dr hab. Wojciech Guzicki
Instytut Matematyki
Uniwersytet Warszawski

Warszawa, 3 lutego 2014 r.

Recenzja rozprawy doktorskiej

Tytuł: Algorytmy ukrywania informacji

Autor: Tomasz Strumiński

Recenzowana rozprawa doktorska składa się z trzech części. W każdej z tych części rozpatrywane jest inne zagadnienie związane z bezpieczeństwem przechowywanych informacji. Cała rozprawa ma dużą objętość: 154 strony. Na tę objętość składają się:

1. spis treści, wstęp i oznaczenia stosowane w pracy — 12 stron,
2. część I poświęcona problemowi usuwania danych z dysków — 50 stron,
3. część II poświęcona problemowi porównywania baz danych — 28 stron,
4. część III poświęcona problemowi udostępniania oddelegowanych danych — 50 stron,
5. streszczenie w języku angielskim i bibliografia — 14 stron.

Omówię teraz krótko problemy, którymi doktorant zajmuje się w rozprawie. Problem pierwszy dotyczy usuwania danych z dysków magnetycznych. Osobie, która nie zna się na architekturze komputerów, może wydawać się, że operacja usunięcia pliku z dysku rzeczywiście powoduje wymazanie go z tego dysku i że odtworzenie jego zawartości będzie już niemożliwe. To oczywiście nie jest prawdą. Skasowany zostaje wyłącznie dostęp do części pamięci zajętej przez usunięty plik, ale fizycznie zapis bitów, z których ten plik się składał, pozostaje na dysku. Osoba, która o tym się dowie, może zaproponować lepszy sposób usunięcia pliku z dysku: nadpisanie go innymi danymi. Może ona sądzić, że w ten sposób każdy bit pliku, znajdujący się fizycznie na dysku, zostanie zastąpiony nowym i stary plik w ten sposób zniknie fizycznie z dysku. To też nie jest prawdą. Urządzenie zapisujące poszczególne bity pliku na dysku magnetycznym nigdy nie jest wystarczająco precyzyjne, by każdy kolejny bit o danym adresie zapisać dokładnie w tym samym miejscu. Powstają niewielkie przesunięcia powodujące, że pewne pozostałości bitów zapisanych w przeszłości są nadal możliwe do odczytania. W ten sposób możemy w danym miejscu (tzn. w miejscu o danym adresie) odczytać nawet wszystkie bity tam zapisane. Autor rozważa zatem dwa modele matematyczne takiego dysku i zastanawia się, czy możliwe jest takie nadpisywanie plików, by — mimo odczytania całej historii dysku — niemożliwe było zidentyfikowanie poszczególnych plików. Inaczej mówiąc, by niemożliwe było rozpoznanie, który bit tak naprawdę pochodzi z którego pliku.

Oba modele matematyczne zakładają, że na dysku zapisuje się nie każdy bit pliku, ale tylko bity różne od istniejących pod danym adresem. Na przykład, jeśli na dysku zapisano kolejno bity 11100010, a następnie nadpisano na tym miejscu bity 10000001, to pierwszy bit nie zostanie nadpisany (bo nie ma takiej potrzeby), następne dwie jedynki zostaną nadpisane zerami, potem trzy zera nie zostaną nadpisane, znów jedynka zostanie nadpisana zerem i wreszcie zero zostanie nadpisane jedynką. W ten sposób niektóre bity mogą bardzo długo nie być nadpisywane, podczas gdy inne były wielokrotnie zmieniane.

W ten sposób nie będzie wiadomo, w którym momencie dany bit został nadpisany, a więc z którego pliku on pochodzi.

Model pierwszy polega na tym, że znane są tylko historie nadpisywania poszczególnych bitów (a więc liczba zmian bitu); model drugi pozwala także poznać czas, w którym dane zapisy powstały. Ten drugi model jest motywowany tym, że zapis na dysku magnetycznym „starzeje się", co pozwala oszacować czas jego „przebywania" na dysku.

Doktorant analizuje problem, czy adwersarz znający historię dysku, może odczytać zapisane na nim pliki. W tym celu doktorant szacuje wielkość oznaczoną w pracy symbolem $\Delta(m, d, t)$ (m oznacza liczbę zapisów na dysku, d liczbę zmian rozważanego bitu, t numer, w kolejności zapisywania, rozważanego bitu). Ta wielkość ma służyć znalezieniu prawdopodobieństwa $P_t = P[X_t = 0 \mid m, d]$ tego, że t-ty zapisany bit ma wartość 0 pod warunkiem, że dokonano w sumie m zapisów na dysku, przy czym rozważany bit był zmieniany d razy. Naturalne rozważania kombinatoryczne pozwoliły doktorantowi oszacować rozważaną wielkość Δ i w konsekwencji udowodnić twierdzenie (w pracy twierdzenie 15) szacujące bezpieczeństwo bitów nadpisanych wielokrotnie. Konsekwencją udowodnionego twierdzenia jest zaproponowany algorytm zabezpieczania plików w praktyce. Polega on na zapisaniu t losowych plików, potem m plików użytkownika i następnie nadpisaniu ich t losowymi plikami. Zbadane wartości t są niewielkie, co dowodzi przydatności tego algorytmu w praktyce.

Znacznie większe problemy stwarza model drugi (w którym adwersarz może ustalić czas zapisania poszczególnych bitów). W tym modelu zachowanie bezpieczeństwa usuwanych plików jest znacznie trudniejsze. Rozważane jest zapisywanie danych z wykorzystaniem niewielkiej pamięci zewnętrznej, do której adwersarz nie ma dostępu oraz zapisywanie bitów nie następuje sekwencyjnie, ale następuje w kolejności losowej. To dopiero pozwala na uzyskanie satysfakcjonującego poziomu bezpieczeństwa. Udowodnione twierdzenie 23 i wnioski z niego pokazują, że bez odpowiedniego sposobu zabezpieczenia adwersarz może odczytać nadpisywane pliki.

Drugi rozważany problem polega na porównywaniu dwóch baz danych. Przypuśćmy, że mamy udostępnioną bazę danych, w której dane osób zostały zanonimizowane. Mamy jednak dostęp do podobnej bazy danych, w której te dane są umieszczone. Przez porównanie znanych części obu baz można odzyskać część danych, która została z pierwszej bazy usunięta przed udostępnieniem jej. Doktorant przytacza dwa rzeczywiste przypadki takiego skompromitowania baz danych.

Doktorant rozważa pojęcie tzw. prywatności różnicowej, zajmuje się kwestią tzw. próbkowania baz danych i poszukuje możliwie największej wartości prawdopodobieństwa próbkowania p (czyli prawdopodobieństwa włączenia danych z bazy do próbki), dla którego prywatność różnicowa jest zachowana. Doktorant dowodzi twierdzenia 38 i wynikającego z niego twierdzenia 40 szacującego to prawdopodobieństwo p. Uzyskany wynik istotnie wzmacnia dotychczasowy wynik uzyskany przez Chaudhuriego i in. w pracy [16].

Wreszcie zagadnienie trzecie, którym doktorant się zajmuje, polega na zbadaniu bezpieczeństwa danych udostępnionych użytkownikom za pośrednictwem stron niezaufanych. Rozważany jest tzw. system PIR (*private information retrieval*). Doktorant szczegółowo analizuje system zaproponowany przez Yanga i in. oraz wykazuje, że system ten cha-

rakteryzuje się zbyt małym współczynnikiem „wymieszania" danych. Dowodzi także, że w silniejszym modelu (wykorzystującym niedostępną adwersarzowi pamięć dodatkową) ten współczynnik „wymieszania" zostanie znacznie zwiększony (lemat 52).

Część I ma charakter zdecydowanie matematyczny i jest — moim zdaniem — najciekawszym matematycznie zagadnieniem badanym w pracy. Zręczne, a przy tym bardzo naturalne rozumowania kombinatoryczne dają wyraźne odpowiedzi na dobrze sformułowane pytania. Badania podjęte w pracy mogą mieć naturalną kontynuację. Na stronie 35 (wiersze pierwszy i drugi od góry) Autor pisze, że faktyczna wartość współczynnika Δ jest zwykle znacząco mniejsza niż to wynika z oszacowania. Powstaje naturalne pytanie o to, czy inne, być może subtelniejsze, metody matematyczne nie doprowadzą do lepszych oszacowań, bliższych rzeczywistości. Doktorant rozważa naturalny z kryptologicznego punktu widzenia problem zapisywania plików losowych. Pliki zaszyfrowane tak z reguły wyglądają. Zupełnie innym problemem byłoby zbadanie bezpieczeństwa plików niewiele się od siebie różniących. Na przykład użytkownik pracujący długo nad kolejnymi wersjami jakiegoś pliku i zmieniający je w niewielkim stopniu, może chcieć ukryć efekty swojej pracy. Interesujące byłoby zatem zbadanie, czy zaproponowane metody nadpisywania bedą należycie chroniły kolejne wersje tego samego pliku, różniące się od siebie tylko niewielką liczbą bitów.

Część II jest dość techniczna i — jak wspomniałem wyżej — miała wzmocnić wyniki uzyskane wcześniej przez innych autorów. Ten cel został osiągnięty. Część III ma z kolei charakter bardziej informatyczny. Analizowany jest jeden konkretny algorytm i jego modyfikacje. Za pomocą zręcznego indukcyjnego rozumowania (również o charakterze kombinatorycznym) Autor także poprawia rozumowania poprzedników. Pracę kończy bardzo obszerna, dobrze zestawiona bibliografia. Główne wyniki pracy były już publikowane w czterech pracach pisanych zespołowo (głównie z promotorem i innymi osobami). Autor wskazuje także wyraźnie, które części tych publikacji pochodzą od niego.

We wszyskich trzech częściach doktorant stosuje rozumowania matematyczne o charakterze kombinatorycznym do rozwiązania konkretnych, dobrze postawionych problemów. Doktorant wykazuje się dobrą znajomością wyników uzyskanych w uprawianej dziedzinie matematyki i informatyki (w szczególności kryptografii). Wykazuje się również umiejętnością analizowania algorytmów kryptograficznych. Czytając rozprawę zwracałem szczególną uwagę na stosowane metody matematyczne. Są to — jak wspomniałem — metody kombinatoryki wykorzystywane przede wszystkim do szacowania prawdopodobieństw. Ponadto doktorant stosuje różnorodne twierdzenia rachunku prawdopodobieństwa. Uważam, że w całej pracy doktorant wykazał się dobrą znajomością metod matematycznych, które zastosował do rozwiązania dobrze postawionych naturalnych problemów matematyki stosowanej o istotnym znaczeniu praktycznym. Uważam, że recenzowana rozprawa i własny wkład doktoranta w wyniki uzyskane w pracach współautorskich spełniają wszystkie wymagania stawiane rozprawom doktorskim przez Ustawę o stopniach i tytule naukowym. Wobec tego wnoszę o dopuszczenie p. Tomasza Strumińskiego do dalszych faz przewodu doktorskiego.