RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasza na

PUBLICZNĄ OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr. inż. Piotra Białczaka

która odbędzie się w dniu **4 kwietnia 2023 roku,** o godzinie **11:00** w trybie zdalnym

Temat rozprawy:

„Wykorzystanie protokołu HTTP do identyfikacji i klasyfikacji złośliwego oprogramowania”

Promotor: dr hab. inż. Wojciech Mazurczyk, prof. uczelni – Politechnika Warszawska

Recenzenci: prof. dr hab. inż. Sławomir Sujecki – Politechnika Wrocławska

dr hab. inż. Rafał Kozik, prof. uczelni – Politechnika Bydgoska

Obrona odbędzie się zdalnie na platformie MS Teams. Osoby zainteresowane uczestnictwem w obronie proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji: dr hab. inż. Andrzeja Bęben, email: andrzej.beben@pw.edu.pl, do dnia 3.04.2023 r., godz. 23:59.

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: https://www.bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-do-30-kwietnia-2019-r/Dyscyplina-informatyka-techniczna-i-telekomunikacja-dziedzina-nauk-inzynieryjno-technicznych/mgr-inz.-Piotr-Bialczak

Przewodniczący Rady Naukowej Dyscypliny
Informatyka Techniczna i Telekomunikacja
Politechniki Warszawskiej
dr hab. inż. Jarosław Arabas, prof. uczelni

Streszczenie

Złośliwe oprogramowanie jest poważnym zagrożeniem współczesnego Internetu. Przestępcy używają go do wysyłania niechcianych wiadomości, wymuszania okupu przez zaszyfrowanie plików czy wykradania danych logowania do banku. Do komunikacji wykorzystywane są w nim popularne protokoły sieciowe, w tym często protokół HyperText Transfer Protocol (HTTP). Celem niniejszej rozprawy doktorskiej jest wykazanie, że żądania tego protokołu wygenerowane przez różne rodziny złośliwego oprogramowania mogą być użyte do ich identyfikacji i klasyfikacji. Do przeprowadzenia oceny eksperymentalnej stworzono zbiory danych ruchu sieciowego obejmujące 121 rodzin złośliwego oprogramowania oraz zestaw popularnych aplikacji niezłośliwych. Przeprowadzone badania podzielono na trzy części. W części pierwszej dokonano identyfikacji charakterystycznych cech żądań HTTP umożliwiających odróżnienie złośliwego oprogramowania od aplikacji niezłośliwych. Cechy te stały się bazą do drugiej części analizy, której efektem było stworzenie narzędzia *Hfinger* umożliwiającego tworzenie unikalnych reprezentacji żądań HTTP. Reprezentacje te można wykorzystać do identyfikacji złośliwego oprogramowania przez rozróżnienie jego rodzin, a także jego konkretnych działań, np. ataków lub pobierania rozkazów. W części trzeciej skupiono się natomiast na problemie klasyfikacji złośliwego oprogramowania przy użyciu algorytmów uczenia maszynowego, tzn. przypisania nazw konkretnych rodzin do analizowanego ruchu sieciowego. Problem ten został rozszerzony o rozpoznanie obecności klas, które były nieznane w trakcie treningu klasyfikatora, czyli tzw. rozpoznawanie otwartozbiorowe (Open Set Recognition). Wykorzystano przy tym dwa sposoby reprezentacji żądań HTTP: bazujący na narzędziu *Hfinger* oraz na analizie n-gramowej. Według wiedzy autora niniejszej rozprawy jest to pierwsza praca wykorzystująca rozpoznawanie otwartozbiorowe do klasyfikacji ruchu protokołu HTTP złośliwego oprogramowania.

Słowa kluczowe: złośliwe oprogramowanie, analiza ruchu sieciowego, protokół HTTP, klasyfikacja, rozpoznawanie otwartozbiorowe.

dr hab. inż. Rafał Kozik, profesor uczelni
Politechnika Bydgoska
im. Jana i Jędrzeja Śniadeckich w Bydgoszczy,
Wydział Telekomunikacji, Informatyki i Elektrotechniki,
Al. prof. S. Kaliskiego 7,
85-796 Bydgoszcz

Bydgoszcz, 18.12.2022

RECENZJA ROZPRAWY DOKTORSKIEJ WYKONANA DLA RADY NAUKOWEJ DYSCYPLINY INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

Tytuł rozprawy: Wykorzystanie protokołu HTTP do identyfikacji i klasyfikacji złośliwego oprogramowania

Autor rozprawy: mgr inż. Piotr Białczak

1. **Jakie zagadnienie naukowe jest rozpatrzone w pracy (teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez Autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, itd.)?**

Zakres rozprawy, której Autorem jest Pan mgr inż. Piotr Białczak, dotyczy wykorzystania protokołu HTTP do identyfikacji i klasyfikacji złośliwego oprogramowania. Autor niniejszej pracy zakłada również, iż możliwym jest rozpoznanie rodziny tego oprogramowania oraz wskazanie obecności przedstawicieli tych rodzin, których (w danej chwili) pochodzenie nie jest znane. Wymaga to istnienia pewnych cech w żądaniu HTTP, które występują dla oprogramowania złośliwego, ale są niezwykle rzadkie dla oprogramowania niezłośliwego.

W rozdziale pierwszym, Autor rozprawy precyzyjnie wskazał kluczowe wyzwania i problemy związane z tematyką rozprawy. W szczególności, Autor słusznie wskazuje jak istotnym zagrożeniem dla użytkowników Internetu jest złośliwe oprogramowanie i jednym ze skutecznych narzędzi do walki z nim jest efektywna jego identyfikacja.

Ponadto, w rozdziale 1.2 Autor rozprawy bardzo szczegółowo nakreślił motywację i cel badań. Zidentyfikowane zostały kluczowe problemy dotychczasowych badań nad detekcją i klasyfikacją złośliwego oprogramowania przy użyciu protokołu HTTP. W szczególności Autor rozprawy słusznie wskazuje na brak możliwości pracy w tzw. scenariuszu otwarto-zbiorowym.

W świetle niniejszego wprowadzenie, określenia celu badań oraz motywacji, postawiona zostaje teza rozprawy. Jest ona sformułowana przez Autora w sposób jasny.

W mojej ocenie rozprawa ma charakter teoretyczno-eksperymentalny. W rozdziale 1.3, Autor rozprawy definiuje kilka szczegółowych zagadnień badawczych (tzw. podtez), dla których jasno zostaje zdefiniowany sposób rozwiązania (udowodnienia). Dla każdego z zagadnień zaplanowany i wykonany zostaje odpowiedni eksperyment.

2. **Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle) świadczący o dostatecznej wiedzy Autora. Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?**

W mojej ocenie analiza literaturowa podzielona została na dwie części: ogólny zarys problematyki związanej z rozprawą doktorską (rozdział 1) oraz właściwa analiza i przegląd aktualnego stanu wiedzy (rozdział 3).

Pierwsza część analizy, pozwala czytelnikowi zrozumieć kluczowe problemy i ograniczenia dotychczasowych badań nad detekcją i klasyfikacją złośliwego oprogramowania przy użyciu protokołu HTTP. W szczególności Autor rozprawy nakreśla tło badań, powód i słuszność wyboru protokołu HTTP do detekcji złośliwego oprogramowania. Pozwala to w jasny i czytelny sposób określić cel niniejszej rozprawy.

Druga część analizy (zawarta w rozdziale 3) stanowi analizę najważniejszych prac i badań, które związane są z problematyką rozprawy. Bibliografia niniejszej rozprawy zbudowana jest z ponad 100 pozycji. Blisko połowa z nich została opublikowana w latach 2017-2022. W szczególności Autor przeanalizował sporą cześć źródeł, które oblikowane zostały w latach 2020 oraz 2021. Niniejsza analiza przeprowadzona jest w sposób właściwy i nawiązuje do charakterystyki ruchu sieciowego protokołu HTTP, problematyki identyfikacji żądań oraz wykrywaniu złośliwego oprogramowania.

W mojej ocenie wnioski z analizy literatury są sformułowane w sposób jasny i są trafne. W szczególności, w świetle przedstawionego aktualnego stanu wiedzy, słuszne jest stwierdzenie, że niewiele rozwiązań bazujących na protokole HTTP ma możliwości pracy w tzw. scenariuszu otwarto-zbiorowym.

3. **Czy Autor rozwiązał postawione zagadnienia? Czy użył do tego właściwych metod i czy przyjęte założenia są uzasadnione?**

W mojej ocenie Autor w sposób właściwy rozwiązał postawione zagadnienia. W procesie tym użyte zostały odpowiednie narzędzia i metody. W szczególności badania i eksperymenty osadzone zostały na obiektywnej analizie literaturowej.

Autor rozprawy odniósł się także do aktualnych trendów i ograniczeń metod typowo wykorzystywanych przy klasyfikacji złośliwego oprogramowania. Ponadto, w rozdziale 4 przeprowadzono szczegółową analizę ruchu sieciowego protokołu HTTP. Pozwoliło to Autorowi niniejszej pracy na a sformułowanie i skonstruowanie odpowiednich cech żądań HTTP, które docelowo umożliwiły stworzenie unikatowej reprezentacji żądań złośliwego oprogramowania oraz modeli klasyfikacji rodzin złośliwego oprogramowania.

W mojej ocenie, zaplanowane i przeprowadzone eksperymenty są wiarygodnym dowodem zrealizowania przez Autora celów badawczych. Tym samym Autor dowiódł, że posiadł umiejętności związane z metodyką prowadzenia badań.

4. **Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek Autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanej przez literaturę światową?**

Oryginalność rozprawy oraz samodzielny dorobek Autora, stanowi propozycja wykorzystania analizy żądań HTTP do identyfikacji i klasyfikacji złośliwego oprogramowania. W szczególności, w trakcie prowadzanych prac badawczych, Autor rozprawy:

- Eksperymentalnie udowodnił, iż ruch sieciowy protokołu HTTP złośliwego oprogramowania zawiera specyficzne cechy, które umożliwiają odróżnienie ruchu sieciowego aplikacji złośliwych i niezłośliwych.
- Opracował nowatorskie narzędzie Hfinger, które umożliwia tworzenie unikalnej reprezentacji żądań protokołu HTTP w efektywniejszy sposób niż inne popularne narzędzia.
- Eksperymentalnie dowiódł, że reprezentacja żądań HTTP tworzona przez narzędzie Hfinger może być wykorzystana do skutecznej klasyfikacji otwarto-zbiorowej złośliwego oprogramowania.

Oceniając pozycję niniejszej rozprawy w odniesieniu do innych prac, można stwierdzić, iż jest ona zgodna z aktualnym stanem wiedzy oraz poziomem techniki reprezentowanej przez literaturę światową. Punktowane publikacje, w których Autor niniejszej rozprawy jest pierwszym autorem, pozwalają stwierdzić, iż zademonstrowane osiągnięcia naukowe stanowią istotny wkład w literaturze międzynarodowej.

5. **Czy Autor wykazał umiejętności poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?**

Nie mam wątpliwości, że Autor rozprawy posiada dużą wiedzę dotyczącą zagadnień związanych z tematem rozprawy doktorskiej a w szczególności w obszarze cyberbezpieczeństwa. Doktorant wykazał się przede wszystkim umiejętnościami dotyczącymi prowadzenia badań naukowych, eksperymentów oraz projektowania algorytmów.

Również zrealizowane przez Autora rozprawy eksperymenty są zgodne z dobrymi praktykami w zakresie pomiaru skuteczności systemów klasyfikacji. Autor w tym zakresie użył znanych metryk (np. F1-score) co pozwala odnieść się do innych rozwiązań dostępnych w literaturze.

Na zwrócenie uwagi zasługuje także fakt, iż Autor rozprawy w poprawny sposób posługuje się różnego rodzaju diagramami i wykresami (np. strona 72 rozprawy) do analizy danych. Pozwala to skondensować stosunkowo dużą ilość informacji na diagramie o małej powierzchni, co docelowo umożliwia wizualną eksplorację żądań HTTP.

6. **Jakie są wady i słabe strony rozprawy?**

Rolą recenzenta jest zauważenie ewentualnych niedociągnięć i mankamentów przedstawianej pracy, oraz zgłoszenie uwag, które mogą być pomocne i przydatne w dalszych pracach. W szczególności:

- Przy opracowywaniu wyników, Autor mógł przedstawić wartości średnie i odchylenia standardowe z uwzględnieniem istotności statystycznej przedstawianych różnic. Przykładowo dla tabeli 6.3 łatwiej byłoby porównać czy poszczególne metody różnią się między sobą pod względem wartości F1 oraz MCC.
- Algorytm 1 przedstawia autorską propozycję protokołu badań, której podstawą jest losowy podział danych na testowe i treningowe. Zastanawiające jest dlaczego Autor nie dokonał modyfikacji 5-krotnej walidacji krzyżowej dla problemu otwarto-zbiorowego. Pozwoliłoby to zapewnić, że przy każdej próbie testowania wykorzystywane byłyby inne próbki.
- W przypadku analizy żądań HTTP, jego ciało (payload) reprezentowane jest tylko przez trzy wartości (obecność znaków specjalnych, wartość entropii Shannona oraz długość). Zastanawiające jest, czy Autor rozważał bardziej szczegółowy zapis cech, który pozwoliłby uchwycić więcej detali (argumenty, format, itd.).
- W pracy można także dopatrzeć się kilku, mało istotnych pomyłek edytorskich, takich jak: „wykorzystanietcyh cech" czy „sandboksach".

7. **Jaka jest przydatność rozprawy dla nauk technicznych?**

Pomimo przedstawionych powyżej uwag, rozprawa mgr inż. Piotra Białczaka posiada wiele silnych stron. Przede wszystkim rozprawa dotyczy ważnej tematyki z zakresu cyberbezpieczeństwa i wnosi interesujący wkład do zagadnień związanych z klasyfikacją

złośliwego oprogramowania. W szczególności, zawiera ona autorską propozycję narzędzia Hfinger. Jak pokazały wyniki z eksperymentów, proponowane rozwiązanie można wykorzystać do efektywnego identyfikowania żądań HTTP wysyłanych przez złośliwe oprogramowanie.

Na podkreślenie zasługuje również fakt udostępnienia przez Autora rozprawy kodu źródłowego, który powstał w trakcie badań. Z jeden strony pozwala to innym badaczom na przejrzyste odtworzenie najważniejszych wyników. Z drugiej strony stanowi to przyczynek do dalszych eksperymentów, proponowania innych algorytmów klasyfikacji, itd.

Również stworzony (zgromadzony) przez Autora zbiór danych ruchu sieciowego HTTP złośliwego oprogramowania (oraz aplikacji niezłośliwych), może stanowić interesujący wkład do dalszych badań i z powodzeniem powinien być wykorzystane przez innych badaczy.

Warto także zauważyć, że Autor ma udokumentowany dorobek publikacyjny w zakresie rozprawy. W publikacjach Doktorant jest pierwszym autorem, więc jego wkład można uznać za znaczący.

Wniosek

Niewątpliwie recenzowana rozprawa doktorska dowodzi dużej wiedzy Autora dotyczącej zagadnień związanych z rozprawą doktorską. Liczne i rozbudowane eksperymenty dowodzą skrupulatności i pozwalają stwierdzić, że Autor rozprawy opanował technikę planowania i prowadzenie badań naukowych. Ponadto, recenzowana praca jasno formułuje tezę, która została udowodniona poprzez badania eksperymentalne i realizację wszystkich celów badawczych.
Wobec powyższego stwierdzam, że recenzowana praca **spełnia wymagania stawiane rozprawom doktorskim** przez obowiązujące przepisy. Dlatego wnoszę o przyjęcie niniejszej rozprawy i **dopuszczenie mgr inż. Piotra Białczaka do publicznej obrony.**

Wrocław, 28.11.2022

Prof. dr hab. inż. Sławomir Sujecki
Katedra Telekomunikacji i Teleinformatyki
Politechnika Wrocławska

Recenzja rozprawy doktorskiej mgr inż. Piotra Białczaka

Wykorzystanie protokołu http do identyfikacji i klasyfikacji złośliwego oprogramowania

1. Jakie zagadnienie naukowe/badawcze jest rozpatrywane w pracy (cel i teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez autora?

W pracy sformułowano następującą **tezę**: Wykorzystanie żądań protokołu HTTP umożliwia identyfikację złośliwego oprogramowania poprzez stworzenie unikalnej reprezentacji jego ruchu sieciowego oraz skuteczne rozpoznanie rodziny złośliwego oprogramowania, a także obecności przedstawicieli rodzin dotychczas nieznanych. Sformułowano także trzy podtezy:

1. Ruch sieciowy żądań protokołu HTTP złośliwego oprogramowania zawiera cechy charakterystyczne, które umożliwiają jego odróżnienie od ruchu sieciowego aplikacji niezłośliwych.
2. Możliwe jest stworzenie unikalnej reprezentacji żądań protokołu HTTP, która umożliwia identyfikację złośliwego oprogramowania.
3. Wykorzystanie odpowiednio stworzonej reprezentacji żądań HTTP umożliwia skuteczne rozpoznanie rodzin złośliwego oprogramowania, w tym także istnienia klas dotychczas nieznanych.

Natomiast zasadniczym **celem pracy** jest stworzenie oprogramowania do detekcji i klasyfikacji złośliwego oprogramowania przy użyciu protokołu HTTP i metod uczenia maszynowego operujących w scenariuszu otwartozbiorowym. Opracowane oprogramowanie ma umożliwić klasyfikację znanych apriori klas złośliwego oprogramowania oraz wykrywać obecność klas nieznanych, tzn. takich które nie były uwzględnione w zbiorze uczącym danej metody uczenia maszynowego.

Cel i teza rozprawy zostały sformułowane jasno.

2. Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle?

W spisie literatury autor rozprawy powołał się na 117 pozycji przy czym około 25 razy odniósł się do wiodących czasopism naukowych z zakresu informatyki i telekomunikacji, t.j. IEEE Access, IEEE Transactions on Network and Service Management, IEEE Transactions on Pattern Analysis and Machine Intelligence. Ponadto spis literatury zawiera odnośniki do istotnych norm RFC, zasobów udostępnionych na portalu github, konferencji naukowych z zakresu telekomunikacji

i informatyki i książek. Autor powołał się wielokrotnie na dostępne jedynie w internecie. Co zważywszy, że tematyka pracy dotyczy cyberbezpieczeństwa w mojej opinii jest uzasadnione. Stwierdzam zatem, że w mojej opinii analiza źródeł została przeprowadzona w sposób właściwy i stanowi wierne odwzorowanie obecnego stanu wiedzy w zakresie rozważanej tematyki.

3. Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?

Autor stworzył pakiet oprogramowania do identyfikacji złośliwego oprogramowania poprzez analizę żądań protokołu HTTP. W ramach badań opracował we własnym zakresie opracował nowatorskie narzędzie Hfinger, które tworzy reprezentacje pojedynczych żądań HTTP, a to z kolei umożliwia identyfikację złośliwego oprogramowania. Przypisanie jednoznacznych i unikalnych reprezentacji do żądań wygenerowanych przez próbkę złośliwego oprogramowania pozwala stworzyć tzw. odcisk palca (ang. fingerprint), który umożliwia identyfikację i klasyfikację złośliwego oprogramowania. W drugim kroku autor zastosował metody uczenia maszynowego do otwartozbiorowej klasyfikacji złośliwego oprogramowania.

Według mojej opinii autor rozwiązał postawione zagadnienia i użył właściwej metody. Przyjęte założenia są uzasadnione.

4. Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy i poziomu techniki reprezentowanych przez literaturę światową?

Autor wykorzystał i opisał w rozprawie szereg nowych rozwiązań, w tym projekt oraz implementację narzędzia Hfinger, metody detekcji oraz klasyfikacji złośliwego oprogramowania wykrytego w ruchu HTTP. Ponadto zastosował metody uczenia maszynowego do otwartozbiorowej klasyfikacji złośliwego oprogramowania. Są to rozwiązania oryginalne i nowatorskie. Stanowią one zatem samodzielny i oryginalny dorobek autora. Zaprezentowane rozwiązanie zastosowane do wykrywania i klasyfikacji złośliwego oprogramowania nie było według mojej wiedzy wcześniej publikowane przez innych autorów w Polsce lub na świecie.

5. Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

W mojej opinii autor wykazał się umiejętnością poprawnego i przekonującego przedstawienia uzyskanych przez siebie wyników. Praca jest napisana starannie choć pozostało w tekście kilka literówek i możnaby jeszcze popracować nad poprawą stylu. Doktorant poprawnie wykorzystał tabele i różne rodzaje wykresów w celu poparcia wniosków. Istotną część 4 i piątego rozdziału stanowią artykuły napisane w języku angielskim. Takie podejście 'hybrydowe' trochę narusza ciągłość narracji w

i prawdopodobnie lepiej byloby dokonać tłumaczenia na język polski ale w mojej opinii jest akceptowalne w przypadku szybko rozwijających się dziedzin nauki takich jak cyberbezpieczeństwo.

6. Jaka jest przydatność rozprawy dla nauk inżynieryjno-technicznych?

Ogólnie opracowane rozwiązania mogą być wykorzystane w narzędziach monitorujących ruch sieciowy. Moim zdaniem zaletą zaproponowanej metody jest to, iż nie spowalnia ona przesyłu informacji w sieci gdyż polega na jedynie biernej analizie żądań protokołu HTTP.

Ponadto należy zauważyć, że bardzo ważnym aspektem rozprawy jest jej wartość praktyczna. Opracowane oprogramowanie zostało bowiem przetestowane w ramach projektu Unii Europejskiej Horyzont dla CERT Polska. Świadczy to o tym, że opracowane oprogramowanie stanowi praktyczne narzędzie do wykrywania i klasyfikacji złośliwego oprogramowania i w związku z tym są duże możliwości jego potencjalnego wykorzystania w systemach cyberbezpieczeństwa.

Dodatkowo autor umieścił własne kody źródłowe w otartych repozytoriach (github.com). Umożliwia to innym specjalistom z tej dziedziny weryfikacje wyników uzyskanych przez autora oraz wykorzystanie opracowanego oprogramowania przez innych naukowców we własnych pracach badawczych.

Biorąc pod uwagę przedstawioną przez Doktoranta rozprawę stwierdzam, że recenzowana praca spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy. Dlatego wnoszę o przyjęcie niniejszej rozprawy i dopuszczenie mgr inż. Piotra Białczaka do publicznej obrony.

Sławomir Kujawa