Warszawa, 08.10.2021

dr hab. inż. Arkadiusz Orłowski, prof. SGGW

Katedra Sztucznej Inteligencji
Instytut Informatyki Technicznej
Szkoła Główna Gospodarstwa Wiejskiego
ul. Nowoursynowska 159
02-787 Warszawa

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy:	**Inżynieria odwrotna przetwarzania informacji w sieciach złożonych z wykorzystaniem wnioskowania statystycznego**
Autor rozprawy:	**mgr inż. Robert Paluch**
Promotor:	**prof. dr hab. Janusz A. Hołyst**
Promotor pomocniczy:	**dr inż. Krzysztof Suchecki**

Recenzowana rozprawa doktorska została napisana w języku angielskim jako **Reverse engineering of information processing in complex networks using statistical inference**. Zasadnicza część rozprawy liczy 93 strony i składa się na nią sześć rozdziałów merytorycznych (w tym trzystronicowe **Summary**) oraz liczący 102 pozycje spis literatury przedmiotu (**Bibliography**). Dodatkowo, na początku rozprawy umieszczono oddzielnie numerowane elementy: **Abstract**, **Streszczenie** i **Preface**.

Warto podkreślić, że rozprawa jest w pewnym sensie podsumowaniem wyników badań otrzymanych i opublikowanych wcześniej w trzech recenzowanych artykułach naukowych. W każdym z tych wieloautorskich artykułów mgr inż. Robert Paluch jest pierwszym autorem. Fakt opublikowania w latach 2018-2020 części wyników w recenzowanych wydawnictwach o zasięgu międzynarodowym dodatkowo podkreśla ich wartość naukową oraz aktualność problematyki badawczej.

Jak stwierdzono w przedmowie (**Preface**) autor rozprawy postawił przed sobą dwa zasadnicze cele. Cel pierwszy, to **stworzenie nowych narzędzi i rozwinięcie istniejących metod służących do identyfikacji źródeł informacji** (w domyśle, informacji propagujących się w dużych sieciach złożonych). Cel drugi, zdaniem autora nie mniej ważny, to **udzielenie odpowiedzi na pytania, jakie czynniki i zjawiska wpływają na efektywność detekcji źródła informacji oraz jak można tę efektywność poprawić**. W rozprawie nie sformułowano w sposób jawny hipotez badawczych, ale można przyjąć, że chodzi właśnie o możliwość stworzenia nowych i udoskonalenia istniejących metod wykrywania źródeł informacji propagującej się w sieci, zwłaszcza pod kątem ich efektywności.

Dwa pierwsze rozdziały recenzowanej rozprawy stanowią niezbędne wprowadzenie do dalszej części pracy. Mają one w związku z tym głównie charakter kompilacyjny i przeglądowy, aczkolwiek zawierają też autorskie obserwacje i komentarze.

W rozdziale pierwszym wprowadzono definicje podstawowych pojęć z zakresu sieci złożonych, modeli propagacji informacji w tego typu sieciach, metod lokalizacji źródeł informacji oraz miar jakości detekcji takich źródeł.

W rozdziale drugim przedstawiono elementy Bayesowskiego podejścia do lokalizacji źródeł informacji. Skoncentrowano się na szczegółowym omówieniu i analizie znanego algorytmu Pinto-Thiran-Vetterli (PTVA) w wersji ograniczonej do sytuacji, w której „obserwatorzy" rejestrują czas przybycia informacji, ale nie identyfikują węzła sieci z którego ta informacja przychodzi (LPTVA, od *Limited* PTVA). Po precyzyjnym sformułowaniu problemu, zaprezentowano rozwiązanie z wykorzystaniem estymacji metodą największej wiarygodności (MLE, od *Maximum Likelihood Estimation*). Następnie przeanalizowano złożoność obliczeniową algorytmu i jego stabilność numeryczną.

W rozdziale trzecim, najobszerniejszym w rozprawie, autor koncentruje się na ważnym aspekcie problemu lokalizacji źródeł jakim jest właściwy (optymalny) dobór (rozmieszczenie) „obserwatorów" (czujników) w sieciach złożonych. Przeanalizowano kilka metod rozmieszczania czujników pod kątem ich użyteczności (skuteczności) w detekcji źródeł informacji. Cztery z tych metod to najciekawsze, zdaniem autora, podejścia znane z literatury przedmiotu, piąta metoda jest podejściem autorskim, opartym na maksymalizacji nowej miary nazwanej „kolektywnym pośrednictwem" (*Collective Betweenness*), zaś ostatnia, szósta metoda, polega na wyborze czysto losowym i służy jako punkt odniesienia. Każdą z tych metod przebadano na różnych realizacjach pięciu typów sieci syntetycznych i na trzech sieciach rzeczywistych. Dla sieci syntetycznych zmieniano dwa parametry: poziom losowości (mierzony wariancją transmisji) oraz gęstość czujników. W przypadku trzech sieci rzeczywistych, o ustalonej strukturze węzłów i połączeń, zmieniano oczywiście tylko drugi z tych parametrów. Bardzo obszerne, szczegółowe i interesujące wyniki przedstawiono głównie w postaci tabel i różnego typu wykresów. Rozdział kończy się syntetyczną dyskusją otrzymanych wyników w podrozdziale 3.4. Z wyników, które sam autor uznał za najciekawsze, warto wymienić następujące dwie obserwacje: (a) istnieje wyraźna różnica w wydajności testowanych metod doboru czujników w zależności od charakteru i typu sieci (sieci rzeczywiste i bezskalowe sieci syntetyczne *versus* sieci o wąskim rozkładzie stopni wierzchołków) oraz (b) autorska metoda wykorzystująca „kolektywne pośrednictwo" bardzo dobrze radzi sobie w przypadku sieci o wysokiej stochastyczności a w przypadku sieci o niskiej stochastyczności dobrze radzi sobie metoda *High Variance Observers*.

Rozdział czwarty podejmuje próbę zaradzenia słabej skalowalności metody LPTVA. Jej złożoność obliczeniowa sprawia, że przestaje być ona efektywna dla sieci o dużych rozmiarach. Zaproponowana przez autora w tym rozdziale metoda GMLA (od *Gradient Maximum Likelihood Algorithm*) ignoruje informacje otrzymane od tych „obserwatorów", którzy otrzymali wiadomość na tyle późno, że można podejrzewać, iż nie są już one niezależne. Innymi słowy, w odróżnieniu od metody LPTVA, która korzysta ze wszystkich informacji, dostarczanych przez wszystkie czujniki, w metodzie GMLA wykorzystujemy tylko informacje najlepszej jakości. Takie podejście pozwala istotnie zredukować złożoność obliczeniową algorytmu. Jak wynika z symulacji numerycznych autora, różnica między GMLA a LPTVA w najgorszym przypadku jest taka, jak między sortowaniem szybkim (*quicksort*) a sortowaniem np. przez wstawianie. Przeprowadzone przez autora testy i symulacje jednoznacznie wskazują, że w przypadku sieci bezskalowych, metoda GMLA, mimo korzystania z mniejszej liczby

danych, daje wyniki lepsze niż metoda LPTVA. Zdaniem autora może to wynikać z faktu, że w sieciach bezskalowych istnieją huby, czyli węzły wysokiego stopnia. Uprawdopodabnia to scenariusz, że wyeliminowane czujniki, czyli te, do których informacja dotarła późno, znajdują się za hubami, a obecność hubów może utrudniać lokalizację źródła. Ponieważ GMLA wykorzystuje czujniki, do których informacja dotarła najwcześniej (tzw. *first observers*), dla sieci bezskalowych uzyskuje się tą metodą wyniki lepsze niż metodą LPTVA, gdyż dla dużych sieci o tej własności rzadko kiedy *first observers* znajdują się za hubami.

W rozdziale piątym analizowany jest problem detekcji źródeł informacji w sieciach wielowarstwowych. Jest to sytuacja bardziej realistyczna, ale też i bardziej skomplikowana. Autor rozprawy prezentuje w tym rozdziale własną metodę odpowiedniego rozszerzenia metody LPTVA na sieci wielowarstwowe. Przeprowadza w tym rozdziale eksperymenty numeryczne pokazujące jak jakość lokalizacji źródeł informacji zależy od takich parametrów jak liczba warstw, gęstość czujników i tzw. wskaźnika infekcji pomiędzy warstwami. Autor zaobserwował istnienie dwóch różnych zakresów badanych parametrów, dla których zachowanie się sieci jest odmienne. W jednym przypadku warstwy interferują destrukcyjne, a w drugim ma miejsce synergia. Jeśli wartość wskaźnika infekcji między warstwami jest niewielka, obserwacje z różnych warstw zakłócają się wzajemnie. W przeciwnym przypadku, ma miejsce synergia, dzięki której dokładność lokalizacji jest większa niż dla sieci jednowarstwowej o tej samej wielkości i gęstości czujników. Autor zaproponował też heurystyczną metodę określania, w którym z tych dwóch zakresów funkcjonuje system.

Rozdział szósty (**Summary**) zawiera podsumowanie osiągniętych wyników, najważniejsze wnioski z nich wynikające i pogłębioną dyskusję. W szczególności pojawiają się tutaj trzy główne, zdaniem autora, osiągnięcia rozprawy: (a) rozwinięcie nowej metody efektywnego rozmieszczania czujników z wykorzystaniem miary *Collective Betweenness*, która przewyższa inne metody w przypadku dużej stochastyczności procesów propagowania się informacji (b) opracowanie GLMA - nowej, szybszej metody lokalizacji źródeł informacji, która istotnie przewyższa jakością metodę LPTVA dla sieci bezskalowych, gdyż jest mniej czuła na istnienie hubów (c) odkrycie efektów interferencji destruktywnej i synergii dla różnych parametrów w sieciach wielowarstwowych i opracowanie heurystycznej metody diagnozowania tych efektów.

Rozprawę zamyka dziewięciostronicowa bibliografia (**Bibliography**), zawierająca właściwie dobrane i w większości starannie zredagowane 102 pozycje literatury przedmiotu.

Przejdę teraz do bardziej szczegółowych uwag i pytań, które nasunęły mi się w trakcie lektury rozprawy.

Autor rozprawy wydaje się utożsamiać pojęcia **identyfikacji**, **detekcji** i **lokalizacji** źródeł informacji propagującej się w sieci. Aczkolwiek dla celów tej rozprawy nie ma to większego znaczenia i nie prowadzi do nieporozumień, to już z pobieżnej analizy zakresów znaczeniowych tych trzech terminów (nawet w standardowej wersji słownikowej) wynika, że można by się w przyszłości pokusić o bardziej subtelne rozróżnienie i zniuansowanie. Nie mam wątpliwości, że doprowadziłoby to do dalszego postępu badań w tym obszarze.

W podrozdziale 1.4 autor pisze: 'The precision of a single test is defined as the ratio between the number of correctly located sources (i.e., true positives, which here equals either zero or one) and the number of seeds found by the method (i.e., true positives plus false positives, which here is at least one).' Nie jest dla mnie zupełnie jasne jak było wybierane źródło s* wypływu informacji. Czy było ono losowo wybierane w grafie? Czy dla każdej realizacji symulacji numerycznej wybierano losowo dokładnie jedno źródło? Czy autor badał może (nie ma tego w rozprawie) jak wyniki symulacji zależą od wyboru/lokalizacji źródła?

Nieco dalej czytamy: 'Distance error is an average distance between the actual and predicted source.' Czy jest oczywiste w jakich jednostkach mierzona jest ta odległość? Czy na pewno spełnia ona wszystkie warunki bycia metryką i jakie ma dodatkowe własności?

Wydaje mi się, że przejście od wzoru (2.10) do wzoru (2.11) jest nieuzasadnione. Wymaga ono bowiem założenia, że macierz kowariancji nie zależy od s. Jeśli przeanalizujemy analogiczne wzory z *Suplemental Materials* do pracy [32], dostępne na stronie https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.109.068702 (w szczególności strony 2-3 i wzór 6), to zobaczymy, że macierz kowariancji jak najbardziej zależy od s.

W podrozdziale 2.3 czytamy: 'The universal SIR model is chosen for transmission modeling with recovery rate $\gamma = 0$. This model is also called the SI since the compartment R does not occur here. The SI is here implemented explicitly as an agent-based model with synchronous dynamics. Each infected node has a chance to pass the information to its neighbor at each subsequent time step. The number of chances per time step is equal to the number of neighbors, and for each neighbor, the probability of success β is the same. Since the number of time steps

needed to pass the information from node i to its neighbor j is equal to the number of independent trials (with the probability β) needed for the first occurrence of success, the delay θij follows a geometric distribution with mean $\mu = 1/\beta$ and variance σ^2 = (1 − β)/β^2.' Oczywiście ten model transmisji informacji (SIR) nie spełnia założenia o normalności czasów opóźnień informacji. Symulacje autora rozprawy można więc potraktować jako próbę badania odporności algorytmów detekcji źródła, które korzystały o założenia o normalności. Pozostają jednak inne pytania. W szczególności, jak konkretnie były obliczane czy generowane czasy t_ij (a w konsekwencji d) potrzebne do obliczenie wartości (*score*) wyrażenia danego wzorem (2.10)? Nie jest to dla mnie całkowicie jasne. Czy model SIR był generowany dyskretnie w ustalonych momentach czasowych, w których dokonywała się kolejna propagacja informacji? Dlaczego autor nie przeprowadził porównań dla generacji informacji, tak jak to zrobili autorzy algorytmu PTVA? To by najlepiej pokazywało na ile proponowany algorytm jest lepszy od dotychczasowego.

Porównania metod LPTVA i GMLA nie zawsze wydają się być symetryczne. Np. str. 15-16, Fig. 2.1. i 2.2 dla modelu BA mamy (N=500, m=3) dla algorytmu LPTVA. Ale na str. 55-56, Fig. 4.5 i 4.6 dla modelu BA mamy (N=1000, m=3) dla algorytmu GMLA. Od strony 65 wszystkie porównania są już robione na takich samych zestawach danych.

W rozprawie zaprezentowano wiele bardzo ciekawych wyników symulacyjnych. Nie ma jednak praktycznie wyników teoretycznych. Nie jest to oczywiście poważny zarzut, w końcu to rozprawa z fizyki a nie algorytmiki czy statystyki matematycznej. Zastanawia mnie jednak, dlaczego autor nie podjął próby teoretycznego uzasadnienia metody GMLA, analogicznie do *Proposition* 2 ze wspomnianej już pracy [32].

Z sugestii bardziej ogólnych, w przyszłości warto się pokusić o uogólnienia na przypadek modeli, gdzie emisja ma wiele źródeł. W końcu większość cyberataków i prawie wszystkie *fake news* tak właśnie się propagują.

Praktycznie wszystkie znane metody lokalizacji źródeł informacji w sieci, w tym GMLA, wymagają znajomości topologii odpowiedniego grafu. Jednak w praktycznych zastosowaniach rzadko kiedy mamy pełną wiedzę o topologii sieci, a nawet o tym, czy opisujący ją graf jest na pewno spójny. Opracowanie metod, które efektywnie rozwiązują problem lokalizacji źródeł bez pełnej wiedzy o topologii jest trudnym i ambitnym zadaniem. Natomiast w zasięgu możliwości autora rozprawy powinna być jakościowa analiza

problemu wpływu ewentualnego braku spójności grafu na strategię rozmieszczania czujników. Innymi słowy, jak ustawiać czujniki nie będąc pewnym, czy w (dużym) grafie nie ma niespójnych komponentów?

Powyższe drobne uwagi krytyczne nie obniżają wartości naukowej pracy. Rozprawa jest dobrze napisana i starannie zredagowana a uzyskane przez autora wyniki zostały zaprezentowane właściwie i czytelnie. Zauważyłem w pracy tylko kilka literówek oraz drobne niedociągnięcia redakcyjne w bibliografii (np. błąd w nazwisku Paula Erdősa w pozycji [16] na stronie 86).

Tematykę rozprawy, która ma charakter interdyscyplinarny, można bez wątpienia zaliczyć do dyscypliny naukowej *nauki fizyczne* w dziedzinie nauk ścisłych i przyrodniczych a przedstawione rozwiązania analizowanych w rozprawie problemów badawczych pozwalają uznać, że mgr inż. Robert Paluch jest właściwie przygotowany do pracy naukowej.

Podsumowując, uważam, że rozprawa doktorska magistra inżyniera Roberta Palucha stanowi bardzo interesujący wkład w rozwój problematyki propagacji informacji w dużych sieciach złożonych. Tematyka podjęta w rozprawie jest ważna i aktualna. Biorąc pod uwagę dynamiczny rozwój sieci złożonych oraz ich powszechność, nie mam wątpliwości, że problematyka ta będzie w najbliższych latach intensywnie rozwijana, dając autorowi szansę na dalsze doskonalenie zaproponowanych rozwiązań. Uważam także, że osiągnięte wyniki pozwalają na wystąpienie z wnioskiem o wyróżnienie rozprawy. Bardzo pozytywnie oceniając poziom naukowy recenzowanej rozprawy doktorskiej i stwierdzając spełnienie wymogów ustawowych, wnoszę o dopuszczenie jej autora do kolejnych etapów przewodu doktorskiego.

Arkadiusz Orłowski