanie $\gamma > \frac{\epsilon(1-p)}{2p(1+\epsilon)}$. Okazuje się, że dokładniej ta konstrukcja została opisana w publikacji konferencyjnej, z której pochodzi rezultat. Trochę to dziwi, szczególnie w porównaniu ze szczegółowością pozostałej części rozprawy. Tym bardziej, że uzyskany w efekcie wynik jest jednym z ciekawszych rezultatów rozprawy.
- W części trzeciej powtarza się określenie „operator baza danych", zręczniejszym byłoby po prostu operator bazy danych. Sądzę, że taka była intencja (?). W opisie algorytmu 3 kilka nierówności jest niepoprawnie skierowanych (algorytm pochodzi z pracy [91], więc błędy nie są istotne merytorycznie, ale uciążliwe dla czytelnika).

Konkluzja

Rozprawa Pana Tomasza Strumińskiego zawiera oryginalne i ciekawe wyniki dotyczące dobrze postawionych i uzasadnionych praktyką problemów z zakresu bezpieczeństwa danych. Uzyskanie przedstawionych wyników wymagało rozległej wiedzy z zakresu probabilistyki, algorytmiki i kombinatoryki, a także pomysłowości w używaniu narzędzi matematycznych. Rezultaty zaprezentowane w rozprawie zostały przyjęte do publikacji w recenzowanych sprawozdaniach międzynarodowych konferencji. Na uwagę zasługuje również umiejętność matematycznego modelowania zagadnień technicznych, szczególnie widoczna w części pierwszej.

W mojej ocenie, w prezentowanej rozprawie Pan Strumiński wykazał szeroką wiedzę oraz umiejętność samodzielnego prowadzenia pracy naukowej. Uważam, że rozprawa w pełni spełnia wymogi ustawy o stopniach i tytule naukowym i wnioskuję o dopuszczenie do dalszych etapów przewodu doktorskiego.