dr hab. inż. Wojciech Mazurczyk

Instytut Telekomunikacji
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Warszawa, 18.08.2015

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY WYDZIAŁU ELEKTROTECHNIKI, AUTOMATYKI, INFORMATYKI I INŻYNIERII BIOMEDYCZEJ AKADEMII GÓRNICZO-HUTNICZEJ W KRAKOWIE

Tytuł rozprawy: **Entropy-based network anomaly detection**

Detekcja anomalii w ruchu sieciowym z wykorzystaniem miar entropijnych

Autor rozprawy: **mgr inż. Przemysław Bereziński**

1. Problematyka naukowa oraz przedmiot rozprawy

Tematyka rozprawy dotyczy detekcji anomalii sieciowych, a w szczególności koncentruje się na wykrywaniu incydentów mających wpływ na bezpieczeństwo sieci teleinformatycznych. Tematyka ta jest ważna i akutalna, ponieważ w ostatnich latach można zaobserwować lawinowy wzrost udanych cyberataków na infrastrukturę sieciową firm, instytucji, państw, a także użytkowników indywidualnych. Realnym skutkiem takiego stanu rzeczy może być w przyszłości spadek zaufania użytkowników do wykorzystywania sieci teleinformatycznych, a co za tym idzie potencjalne straty socjo-ekonomiczne. Należy także podkreślić, że obecnie istniejące systemy zabezpieczeń są w stanie wykrywać, jak się szacuje, jedynie ok. 50% zagrożeń, co dodatkowo stanowi o istotności badań nad nowymi sposobami detekcji ataków sieciowych.

Detekcja anomalii sieciowych to bardzo szeroka dziedzina nauki, w której pierwsze metody opracowano prawie 40 lat temu. Od tego czasu zarówno sposoby wykrywania anomalii, ale przede wszystkim ataki sieciowe ewoluowały w sposób znaczący. Należy także zaznaczyć, iż obecnie nie ma jednego sposobu detekcji, który byłby skuteczny dla każdego rodzaju zagrożeń, tworzone są raczej dedykowane rozwiązania efektywne dla określonych typów anomalii związanych z aktywnością potecjalnego cyberprzestępcy.

Recenzowana rozprawa doktorska poświęcona jest metodom detekcji bazującym na miarach entropijnych, które w ostatnim czasie zyskują na popularności w literaturze światowej.

W rozprawie Autor stawia tezę, w której stwierdza co następuje:

„Podejście bazujące na miarach entropijnych jest odpowiednie do wykrywania anomalii charakterystycznych dla złośliwego oprogramowania typu botnet na podstawie analizy ruchu w sieciach lokalnych."

Uważam, że teza rozprawy została postawiona prawidłowo i jest jasno sformułowana.

Żeby osiągnąć postawiony cel, Doktorant najpierw dokonał porównania kilku miar entropijnych pod kątem ich przydatności do detekcji anomalii sieciowych. Następnie zaproponował, zaimplementował i zweryfikował autorską metodę bazującą na

parametryzowanych entropiach i nadzorowanym uczeniu maszynowym.

Ocena efektywności zaproponowanego rozwiązania została przeprowadzona na podstawie reprezentatywnego zbioru danych stworzonego przez Doktoranta na potrzeby rozprawy, ze względu na to, że obecnie publicznie dostępne zbiory danych, które można by wykorzystać do testowania metod detekcji ataków sieciowych są niekompletne i/lub przestarzałe.

Dodatkowo, Autor przeprowadził szczegółowe badania zaproponowanego rozwiązania pozwalające na wyznaczenie najbardziej efektywnych wartości parametrów konfiguracyjnych metody, stosownych cech ruchu sieciowego oraz rodzaju wykorzystywanego klasyfikatora.

Przeprowadzone badania eksperymentalne skuteczności zaproponowanej metody detekcji anomalii wykazały, że wykorzystanie parametryzowanej entropii Renyi'ego lub Tsallisa razem z klasyfikatorem bazującym na regresji logicznej pozwalają na efektywne wykrywanie anomalii związanych z aktywnością złośliwego oprogramowania lub atakującego przy jednoczesnym zachowaniu niskiego poziomu fałszywych alarmów. Doktorant porównał także zaproponowane przez siebie rozwiązanie z metodami bazującymi na: 1) entropii Shannona oraz 2) podejściu wolumenowym wykorzystującym proste liczniki takie jak: liczba przepływów, pakietów i bajtów. Wyniki obu wymienionych rozwiązań okazały się gorsze (w niektórych przypadkach znacznie) od osiągniętych przez metodę detekcji zaproponowaną w rozprawie.

2. Analiza treści rozprawy oraz uzyskanych wyników

Rozprawa została przygotowana w języku angielskim i składa się z 9 rozdziałów, oraz bibliografii. Całość pracy obejmuje ok. 100 stron. Wyniki własne doktoranta przedstawione są w rozdziałach 4-9.

Pierwszy rozdział rozprawy stanowi wprowadzenie do poruszanej problematyki i obejmuje przedstawienie motywacji, tezy i celu pracy oraz oryginalnego wkładu Doktoranta w aktualny stan wiedzy w dziedzinie detekcji anomalii sieciowych.

Stan dotychczasowych badań w głównym zakresie tematycznym rozprawy obszernie przedstawiono w rozdziale stanowiącym wprowadzenie do pracy (rozdz. 2) oraz częściowo także w innych rozdziałach (np. w rozdziale 5). Rozprawa zawiera ok. 180 pozycji bibliografii. Doktorant wykazał się znakomitą znajomością stanu wiedzy w dziedzinie będącej przedmiotem pracy oraz umiejętnością analizy literatury i poprawnego formułowania wniosków na jej podstawie.

W rozdziale 3 Autor przedstawił ogólny opis proponowanej metody, wskazał jej cechy oraz umiejscowił na tle ogólnej klasyfikacji tego typu rozwiązań.

Rozdział 4 wprowadza teorię i definicje związane z miarami entropijnymi. Zawiera także porównanie miar entropijnych Shannona, Renyi'ego i Tsallisa uzyskane na podstawie eksperymentów symulacyjnych.

Szczegółową koncepcję zbierania i analizy ruchu sieciowego na potrzeby przeprowadzonych w ramach rozprawy badań eksperymentalnych zaprezentowano w rozdziale 5. Wykorzystane podejście zostało zestawione i przedyskutowane na tle innych dostępnych rozwiązań.

W kolejnym rozdziale Doktorant zwięźle przedstawia architekturę oraz sposób działania zaproponowanej metody detekcji złośliwego oprogramowania. Ponadto, scharakteryzowana została implementacja kompleksowego narzędzia o nazwie Anode, w którym zrealizowano moduł detekcji bazujący na parametryzowanej entropii.

Rozdział 7 przedstawia motywację oraz sposób przygotowania zbioru danych testowych wykorzystanych do oceny efektywności zaproponowanej metody detekcji anomalii sieciowych. Autor krok po kroku opisuje źródło i charakterystykę ruchu sieciowego oraz w jaki sposób wygenerowany został profil ruchu normalnego (tj. nie zawierającego anomalii). Następnie Doktorant zaproponował trzy scenariusze (a co za tym idzie trzy zbiory danych) imitujące funkcjonowanie sieci botnet. Wspomniane zbiory danych zawierają „wymieszany" prawdziwy ruch sieciowy wraz z wygenerowanymi sztucznie atakami takimi jak: skanowanie, ataki typu „brute force", ataki (D)DoS oraz komunikacja w ramach sieci botnet.

Dla każdego ze wspomnianych powyżej scenariuszy w rozdziale 8 Autor zawarł ocenę skuteczności zaproponowanej metody detekcji anomalii dla różnych rodzajów wykorzystanych klasyfikatorów. Przedstawono w nim także wyniki korelacji mające na celu wskazanie właściwego zbioru parametrów dla miar entropijnych oraz optymalnego zbioru cech ruchu sieciowego dla skuteczniejszego działania tego rozwiązania.

W rozdziale 9 Doktorant dokonał podsumowania wykonanych eksperymentów i uzyskanych rezulatów a także przedstawił potencjane, dalsze kierunki badawcze.

3. Najistotniejsze osiągnięcia przedstawione w rozprawie

Oceniając dorobek rozprawy stwierdzam, że mgr inż. Przemysław Bereziński wniósł oryginalny wkład w dziedzinę wykrywania anomalii sieciowych. Na podstawie aktualnego stanu wiedzy zawartego w literaturze oraz własnych doświadczeń i przemyśleń, w rozprawie doktorskiej opracował koncepcję, przeprowadził implementację oraz badania eksperymentalne efektywności autorskiej metody detekcji anomalii w ruchu sieciowym bazującej na parametryzowanych entropiach Tsallisa i Renyi'ego. Metoda ta jest wartościowa, gdyż koncentruje się na próbie detekcji zagrożeń, które nie generują znacznego wolumenu ruchu (co jest zgodne z obserwowalnym obecnie trendem rozwoju zagrożeń sieciowych).

W celu realizacji zaproponowanej metody oraz wykazania jej skuteczności Doktorant w szczególności:

1) wykonał reprezentatywny zbiór danych testowych zawierający rzeczywisty ruch sieciowy połączony z syntetycznymi anomaliami reprezentującymi działanie złośliwego oprogramowania w ramach sieci botnet.

2) przeprowadził implementację systemu, w którym proponowana metoda mogła zostać przetestowana w praktyce.

3) opracował narzędzie pozwalające w statystyczny sposób modelować anomalie sieciowe.

4) poprzez badania eksperymentalne wskazał odpowiedni zestaw wartości parametru α dla wykorzystywanych entropii Tsallisa i Renyi'ego.

5) zaproponował szeroki zakres cech ruchu sieciowego dla metody i uzasadnił

zasadność jego stosowania.

6) wykazał wyższość proponowanego rozwiązania nad metodami wykorzystującymi entropię Shannona czy miary wolumenowe.

7) dla zaproponowanej metody detekcji empirycznie dobrał odpowiedni klasyfikator.

8) zaproponował autorską miarę oceny skuteczności klasyfikacji typu multi-class.

Ponadto, co warte podkreślenia, rezultaty zawarte w rozprawie zostały przedstawone i opublikowane przez Autora w siedmiu współautorskich artykułach naukowych (w tym w czterech występuje on jako pierwszy autor). Na ten dorobek składa się sześć publikacji na konferencjach o zasięgu międzynardowym (za jedną z nich Doktorant otrzymał nagrodę za najlepszy artykuł) oraz jedna w czasopiśmie naukowym z tzw. listy filadelfijskiej (Entropy, IF=1.5, 17(4), 2015).

Cieszy także, wskazana przez Doktoranta w ostatnim rozdziale rozprawy, znaczna liczba dalszych, ciekawych kierunków badawczych, które mogą być realizowane w przyszłości w oparciu o doświadczenia zebrane przy opracowywaniu niniejszej rozprawy.

4. Uwagi merytoryczne, kwestie dyskusyjne

Rozprawa jako całość nie ma istotnych wad. Niemniej jednak w trakcie jej czytania nasuwają się pewne uwagi o charakterze dyskusyjnym. Należą do nich:

- Tworzenie danych testowych dla zaproponowanej metody detekcji anomalii zostało przeprowadzone poprzez „pomieszanie" prawdziwego ruchu sieciowego ze sztucznie wygenerowanym ruchem zawierającym ataki/anomalie. Zdając sobie sprawę z braku dostępności rzetelnych zbiorów danych zawierających prawdziwy ruch wraz z prawdziwymi atakami sieciowymi (co jest szerszym problemem dotyczącym całej dziedziny detekcji anomalii) powstaje pytanie na ile taki sposób generowania danych wpływa na wyniki detekcji? Rozwiązaniem w tym przypadku mogłoby być przedstawienie wyników dokładności wykrywania ataków sieciowych (dla tego samego zbioru danych testowych) dla innych (najlepszych) znanych z literatury metod wykrywania anomalii.

- Użyte przez Doktoranta dane testowe wcale nie zawierają anomalii (prawdziwy ruch sieciowy), bądź zawierają anomalie odwzorowujące ataki sieciowe. Jaki wpływ na rezultaty detekcji miałoby występowanie w ruchu anomalii nie będących atakami sieciowymi np. typu „flash crowd", itp.?

- Doktorant zakłada wykrywanie aktywności sieci typu botnet, natomiast skupia się w zasadzie na wykrywaniu oddzielnie poszczególnych faz jej działania (np. skanu portów, ataku typu „brute force", czy komunikacji w ramach kanału C&C) bez próby korelacji tych złośliwych aktywności ze sobą. W takiej sytuacji, wykryte ataki wcale nie muszą być być rezultatem działania sieci botnet, ale mogą być zdarzeniem jednostkowym spowodowanym w inny sposób przez atakującego (które także oczywiście należy wykrywać!).

- Warto byłoby przeprowadzić badania zaproponowanej metody detekcji dla większego zestawu danych wykorzystywanych do utworzenia profilu ruchu normalnego jak i dla ruchu ze wstrzykniętymi anomaliami (np. co najmniej tydzień/miesiąc ruchu dla utworzenia profilu ruchu normalnego i kolejny tydzień/miesiąc ruchu z anomaliami).

- W założonych trzech scenariuszach badań, a szczególnie w scenariuszach 2 i 3 istnieje regularność przeprowadzania sekwencji ataków (równo co godzinę). Czy taka regularność może wpływać na wyniki detekcji?

- Uzyskane w rozprawie wyniki detekcji anomalii warto byłoby porównać z rezultatami istniejących prac (wskazanych przez Doktoranta w rozdziale 2), w których także wykorzystano parametryzowaną entropię do detekcji ataków sieciowych np. (D)DoS, czy skanowania (prace m.in. Shafiq i in., Lima i in., Tellenbach i in, Yang i in.). Pozwoliło by to na pokazanie na ile uzyskane w rozprawie rezultaty są zgodne z tymi uzyskanymi przez inne grupy badawcze.

- Do detekcji anomalii wykorzystywane jest okno przesuwne, na podstawie którego określane są minimalne i maksymalne wartości entropii. Jak dobór rozmiaru tego okna oraz sposób jego przesuwania wpływa na wyniki detekcji?

5. Uwagi redakcyjne i edytorskie

Struktura pracy jest poprawna, natomiast byłaby jeszcze lepsza, gdyby rodział 3 (o długości 3 stron i zawierający dwa duże rysunki) został umieszczony jako podrozdział lub połączony z innym rozdziałem np. rozdziałem 6 (również bardzo krótkim), który opisuje architekturę i implementację zaproponowanego systemu detekcji anomalii. Poprawiło by to czytelność pracy jako całości.

Rozprawa jest dość starannie zredagowana i napisana precyzyjnym językiem. Doktorant dobrze radzi sobie z konstrukcjami języka angielskiego. Niemniej jednak w rozprawie nadal da się odnaleźć niedociągnięcia edytorskie bądź językowe. Przykłady:

- Str. 30, zdanie "The aim of this experiments is to show, how via entropies, highlight concentration (frequent events forming the main mass) and dispersion (rare events forming the tail) caused by typical network anomalies such as port and network scans" jest dość trudno zrozumiałe i powinno uwzględniać "these experiments" zamiast "this experiments".

- Str. 34, jest "to check if particular serviceon", a powinno być "to check if a particular service on".

- Str. 38 jest "but recently it becomes the popular source", a powinno być "but recently it has become (lub ewentualnie it is becoming) the popular source".

- Str. 43 jest "and stores it into binary files", a powinno być "and stores it in binary files".

- Str. 46 jest "One of the feature of Anode", a powinno być "One of the features of Anode".

- Str. 56-57 jest "Slowrolis", a poprawna nazwa ataku to "Slowloris".

- Str. 59 jest „only o small", a powinno być "only a small".

- Str. 71 jest „which is an unacceptable", a powinno być "which is unacceptable".

- Str. 71 dwukrotnie jest "than BDR will drop to", a powinno być "then BDR will drop

to" (podobny błąd występuje także na tej samej stronie przy "classifier than classify...").

- Str. 71 jest „avaraged" a powinno być „averaged" (dwukrotnie ten sam błąd występuje na stronie 74).
- Str. 76 jest „among popular method employed in Weka", a powinno być „among popular methods employed in Weka".
- Str. 82 jest "For the feature", a powinno być "For the future".
- Str. 82 jest "We are going to put a attention", a powinno być "We are going to pay attention".

Ponadto, w pracy zdarzają się drobne niedociągnięcia, przykładowo Rys. 4.6 przedstawiający rozkład adresów IP oraz portów dla ruchu bez anomalii jest taki sam jak Rys. 4.7 obrazujący te same cechy dla ataku związanego ze skanowaniem portów, co jest oczywistą pomyłką edytorską.

6. Przydatność rozprawy dla nauk technicznych

Powyżej przedstawione uwagi merytoryczne oraz redakcyjne nie mają istotnego wpływu na jakość i wagę przedstawionych rozwiązań i w żadnym stopniu nie obniżają wartości rozprawy. Przedstawione przez Autora badane aspekty zostały ujęte wystarczająco szczegółowo i dokładnie.

Praktyczna przydatność rozprawy dla nauk techninczych jest potencjalnie duża. Detekcja anomalii szczególnie takich, które nie skutkują zbyt dużym wolumenem ruchu (co jest zgodne z istniejącym trendem tworzenia złośliwego oprogramowania przez cyberprzestępców) nie jest trywialna i tym bardziej wartościowe są rozwiązania, w tym to zaproponowane w rozprawie, umożliwiające ich wykrywanie. Warto zauważyć także, że proponowane przez Doktoranta rozwiązanie może być traktowane jako uzupełnienie istniejących metod detekcji bazujących na sygnaturach.

7. Podsumowanie.

Biorąc pod uwagę zaprezentowany dorobek naukowy Doktoranta, a w szczególności jego dorobek publikacyjny, na którym bazuje rozprawa (7 publikacji o zasięgu międzynarodowym, w tym jedna publikacja w czasopiśmie z tzw. listy filadelfijskiej) uważam, że recenzowana praca spełnia w stopniu bardzo dobrym wymagania stawiane rozprawom doktorskim przez obowiązującą ustawę o stopniach i tytule naukowym.

dr hab. inż. Wojciech Mazurczyk