METODY EKSPLORACJI DANYCH W ANALIZIE RUCHU OBSERWOWANEGO PRZEZ SYSTEMY HONEYPOT

Krzysztof Cabaj, Michał Buda

Instytut Informatyki Politechniki Warszawskiej

Streszczenie. Od kilku lat systemy HoneyPot są coraz szerzej wykorzystywane w celu szybkiego zdobywania informacji dotyczących nowych ataków pojawiających się w Internecie. Mimo dużej liczby badań dotyczących nowych systemów HoneyPot, brakuje oprogramowania umożliwiającego analiza danych przez nie uzyskanych. W artykule znajduje się opis systemu WebHP/HPMS (ang. HoneyPot Management System) umożliwiającego analizę z wykorzystaniem metod eksploracji danych, zastosowanych technik oraz rezultaty pierwszych eksperymentów. Uzyskane wyniki są obiecujące, ponieważ w natłoku uzyskanych danych wykryte wzorce umożliwiły szybką identyfikację nowych zagrożeń.

Słowa kluczowe: systemy HoneyPot, analiza danych, eksploracja danych

Potrzeba wdrażania i utrzymywania systemów ochrony informacji, między innymi systemów: zapór ogniowych, wykrywania włamań czy antywirusowych, jest dzisiaj bezdyskusyjna. Aktualnie większość systemów tego typu wykorzystuje wiedzę uzyskaną z analizy różnego typu zagrożeń, rozpowszechnianą w postaci reguł, sygnatur czy szczepionek. Dane do analizy pochodzą z zaatakowanych maszyn użytkowników, jak również coraz częściej z systemów będących pułapkami na atakujących, nazywanych systemami HoneyPot. Na rynku komercyjnym i w środowisku oprogramowania opartego na otwartym kodzie dostępne są różnego typu systemy HoneyPot, poczynając od prostych symulatorów pojedynczych wybranych usług, poprzez symulatory całych sieci a kończąc na systemach monitorujących zintegrowanych na rzeczywistym sprzęcie z popularnym oprogramowaniem systemowym i użytkowym.

Wdrożenie systemu HoneyPot jest relatywnie proste, jednak w związku z ilością danych uzyskiwanych z tego typu systemów coraz większym problemem staje się ich szybka analiza. Motywacją do prezentowanych prac badawczych była obserwacja braku na rynku narzędzi pozwalających wyciągać wnioski z zarejestrowanych danych, które ułatwiłyby stworzenie nowych sygnatur i reguł dla używanych systemów bezpieczeństwa. W artykule znajduje się opis zaproponowanej i zaimplementowanej w eksperymentalnym systemie HPMS (ang. HoneyPot Management System) metody analizy danych. Danymi wejściowymi w tym systemie są żądania skierowane do

serwera WWW. Ich analiza wykonywana jest za pomocą metod eksploracji danych. Wykorzystano tu autorskie algorytmy podziału danych i dalszej analizy wyników częściowych. Zaletą użytego w HPMS wykrywania wzorców z wykorzystaniem metody zbiorów częstych jest czytelność, łatwość i intuicyjność zrozumienia uzyskanych wzorców. W pracy zostanie dokładnie omówiona zaproponowana metoda oraz wnioski z wdrożenia prototypu systemu w sieci Instytutu Informatyki Politechniki Warszawskiej. Dodatkowo zostaną przedstawione wyniki uzyskane z eksperymentów na rzeczywistych danych uzyskanych z prawie rocznego okresu działania systemu HPMS, które potwierdzają przydatność metody. W tym czasie, między innymi, wykryto maszyny dokonujące masowych ataków na serwer, pojawienie się w ruchu sieciowym aktywności związanej z nowym skanerem podatności jak również ataki niedawno wykrytego robaka „The Moon".

Praca posiada następujący układ. Rozdział pierwszy poświęcony jest wprowadzeniu do idei systemów HoneyPot. Drugi rozdział zawiera opis metod eksploracji danych, z naciskiem na metodę zbiorów częstych wykorzystywaną w opisywanym systemie HMPS. Następny rozdział poświęcony jest opisowi prototypowej instalacji systemu WebHP wraz z systemem zarządzania i analizy danych HPMS. Kolejny, czwarty rozdział zawiera opis eksperymentów oraz uzyskanych wyników przeprowadzonych na danych zebranych przez system WebHP. Ostatni piąty rozdział zawiera, krótkie podsumowanie wykonanych prac wraz z zarysowaniem kierunków dalszych prac.

1. Systemy HoneyPot

Systemy HoneyPot są narzędziem umożliwiającym poznanie sposobów działania oraz motywacji atakujących. System HoneyPot nie jest określonym rozwiązaniem sprzętowo programowym i w zależności od potrzeby może być zbudowany na różne sposoby. Jedyną wspólną cechą jest to, że system HoneyPot nie posiada żadnej produkcyjnej roli w organizacji, która go uruchamia [8]. Jego jedynym zadaniem jest oczekiwanie na atak z zewnątrz. W razie wystąpienia ataku wszelkie informacje, które mogą być przydatne do analizy ataku są zbierane. W zależności od potrzeb, systemem HoneyPot może być program symulujący pewną usługę, specjalnie skonfigurowana maszyna zawierająca rzeczywiste oprogramowanie lub fikcyjnie stworzony cały fragment sieci z maszyn, łączy i urządzeń sieciowych. Wyczerpujący opis najpopularniejszych systemów HoneyPot wraz z propozycją ich taksonomii można znaleźć w pracy [10]. Systemy HoneyPot można podzielić na wysokiej i niskiej interakcji. Pierwsze z nich są

skierowane głównie na atakujących samodzielnie wyszukujących podatności w określonym systemie. W takim przypadku dany zasób musi wydawać się interesujący dla atakującego aby zachęcił go do zbadania podatności i próby ich wykorzystania. Drugi rodzaj systemów, który jest wykorzystywany w związku z niniejszą pracą, skierowany jest głównie do rejestrowania automatycznych prób zbierania informacji i ataków. W tym przypadku nie trzeba się dokładać specjalnych starań aby system HoneyPot wydawał się interesujący. W wielu przypadkach nawet nie trzeba próbować ogłaszać jego istnienia z wykorzystaniem systemu DNS czy z pomocą linków z innych stron. Wystarczy samo podłączenie do sieci Internet. Jak pokazują badania po bardzo krótkim czasie zostaną nawiązane pierwsze połączenia. System taki bardzo szybko zostanie rozpoznany jako system HoneyPot przez człowieka, jednak doskonale nadaje się do zbierania informacji dotyczących automatycznych i masowych aktywności pojawiających się w Internecie. Od początku 21 wieku systemy HoneyPot cieszą się niesłabnącym zainteresowaniem co skutkuje dużą liczbą prac badawczych. Jak pokazuję prace przeglądowe [4, 9] większość aktualnie prowadzonych badań związanych jest z rozwijaniem nowych rodzajów systemów HoneyPot, sposobów ich ukrywania oraz detekcji [6]. Niestety niewiele uwagi poświęcanej jest analizie oraz wizualizacji danych uzyskanych za pomocą tych systemów.

Więcej szczegółów dotyczących zaimplementowanego i wdrożonego systemu HoneyPot, który był źródłem rzeczywistych danych do analizy znajduje się w rozdziale trzecim. Rozdział czwarty poświęcony jest omówieniu wyników uzyskanych z analizy danych, które reprezentują różnego typu aktywności zaobserwowane przez wdrożony system.

2. Metody eksploracji danych

Posiadanie ogromnych zbiorów danych przez różne organizacje spowodowało rozwój technik umożliwiających ich analizę. Jedną z możliwych do zastosowania metod jest wykorzystanie odkrywania wiedzy (ang. Knowledge Discovery in Databases) zakładające, że w danych kryje się jakaś istotna, w związku z ich wielkością niezauważalna na pierwszy rzut oka, interesująca dla ich posiadacza wiedza. Eksploracja danych jest jednym z najważniejszych etapów całego procesu odkrywania wiedzy polegającym na wykorzystaniu określonych algorytmów do właściwej analizy danych, często utożsamiana z całym procesem odkrywania wiedzy. Pozostałe, często pomijane a nie mniej ważne etapy związane są z przygotowaniem wstępnym danych, przygotowaniem uzyskanych wyników do prezentacji człowiekowi oraz weryfikacją i zastosowaniem wykrytej wiedzy. Wśród stosowanych algorytmów eksploracji danych najczęściej wymieniane podejścia

związane są z grupowaniem (ang. clustering) oraz klasyfikacją (ang. classification). Niniejsza praca opisuje praktyczne wykorzystanie mniej popularnego, a w wielu przypadkach bardzo przydatnego podejścia wykrywania wzorców częstych (ang. frequent patterns discovery). W zależności od przyjętej reprezentacji danych wzorcem częstym może być podzbiór [1], sekwencja elementów [2] czy nawet podgraf [11]. Na potrzeby opisywanych w niniejszym artykule eksperymentów skorzystano z używanego podczas analizy koszykowej wzorca jakim jest zbiór częsty. Pierwsze opisane zastosowania tej metody miały określić jakie produkty klienci kupują łącznie, np. w celu zaproponowania odpowiedniej ceny lub rozmieszczenia w sklepie. Na potrzeby tych analiz każde pojedyncze klienckie zakupy, nazywane transakcją, reprezentowane są jako zbiór, w którym elementy odpowiadają poszczególnym zakupionym produktom. Zgodnie z definicją zaproponowaną w pracy [1], zbiorem częstym nazywany jest podzbiór występujący co najmniej w określonej przez analizującego minimalnej liczbie transakcji. Zwyczajowo parametr ten nazywany jest minimalnym wsparciem (ang. minimal support). Na uwagę zasługuje jeszcze jeden, często powodująca pewne nieporozumienia, fakt związany z istnieniem różnych algorytmów do wykrywania tego samego rodzaju wzorca. Przykładowo, wykorzystany w niniejszej pracy wzorzec - zbiór częsty - może być wykrywany za pomocą algorytmu Apriori [1], lub z wykorzystaniem różnego typu drzew FP-Tree [7], CATS [5].

Rozpatrzmy przykładowy zbiór transakcji przedstawiony w tablicy Tabela 1. W każdym wierszu tabeli znajduje się jeden zbiór odpowiadający kolejnym transakcjom. Dla ułatwienia omawiania przykładu elementy transakcji są identyfikowane za pomocą pojedynczych liter. W implementowanych rozwiązaniach w celu zapewnienia możliwie szybkiego i efektywnego porównywania elementów zbiorów są one reprezentowane w postaci liczb całkowitych. W ramach etapu przygotowania wstępnego danych, właściwe dane podlegające analizie zostają przetransformowane od postaci dogodnej do dalszej analizy przez algorytmy eksploracji danych. Więcej szczegółów dotyczących tego typu procesu znajduje się w sekcji trzeciej niniejszego artykułu.

Przy założonym parametrze minimalnego wsparcia o wartości 3, w przedstawionym przykładowym zbiorze danych zbiorami częstymi będą między innymi podzbiory "ab", "efg" oraz "g". Zbiór "ab" występuje w transakcjach 1, 2 i 3 a "efg" i "g" w transakcjach 3, 4 i 5. Podzbiór "abc" nie jest zbiorem częstym ponieważ występuje jedynie w transakcjach 1 i 3, czyli jego wsparcie o wartości dwa jest mniejsze niż założony próg minimalnego wsparcia o wartości trzy.

Tabela 1
Przykładowy zbiór danych wykorzystywanych przez algorytmy wykrywania zbiorów częstych. Dla parametru minSup = 3 maksymalnymi zbiorami częstymi są "ab" oraz "efg"

Identyfikator transakcji	Zawartość
1	(a, b, c)
2	(a, b)
3	(a, b, c, d, e, f, g)
4	(e, f, g, h)
5	(e, f, g, h, i)

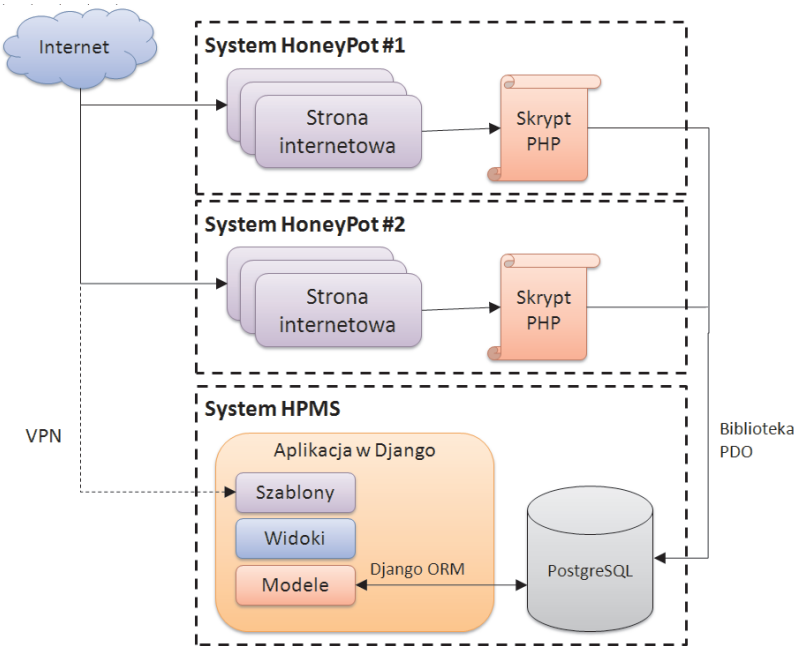
Jak łatwo można zauważyć, jeśli zbiorem częstym jest podzbiór "efg" to zbiorami częstymi także będą jego wszelkie podzbiory - "ef", "fg", "eg", "e", "f" i "g". Jeśli skorzystamy z algorytmu wykrywającego wszystkie możliwe zbiory częste w wyniku uzyskamy zbiór częsty jak i jego wszystkie podzbiory. Z tego powodu często podczas wykrywania uzyskujemy jedynie maksymalne zbioru częste, czyli takie zbiory częste, których wszystkie nadzbiory nie są zbiorami częstymi. W omawianym wcześniej przykładzie, maksymalnymi zbiorami częstymi są zbiory "ab" oraz "efg".

Największą zaletą zastosowania tego typu wzorców jest zmniejszenie liczby informacji, które musi przeanalizować analityk. Nawet w prostym przykładzie zaprezentowanym powyżej z sześciu transakcji otrzymujemy dwa wzorce. Przy rzeczywistych danych z tysięcy zarejestrowanych transakcji uzyskamy kilkadziesiąt wzorców. Dodatkowo pewne nieistotne zmienne elementy analizowanych danych zostaną usunięte, prezentując jedynie najistotniejsze, najczęściej pojawiające się elementy. W kolejnym rozdziale zostanie omówiony zaimplementowany system, umożliwiający wykrywanie omówionych powyżej wzorców w danych zebranych przez systemy HoneyPot.

3. System WebHP/HPMS

Rozdział ten zawiera dokładny opis zaimplementowanego i wdrożonego systemu HoneyPot dedykowanego atakom na aplikacje Webowe wraz z systemem analizy danych, wykorzystującym mechanizmy eksploracji danych. System posiada dwa podstawowe podsystemy - WebHP oraz HPMS (ang. HoneyPot Management System).

Podsystem WebHP jest odpowiedzialny za zbieranie danych dotyczących całego ruchu dochodzącego do sensorów. Jest zaimplementowany w języku php i uruchomiony pod kontrolą specjalnie skonfigurowanego serwera



Rysunek 1. Schemat wdrożonej w Instytucie Informatyki Politechniki Warszawskiej instalacji HoneyPot, składającego się z sensorów WebHP wraz z systemem HPMS

Apache. Wszelka aktywność skierowana do tak przygotowanego serwera WWW zostaje zarejestrowana i po wstępnej analizie zapisana w bazie danych. Aktualnie wdrożony system uruchomiony jest na dwóch niezależnych maszynach i kilku najaktywniejszych portach, między innymi 80, 8080 oraz 5000. System HPMS jest odpowiedzialny za cykliczne wykonywanie analiz oraz prezentację uzyskanych wyników analitykowi. Jest on zaimplementowany w języku Python, a Webowy interfejs użytkownika zaimplementowano z użyciem środowiska Django. Rysunek 1 prezentuje schemat aktualnie wdrożonej w Instytucie Informatyki konfiguracji sensorów WebHP i systemu HPMS wraz z najważniejszymi przepływami danych.

Najistotniejszą częścią wdrożonego systemu jest mechanizm analizy zaobserwowanego przez sensory ruchu z wykorzystaniem metod eksploracji danych. Do tego celu wykorzystany został mechanizm wykrywania zbiorów częstych. Zgodnie z opisem w poprzednim rozdziale wykrywanie tego wzorca wymaga reprezentacji danych w postaci zbiorów. Z tego powodu pierwszym etapem analizy jest proces przygotowania wstępnego danych. Transakcją w tym przypadku jest każde pojedyncze połączenie zarejestrowane przez sensor i reprezentowane jako zbiór. Elementami wchodzącymi

w skład tego zbioru są identyfikatory odpowiadające występującym w żądaniu protokołu HTTP kolejnym atrybutom, rozszerzonym o pewne metadane związane z połączeniem, przykładowo adresem klienta. Dodatkowo, zawarty w żądaniu URL został potraktowany nie jako jeden ciąg a zbiór kilku elementów, każdy odpowiadający części oddzielonej znakiem '/'. Takie podejście pozwoliło wykrywać przykładowo skanowania w których wielokrotnie występowały żądania dotyczące tego samego pliku umieszczonego w różnych lokalizacjach. W wyniku tych operacji każde połączenie reprezentowane jest jako zbiór liczb naturalnych. Dodatkowo, w bazie danych przechowywana jest informacja pozwalająca w późniejszym czasie (po wykryciu wzorców) dokonać mapowania odwrotnego, tj. zamienić identyfikatory poszczególnych atrybutów na postać tekstową zrozumiałą dla człowieka. Tabela 2 prezentuje przykładowy zbiór danych w postaci uzyskanej z systemu WebHP oraz po wykonaniu procesu przekształcenia wstępnego danych. Dla celów omówienia systemu, tabela prezentuje jedynie fragment analizowanych danych: adresy nawiązujących połączenia oraz żądane URI. W zaimplementowanym systemie z każdą transakcja może być związanych nawet do kilkudziesięciu różnych parametrów. W dalszej części pracy (na Rysunku 2) zaprezentowany jest przykładowy zrzut ekranu z systemu HPMS pokazujący przykładowe atrybuty wchodzące w skład wykrytych zbiorów częstych dla rzeczywistych danych.

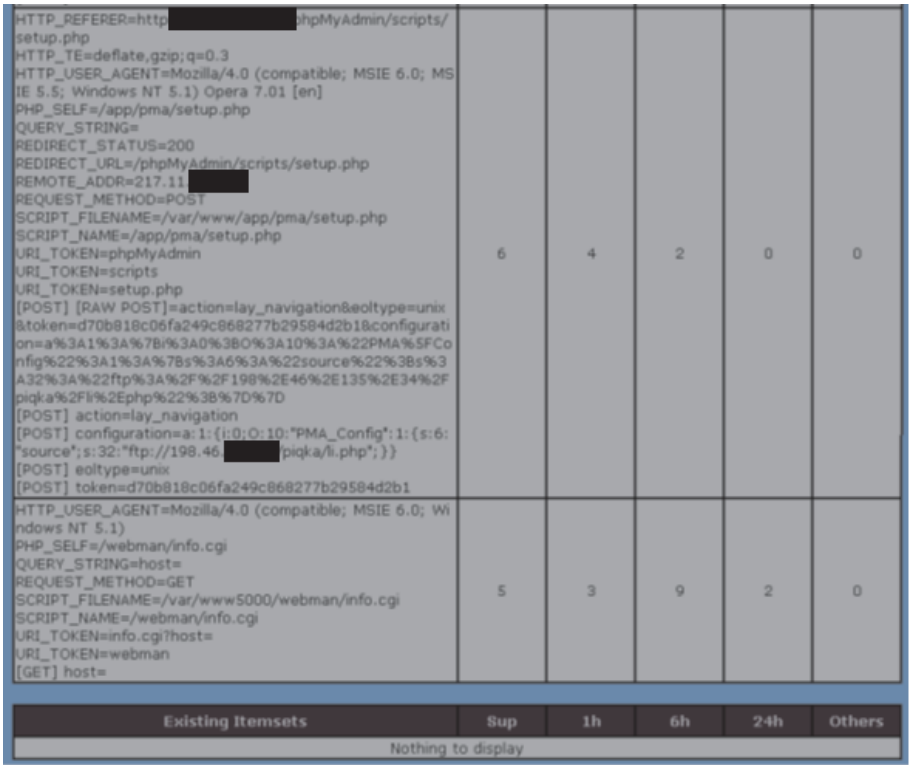
W tak przygotowanych danych zostają wykryte maksymalne zbiory częste z wykorzystaniem algorytmu Max Miner [3]. Dla danych przedstawionych w tabeli 2 i parametrowi minimal support ustalonemu na wartość 3 zostaną wykryte dwa zbiory częste. Pierwszy (REMOTE_ADDR= 217.11.XX.YY, URI_TOKEN=scripts, URI_TOKEN=setup.php) reprezentuje aktywność z adresu 217.11.XX.YY skierowaną na aplikację posiadającą plik setup.php w katalogu scripts. Drugi zbiór częsty (URI_TOKEN=webman, URI_TOKEN=info.cgi?host=) reprezentuje skanowanie przeprowadzane z różnych adresów, o czym świadczy brak elementu zbioru związanego z adresem IP. Skanujące maszyny próbują wykryć czy w danej domenie jest pliki info.cgi znajdujący się w katalogu webman. Biorąc pod uwagę stały napływ nowych danych do systemu HoneyPot oraz mając na uwadze zmiany aktywności różnych działań w czasie chcieliśmy stworzyć system pozwalający je wykrywać. Algorytmy wykrywania zbiorów częstych dokonują wykrywania wzorców w całym dostępnym zbiorze danych. W związku z tym, wszystkie nadchodzące dane dzielimy ze względu na czas ich zarejestrowania i na takich fragmentach danych dokonujemy wykrywania wzorców częstych. W aktualnej wersji systemu HPMS cyklicznie, co godzinę, sześć godzin, raz na dobę oraz tygodniowo wykonywane jest wykrywanie zbiorów częstych. Zastosowanie różnych interwałów

Tabela 2
Przykładowy zbiór danych uzyskany przez sensor WebHP w formie zbliżonej do surowych danych oraz w formie dogodnej do analizy wykorzystującej wyszukiwanie zbiorów częstych

Identyfikator Transakcji	Dane w formie zbliżonej do surowych danych	Dane w postaci dogodnej do analizy – jako zbiory
1	REMOTE_ADDR=217.11.XX.YY URI_TOKEN=phpmyadmin URI_TOKEN=scripts URI_TOKEN=setup.php	(1, 2, 3, 4)
2	REMOTE_ADDR=217.11.XX.YY URI_TOKEN=phpMyAdmin URI_TOKEN=scripts URI_TOKEN=setup.php	(1, 5, 3, 4)
3	REMOTE_ADDR=219.129.AA.BB URI_TOKEN=webman URI_TOKEN=info.cgi?host=	(6, 7, 8)
4	REMOTE_ADDR=217.11.XX.YY URI_TOKEN=pma URI_TOKEN=scripts URI_TOKEN=setup.php	(1, 9, 3, 4)
5	REMOTE_ADDR=58.20.CC.DD URI_TOKEN=webman URI_TOKEN=info.cgi?host=	(10, 7, 8)
6	REMOTE_ADDR=211.27.EE.FF URI_TOKEN=webman URI_TOKEN=info.cgi?host=	(11, 7, 8)

czasowych pozwala wykrywać aktywności o różnej częstotliwości. Przykładowo, wzorce wykryte w czasie jednej godziny reprezentują dość agresywne zachowania, natomiast zastosowanie analizy okresu tygodnia pozwala wykryć np. utajone wolne skanowania. Wykrycie pewnego wzorca w okresie o krótszym interwale implikuje, że zostanie on wykryty także w dłuższym interwale. Aby nie dopuścić do sytuacji, że liczne, wcześniej wykryte wzorce w krótszym interwale przysłonią nam nowe, mniej liczne wykryte w dłuższym interwale, przy prezentacji uzyskanych wyników, najpierw pokazywane są nowe wzorce, a później znajduje się lista już wcześniej wykrytych. Dodatkowo, wprowadzenie statystyki dotyczącej ile razy dany wzorzec został wykryty w interwałach o określonej długości ułatwia analitykowi ocenę z jakim rodzajem aktywności związany jest dany wzorzec. Przykładowo, wzorce dotyczące dość intensywnej aktywności w najkrótszym interwale powtarzające się przez kilka sąsiadujących ze sobą interwałów mogą świad-

czyć o próbie przeprowadzenia ataku odmowy usługi. Rysunek 2 prezentuje przykładowy zrzut ekranu z wykrytymi dwoma zbiorami częstymi.



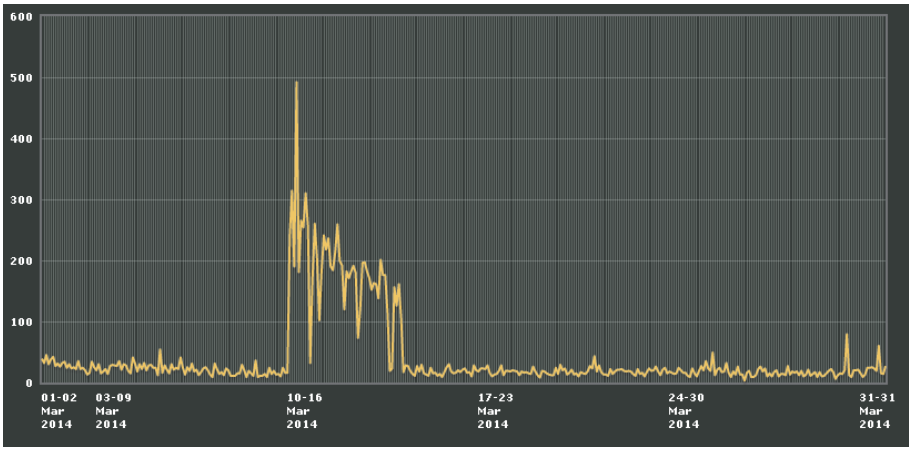
Rysunek 2. Zrzut ekranu z systemu HPMS prezentujący dwa przykładowe zbiory

4. Przeprowadzone eksperymenty

Wykorzystując oprogramowanie opisane w poprzednim punkcie w ramach przeprowadzonych eksperymentów cały ruch zarejestrowany przez system HoneyPot pomiędzy pierwszym marca 2014 a końcem kwietnia 2014 został przeanalizowany w celu wykrycia wzorców. W tym czasie zostało zarejestrowanych ponad 25 tysięcy połączeń do trzech sensorów WebHP. Na potrzeby tych eksperymentów parametr minimalnego wsparcia przyjmował wartość pięć - w wykorzystanej implementacji oznacza to, że pięciokrotne wystąpienie jakiegoś podzbioru oznacza uznanie go za zbiór częsty.

W uzyskanych danych zostały wykryte maksymalne zbiory częste. W wyniku uzyskano niecałe 1050 wzorców, z czego ponad 320 wystąpiło jedynie jednokrotnie, co może sugerować, że te zbiory częste powstały z przypadkowego połączenia zebranych danych. W aktualnej implementacji nie zastosowano żadnego filtrowania uzyskanych zbiorów, przykładowo ze względu na wystąpienie lub brak wystąpienia określonych atrybutów w wykrytym zbiorze. Z tego powodu wykryto szereg niezbyt istotnych wzorców, przykładowo zbiór, który zawiera jedynie atrybuty powiązane z odpowiedzią serwera związaną z brakiem danej strony. Interpretacją takiego zachowania jest złączenie w jeden zbiór częsty pięciu niezależnych połączeń z różnych maszyn, których jedyną cechą wspólną jest to, że nie powiodło się otrzymanie żądanej strony z powodu jej braku na serwerze. Jednak wiele z wykrytych wzorców odpowiadało istotnym i powtarzalnym aktywnościom obserwowanym przez wdrożoną instalację systemu HoneyPot.

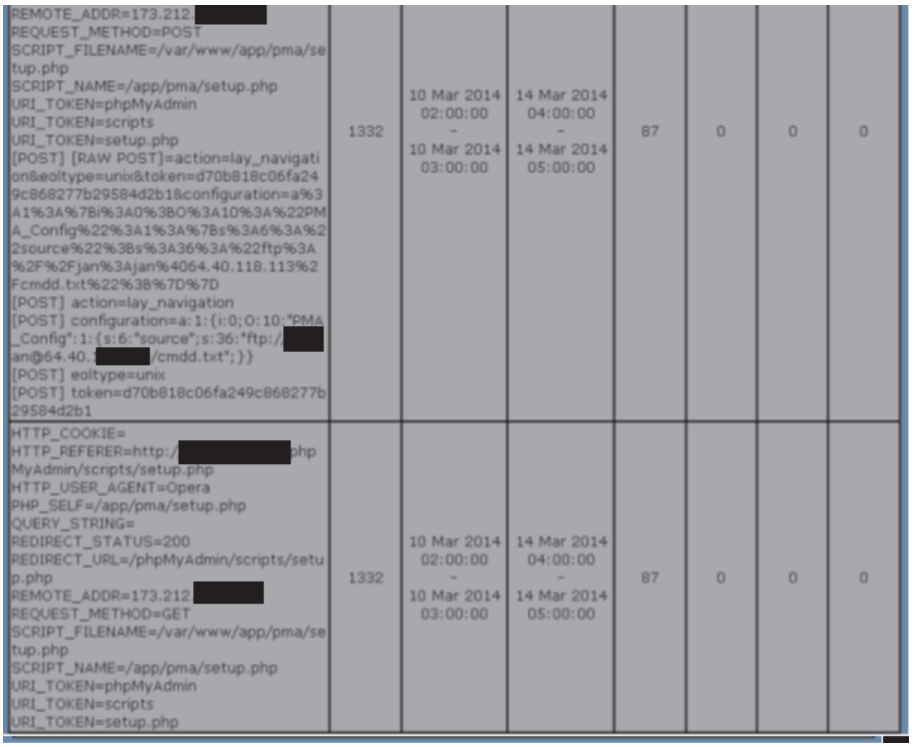
Przykładowo około 10 marca zaobserwowana została wzmożona aktywność na jednym z sensorów symulującym udostępnienie w Internecie dostępu do aplikacji „PHP My Admin". Rysunek 3 przedstawia zrzut ekranu z systemu HPMS zawierający całą aktywność zaobserwowaną w marcu 2014.



Rysunek 3. Zrzut ekranu z systemu HPMS prezentujący wykres aktywności obserwowanej przez system w marcu 2014

Równocześnie 10 marca o godzinie 3:00, podsystem wykrywania zbiorów częstych, wykrył 7 zbiorów, zawierających ten sam zdalny adres. Analiza wszystkich uzyskanych zbiorów częstych w czasie omawianego eksperymentu pokazała, że sześć zbiorów częstych zostało wykrytych 87 razy

w okresach o długości 1 godziny i ich sumaryczne wsparcie jest zawsze równe 1332. Dodatkowo wszystkie zostały pierwszy raz wykryte między 2:00 a 3:00 10 marca a ostatni raz między 4:00 a 5:00 14 marca. Zrzut ekranu pokazujący dwa przykładowo wykryte zbiory częste z tej aktywności zaprezentowany jest na rysunku 4.



Rysunek 4. Zrzut ekranu z systemu HPMS prezentujący dwa zbiory częste wykryte po raz pierwszy 10 marca między godziną 2:00 a 3:00

Manualna analiza wszystkich zarejestrowanych danych wykazała, że adres atakującego 172.212.XX.YY został zarejestrowany 8076 razy, pierwszy raz o 2:06 10 marca a ostatni o 5:17 14 marca 2014. Dane te potwierdzają bardzo dużą dokładność wykrytych automatycznie wzorców, można zauważyć, że $6 * 1332 = 7992$, a czas początku i końca wykrycia aktywności jest prawie identyczny. Podczas analizy danych z tych dwóch miesięcy wykryto jeszcze kilka adresów, które masowo próbowały dokonywać ataków na wybrane sensory systemu HoneyPot. Wykrycie faktu masowego ataku z jednego adresu IP jest możliwe za pomocą dostępnych już od lat syste-

mów analizujących logi serwerów lub zapór ogniowych. Jednak zastosowanie zaproponowanej metody wykorzystującej zbiory częste pozwala wykrywać bardziej finezyjne zmiany w postępowaniu atakujących. Przykładowo, w okresie dokonywania analizy wykryto ponad 25 zbiorów częstych zawierających atrybut związany z jednym adresem IP - 217.11.XX.YY. Analiza uzyskanych wyników pokazała, że w czasie dwutygodniowej aktywności, podczas ataków wielokrotnie były zmieniane adresy serwerów i nazwy plików zawierających ściąganego po infekcji oprogramowania bota. Tak niewielkich z punktu widzenia ruchu sieciowego a bardzo istotnych z punktu widzenia zmian postępowania atakującego nie byłyby w stanie wykryć metody bazujące na prostym zliczaniu ruchu pochodzącego od wybranego adresu IP.

Omawiane do tej pory wzorce dotyczyły pojedynczych adresów dodatkowo wykazujących dość intensywną działalność. Dopiero przy analizie danych występujących rzadziej można zaobserwować zalety zaproponowanej metody. Przykładem ciekawej aktywności wykrytej przez zaimplementowany system, jest skanowanie w poszukiwaniu podatnych na atak urządzeń typu NAS (ang. Network Attached Storage) firmy Synology. Rysunek 5 przedstawia wykryty przez system analizy zbiór częsty reprezentujący daną aktywność.

HTTP_USER_AGENT=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1) PHP_SELF=/webman/info.cgi QUERY_STRING=host= REQUEST_METHOD=GET SCRIPT_FILENAME=/var/www5000/webman/info.cgi SCRIPT_NAME=/webman/info.cgi URI_TOKEN=info.cgi?host= URI_TOKEN=webman [GET] host=	119	06 Apr 2014 12:00:00 - 06 Apr 2014 18:00:00	28 Apr 2014 12:00:00 - 28 Apr 2014 18:00:00	5	12	2	0

Rysunek 5. Zrzut ekranu z systemu HPMS prezentujący zbiór częsty związany z poszukiwaniem podatnych urządzeń firmy Synology

W przeciwieństwie do poprzednio prezentowanych zbiorów częstych, które występowały tylko w okresach o jednakowej długości, można zaobserwować, że zbiór ten został wykryty odpowiednio 5 razy w okresach 1-godzinnych, 12 w okresach 6h a nawet dwa razy w okresach dobowych. Ta informacja, wraz z brakiem w wykrytym zbiorze częstym atrybutów związanych z adresem źródłowym, może potwierdzać, że zaprezentowany wzorzec dotyczy aktywności spowodowanej przez wiele niezależnych maszyn. Jeśli weźmiemy pod uwagę, że w okresie analizy zaobserwowano ponad 25 tysięcy zdarzeń, najprawdopodobniej te 119 powiązanych z poszukiwaniem danego adresu URL, związanego jednoznacznie z podatnym urządzeniem

firmy Synology, zostałoby pominiętych. Na uwagę zasługuje jeszcze jeden zestaw meta-informacji związany z wykrytym wzorcem - zapisanie czasu kiedy pierwszy i ostatni raz został zaobserwowany. Na zaprezentowanym przykładzie, pierwszy wzorzec dotyczący urządzeń firmy Synology został wykryty 6 kwietnia w okresie 6 godzinnym między 12 a 18. Pokrywa się to z czasem uruchomienia sensora dla tego zagrożenia (port nr 5000), które nastąpiło w sobotę 5 kwietnia.

Analiza uzyskanych danych pokazała jeszcze rzadsze zdarzenia, które zostały wykryte przez zaproponowany system oraz potem zweryfikowane jednoznacznie przez analityka jako powiązane z działalnością atakujących. Na rysunku 6 zaprezentowany jest wykryty wzorzec reprezentujący skanowania związane z robakiem „The Moon" atakującym urządzenia sieciowe firmy Linksys. Odpytania dotyczące adresu URL „HNAP1" związane są z próbą pobrania informacji dotyczących dokładnej konfiguracji urządzenia i pozwalających zweryfikować, czy jest ono podatne na atak przepełnienia bufora w jednym ze skryptów interfejsu graficznego.

PHP_SELF=/errors/404.html QUERY_STRING= REDIRECT_REQUEST_METHOD=GET REDIRECT_STATUS=404 REDIRECT_URL=/HNAP1/ REQUEST_METHOD=GET SCRIPT_FILENAME=/var/www/errors/404.html SCRIPT_NAME=/errors/404.html URI_TOKEN= URI_TOKEN=HNAP1	47	24 Feb 2014 00:00:00 - 03 Mar 2014 00:00:00	30 Mar 2014 00:00:00 - 13 Apr 2014 00:00:00	0	0	0	6

Rysunek 6. Zrzut ekranu z systemu HPMS prezentujący zbiór częsty związany z aktywnością robaka „The Moon"

W przypadku tego wzorca w czasie eksperymentu zaobserwowano 47 zdarzeń i wszystkie były obserwowane w oknach siedmiodniowych. Dodatkowo z informacji powiązanych z tym wzorcem można zaobserwować, że około połowy kwietnia aktywność tego zagrożenia przestała być praktycznie obserwowalna. Potwierdza to manualna analiza wszystkich dostępnych danych, w której w okresie od 15 do 29 kwietnia wykryto jedynie cztery zarejestrowane zdarzenia tego typu - po dwa 20 i 23 kwietnia.

5. Podsumowanie

W artykule został omówiony zaproponowany i wdrożony w Instytucie Informatyki Politechniki Warszawskiej prototyp systemu HoneyPot dedykowany aplikacjom Webowym wraz z systemem analizy wyników wykorzystującym metody eksploracji danych. Wykorzystując zaimplementowane

oprogramowanie, wstępnej analizie zostały poddane wszystkie zarejestrowane dane od początku marca do ostatnich dni kwietnia - razem ponad 25 tysięcy rekordów. System automatycznie wykrył około 1050 wzorców. Wśród wykrytych wzorców znajdowały się takie, które dotyczyły masowych aktywności, przykładowo zawierających ponad 8000 transakcji. Jednak system także wykrył wzorce, w porównaniu z poprzednimi występujące nie tak często, a związane z kilkudziesięcioma połączeniami, jednak powtarzającymi się przez okres kilku tygodni. Wśród uzyskanych wyników w związku ze specyfiką działania wykorzystanej metody znalazły się nieistotne wzorce, przykładowo takie, które związane są z kodem błędu w przypadku braku na serwerze żądanego zasobu od różnych, niepowiązanych ze sobą klientów. Jednak po ich manualnym odsianiu pozostałe wzorce okazały się bardzo ciekawe. Wśród nich wykryte zostały masowe próby ataków na różne aplikacje, skanowania w poszukiwaniu podatnych maszyn a nawet aktywność robaków internetowych. Pierwsze eksperymenty potwierdzają wstępne założenia i motywacje do zastosowania tego rodzaju algorytmów. Wykorzystanie wykrywania wzorców w danych wydatnie zmniejszyło liczbę danych, które musi przeanalizować manualnie analityk. Dodatkowo, możliwości wykrycia powtarzalnych, ale relatywnie rzadkich zdarzeń, jak omawiane skanowania w celu wykrycia urządzeń firmy Synology lub Linksys, zmniejsza szansę ich pominięcia przez analityka.

Autorzy zdecydowali się nie porównywać wyników uzyskanych z omówionego systemu z dokładnością i wykrywalnością oferowaną przez dostępne systemy antywirusowe i wykrywania włamań. Na decyzję miał fakt iż większość wykrytych zdarzeń dotyczyła zagrożeń klasy „zero-day-exploit" – wcześniej nie znanych producentom oprogramowania bezpieczeństwa. W związku z tym na pewno nie byłyby one wykryte przez systemu bazujące na manualnie przygotowywanych szczepionkach systemów antywirusowych czy regułach systemów IDS/IPS.

Na podstawie doświadczeń z przeprowadzonych eksperymentów rozpoczęte zostały prace pozwalające dokonać filtracji nieistotnych wzorców, jak również połączyć różne wzorce, przykładowo związane z jednym adresem IP lub zasobem URL. Dodatkowo rozważane jest zastosowanie bardziej skomplikowanych wzorców, przykładowo sekwencji częstych oraz epizodów.

Literatura

[1] R. Agrawal, T. Imielinski, A Swami, *Mining Association Rules Between Sets of Items in Large Databases*, Proceedings of ACM SIGMOD Int. Conf. Management of Data, (1993).

[2] R. Agrawal, R. Srikant, *Mining Sequential Patterns: Generalizations and Performance Improvements*, In Proceedings of the Fifth International Conference on Extending Database Technology (EDBT), (1996).

[3] R. J. Bayardo, *Efficiently mining long patterns from databases*, In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98), Seattle, WA, pp. 85–93, (1998).

[4] M. L. Bringer, C. A. Chelmecki, H. Fujinoki, *A Survey: Recent Advances and Future Trends in Honeypot Research*, I. J. Computer Network and Information Security, 10, 63–75, (2012).

[5] W. Cheung, O. Zaïane, *Incremental Mining of Frequent Patterns Without Candidate Generation or Support Constraint*, 7th International Database Engineering and Applications Symposium (IDEAS 2003), Hong Kong, China. IEEE Computer Society, (2003).

[6] X. Fu, W. Yu, D. Cheng, X. Tan, K. Streff, and S. Graham, *On Recognizing Virtual Honeypots and Countermeasures*, Proceedings of the IEEE International Symposium on Dependable, Autonomic and Secure Computing, pp. 211-218, (2006).

[7] J. Han, J. Pei, Y. Yin, *Mining Frequent Patterns without Candidate Generation*, Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, Texas, United States, (2000).

[8] The Honeynet Project, Know Your Enemy, learning about security threats, Addison-Wesley, ISBN 0-321-16646-9, (2004).

[9] N. Provos, T. Holz, *Praise for virtual HoneyPots*, Pearson Education, ISBN 978-0-321-33632-3, (2007).

[10] C. Seifert, I. Welch, P. Komisarczuk, *Taxonomy of Honeypots*, CS Technical Report TR-06-12, School of Mathematics, Statistics and Computer Science, Victoria University of Wellington, New Zealand., (2006).

[11] Ł. Skonieczny, *Mining for Unconnected Frequent Graphs with Direct Subgraph Isomorphism Tests*, w: Man-Machine Interactions / K. A. Cyran i in. (red.), *Advances in Intelligent and Soft Computing*, vol. 59, 2009, Springer, ISBN 978-3-642-00562-6, ss. 523–531, DOI:10.1007/978-3-642-00563-3-55

ANALYSIS OF THE HONEYPOT SYSTEM DATA USING DATA MINING TECHNIQUES

Abstract. The HoneyPot systems are used From several years to gather data concerning novel attacks appearing in the Internet. Despite the fact that new types of HoneyPots are developed, there is a lack of analytical software, which can be used for analysis of data provided by this kind of systems. The article contains a description of the WebHP/HPMS (HoneyPot Management System) which allows analysis of HoneyPot gathered data. Additionally, the article presents used data mining techniques and conducted experiments. Preliminary results appeared to be very promising. In the vast amounts of data, discovered patterns rapidly reveal signs of new types of attacks.

Keywords: HoneyPot systems, data analysis, data mining