Warszawa, 03.09.2021 r.

prof. dr hab. Ryszard Kutner
Wydział Fizyki Uniwersytet Warszawski
ul. Pasteura 5, 02-093 Warszawa
email:ryszard.kutner@fuw.edu.pl

Przewodniczący Rady Naukowej Dyscypliny Nauki Fizyczne

prof. dr hab. inż. Tomasz Woliński

Opinia o rozprawie doktorskiej mgr Roberta Palucha pt.: "Reverse Engineering of Information Processing in Complex Networks Using Statistical Inference"

Na wstępie pragnę podkreślić, że rozprawa mgr Roberta Palucha o charakterze numerycznym (napisana w języku angielskim) podoba mi się: jest rzetelna i ciekawa, posiada potencjał inspirujący do dalszych badań a także do ważnych zastosowań. Ta rzetelność jest oparta na dużej liczbie benchmarków/testów, co może sprawiać wrażenia rozprawy bardzo technicznej. Rozprawa, jest oparta na alternatywnych spojrzeniach na zagadnienie, co uważam za nadzwyczaj ważne.

Od razu też powiem, że pomimo (w ogromnej większości) starannie wprowadzonych definicji, brakuje słownika używanych przez autora terminów angielskich, który mógłby być zamieszczony na samym wstępie aby ujednolicić komunikację w języku polskim. Dlatego, wszędzie tam gdzie jest to konieczne jestem zmuszony używać nazw angielskich (w wersji podanej w rozprawie).

Należy podkreślić, że praca należy do obszaru zwanego nauką o sieciach (ang. network science) a tam do tematyki rozprzestrzeniania się sygnałów (ang. signal spreading) lub, patrząc od innej strony, rozprzestrzeniania się epidemii (ang. spread of epidemic). Tematyka (ogólnie mówiąc) monitorowania/śledzenia niekontrolowanego rozprzestrzeniania się (delikatnie mówiąc) złośliwych informacji i aplikacji w internecie (w tym różnego rodzaju fake news-ów a także wirusów i robaków – będę je wszystkie w dalszym ciągu określał mianem sygnałów toksycznych) jest wyzwaniem współczesnego, połączonego siecią internetową świata. Jednym z zasadniczych powodów zainteresowania tą tematyką jest destrukcyjna siła oddziaływania sygnałów toksycznych na użytkowników internetu oraz ich

powszechność, prowadzące do poważnych skutków społecznych i ekonomicznych (np. wpływanie na wyniki wyborów a nawet na szeroko rozumiane bezpieczeństwo państw).

Szybkość rozprzestrzeniania się tego typu toksyn, ich anonimowość oraz zdolność do maskowania się stanowi poważne wyzwanie dla ich monitorowania oraz detekcji ich źródeł. Jak się wydaje, wstępnym krokiem we właściwym kierunku jest np. wprowadzenie przez facebook i instagram moderatorów usuwających skrajnie niebezpieczne (ekstremistyczne) toksyny. Dodam, że firmy te stworzyły katalogi cech toksyn częściowo automatyzując proces ich detekcji. Jednakże, jest to wciąż usuwanie skutków a nie leczenie objawów choroby a ponadto ostateczna decyzja o usunięciu danej toksyny jest wciąż podejmowana indywidualnie przez moderatora – zatem jest bardzo czasochłonna/nieefektywna.

Autor pracy podjął się bardziej ambitnego zadania a mianowicie, systemowego dotarcia do statycznych źródeł tego typu treści. Chodzi, po pierwsze o powstrzymanie tych treści u samego źródła zanim ulegną rozprzestrzenieniu i po drugie (co jest nadzwyczaj ważne i już realistyczne na dzień dzisiejszy) stworzenie efektywnego oprogramowania prowadzącego detekcję źródeł (czyli śledzącego/tropiącego źródła). W tym miejscu należy podkreślić, że właśnie efektywność stanowi tutaj wyzwanie, któremu sprostała niniejsza rozprawa. Można ogólnie powiedzieć, że zajmuje się ona tzw. zagadnieniem odwrotnym a mianowicie, namierzeniem lokalizacji źródła, dzięki sygnałom pochodzącym z tego źródła a idącym poprzez wybrane węzły sieci (czujniki). Zagadnieniem prostym nazywamy sytuację, gdy znana jest lokalizacja źródła w sieci i analizowane są wysyłane przez nie sygnały (np. giełda jako źródło produkuje sygnały w postaci notowań, które analizujemy).

Rozprawa prezentuje badania zawarte w trzech publikacjach zamieszczonych w międzynarodowych czasopismach (Scientific Reports, Future Generation Computer Systems oraz Physica A) o wysokim IF ponadto, badania zawarte w jednej pracy stanowiącej rozdział w książce `Simplicity of Complexity in Economic and Social Systems', jaka ukazała się nakładem Wydawnictwa Springera oraz w jednej przygotowywanej aktualnie do druku.

Dokładnie rzecz biorąc, przedmiotem przedstawionej mi do oceny rozprawy jest ulepszanie istniejących metod identyfikacji/detekcji źródeł propagacji sygnałów w dużych sieciach złożonych oraz przedstawienie własnych metodologii, metod/podejść i algorytmów. Ponadto, celem pracy jest unikalne podejście rozszerzające tematykę na sieci wielowarstwowe – najbardziej realistyczne. Właśnie ten element pracy wywarł na mnie największe wrażenie – uważam go za oryginalny i nadzwyczaj cenny.

Aby zrealizować cel rozprawy, autor stosuje z jednej strony rozpowszechnione metody lokalizacji źródła oparte na czujnikach (obserwatorach, detektorach, sensorach), z drugiej ich cenne (własne/autorskie) rozszerzenie np. na sieci wielowarstwowe. W tym kontekście na szczególną uwagę zasługują **rozdziały 3 – 5**. Warto też zwrócić uwagę na model/algorytm LPTVA, który autor omówił w **rozdz. 2** a znacząco rozszerzył w **rozdz.** 5. W **rozdz. 4** autor zaproponował podejście, które nazwał `Gradientowym Algorytmem Maksymalnej Wiarygodnosci (ang. akronim GMLA)'. Ponadto, starannie opracowany **rozdz. 1** zawierający konieczne definicje oraz informacje (a w tym niezbędne odniesienia literaturowe), jest przydatny dla zrozumienia ducha rozprawy a zwłaszcza jej umocowania w głównym nurcie tematycznym. Rozprawę zamykają **rozdz. 6**, stanowiący zgrabne podsumowanie oraz użyteczna bibliografia.

Poniżej omawiam systematycznie, wspomnianą powyżej zawartość rozprawy zaczynając od najważniejszych rozdziałów. Zatem, w **rozdz. 3** autor wprowadził nową wielce użyteczną miarę analizy sieci, którą nazwał Kolektywnym Pośrednictwem (ang. Collective Betweeness) i porównał ją z pięcioma innymi ważnymi tego typu miarami, a mianowicie z: Betweenness Centrality, High Coverage Rate, K-Median, High Variance Observer oraz Randomly Placed Sensors (patrz porównania zamieszczone na wykresie 3.1 oraz w Tabeli 3.1 dla czterech spośród nich). Tutaj chciałbym nadmienić, że w definicji Betweeness Centrality (3.1) nie został zdefiniowany zbiór S – czyżby to był zbiór źródeł? Chyba nie, bo ich nie znamy. A może to jest po prostu zbiór węzłów, których pośrednictwo badamy? Jak widać, potrzebne jest przynajmniej jedno zdanie wyjaśnienia.

Wprowadzona przez autora miara posłużyła do skonstruowania efektywnego algorytmu rozmieszczania optymalnej konfiguracji czujników w różnych sieciach złożonych takich jak

syntetyczne: Erdös-Rényi network, Random Regular Graph, Degree Sequence Algorithm Graph with Poisson distribution, Barabási-Albert evolving network, Configuration Network with power-law degree distribution oraz w trzech rzeczywistych sieciach złożonych: Infection Network (human face-to-face contacts), and two local networks of internet communications between academics at University of California (USA) and University of Rovira & Virgili (Spain). Miara ta została przetestowana w rozprawie na tych charakterystycznych przykładach reprezentujących różne rodzaje sieci złożonych, w szerokim zakresie gęstości czujników oraz losowości używanych sygnałów. Na tej drodze autor wykazał (ogólnie mówiąc) wyższość swojego podejścia optymalnego rozmieszczenia czujników w stosunku do tych już istniejących, przede wszystkim w sytuacji silnie stochastycznego charakteru rozprzestrzeniających się sygnałów, czyli dla dużej wartości parametru transmission variance (stosunek szumu do sygnału) $\xi = \sigma/\mu$, gdzie μ jest średnim czasem transmisji sygnału w sieci, natomiast σ jest odchyleniem standardowym czasów transmisji. Jest to jeden z trzech głównych wyników pracy – oceniam go bardzo pozytywnie, gdyż wydatnie rozszerza możliwości monitoringu.

Chociaż różnice w wynikach testów przeprowadzonych z użyciem powyżej podanych miar nie są duże, co widać na wykresach 3.3-3.10 porównujących jakość lokalizacji źródeł, to jednak spodziewam się, że podejście autora powinno wkrótce być w powszechnym użytkowaniu. Szkoda tylko, że nigdzie w rozprawie nie znalazłem precyzyjnej definicji pojęcia `average precision' przedstawionego na tych wykresach (nawiasem mówiąc, wolałbym aby osie rzędnych sześciu pierwszych wykresów były opisane przez ten właśnie termin a nie przez `precision' jak jest teraz; dotyczy to także innych wykresów rozprawy zawierających taką oś).

Trzeba też w tym kontekście podkreślić bogactwo informacji jakie niosą ze sobą podsumowujące czytelne diagramy zamieszczone na rysunkach 3.11-3.15 oraz 3.16-3.18, wsparte podsumowującymi Tabelami (odpowiednio) 3.3-3.7 oraz 3.9-3.11. Te cenne diagramy i towarzyszące in Tabele zostały przez autora systematycznie przedyskutowane.

Dodam, że wspomniane powyżej sieci syntetyczne autor podzielił na dwie zasadnicze grupy: 1) sieci bezskalowe, dobrze opisujące sytuacje realne oraz 2) sieci quasi-

deterministyczne (o wąskim rozkładzie stopni wierzchołków). Badania autora wpisały się w nurt analiz tych sieci ze względu na gęstość czujników i stopień losowości sygnałów. W rozprawie autor wskazał na wyraźne, charakterystyczne różnice zachodzące pomiędzy obiema grupami sieci. Dla pierwszej grupy jest możliwe wyróżnienie liderów wydajności optymalnego rozmieszczenia czujników w przeciwieństwie do drugiej grupy, gdzie różnice pomiędzy wynikami dostarczanymi przez różne metody są szczątkowe.

Co więcej, autorowi udało się w wyniku prowadzonych przez siebie drobiazgowych testów, wyróżnić metodę `High Variance Observers', która jest zdecydowany ı liderem w analizie propagacji sygnałów o niskiej stochastyczności. W tym kontekście jednym z kluczowych osiągnięć rozprawy jest podejście oparte na wprowadzonym przez autora Kolektywnym Pośrednictwie. Podejście to radzi sobie w sposób wyróżniający w sytuacji gdy rozprzestrzenianie się sygnałów jest wysoce stochastyczne/nieprzewidywalne, jak to ma najczęściej miejsce w realnych sieciach.

Podsumowując, w tym rozdziale autor wykazał że właściwy dobór czujników może wyraźnie zwiększyć jakość identyfikacji źródła. Zatem, ma istotne znaczenie z których miejsc w sieci algorytm czerpie sygnały/informacje. Kolejny rozdział pozwala spojrzeć na to zagadnienie z alternatywnego/dopełniającego punktu widzenia.

Jeżeli chodzi o zaproponowaną przez autora w **rozdz. 4** technikę GMLA, to jest ona, co należy mocno podkreślić, szybką technikę lokalizacji źródła sygnałów. Ponadto, technika GMLA zapewnia wyższą jakość wykrywania tych źródeł w sieciach bezskalowych w porównaniu z tradycyjną metodą LPTVA, ponieważ jest mniej podatna na wpływ sieciowych hub-ów (patrz rozdz. 4.5 a tam wykresy na rysunku 4.18).

Główną wadą tradycyjnej metody LPTVA jest złożoność algorytmiczna jej operacji macierzowych typu $O(K^a)$, $3 \leq a \leq 4$, przypadająca na pojedynczy w zeł sieci, gdzie K oznacza liczbę czujników w sieci. GMLA rozwiązuje ten poważny problem, poprzez 1) ograniczenie liczby detektorów używanych do znalezienia prawdopodobieństwa tego, że dany węzeł sieci jest źródłem toksycznego sygnału oraz poprzez 2) ograniczenie liczby sprawdzanych węzłów (patrz rozdz. 4.2 a tam zwłaszcza rysunek 4.2). Zatem, GMLA działa

jak filtr odrzucając sygnały niskiej jakości pochodzące od wielu odległych czujników/obserwatorów. Pozostają więc tylko te czujniki, które dostarczają sygnały o najmniejszym opóźnieniu. Przy okazji, nie zauważyłem w rozprawie precyzyjnej definicji score of node – przydałaby się.

Warto zwrócić uwagę na ważne wykresy 4.5 i 4.6, które m.in. pokazują, że GMLA nie daje 100-procentowej pewności detekcji źródła sygnału. W rozdz. 4.2 autor bada starannie to i temu podobne zagadnienia pokazując następnie (w rozdz. 4.3) jak podnieść zdolność do poprawnej detekcji, czyli efektywność GMLA. Dobrze to widać na kluczowych wykresach 4.8-4.11. Istotne tutaj są także wykresy 4.12 i 4.13 pokazujące średnią precyzję detekcji w zależności od rozmiaru sieci oraz wymowne wykresy 4.14-4.17 porównujące średnią precyzję GMLA i LPTVA oraz niepewności detekcji tych podejść w zależności od gęstości czujników. Przy okazji, przydałaby się uwaga jasno pokazująca czym różni się podejście LPTVA od PTVA.

Osiągnięcie uzyskane przez GMLA było możliwe dzięki oryginalnej/autorskiej procedurze selekcji wykorzystującej: a) odpowiednią konfigurację startową (chodzi tutaj przede wszystkim o wybór first observers, czyli czujników znajdujących się blisko źródła (tzn. wysyłających sygnały o możliwie krótkich opóźnieniach) , patrz rozdz. 4.1) oraz b) procedurę spadku gradientu prawdopodobieństwa. Doprowadziło to do złożoności algorytmicznej co najwyżej typu O(NlogN). Jest to jeden z najszybszych algorytmów detekcji źródeł toksycznych sygnałów w sieciach generycznych z niepełnym zbiorem czujników.

Jeżeli chodzi o **rozdz. 5,** to uważam go za nadzwyczaj ważny i szczególnie interesujący. W rozdziale tym autor jasno pokazał zależność pomiędzy jakością wykrywania źródła a liczbą warstw sieci złożonej, gęstością obserwatorów oraz wskaźnikiem zarażania/infekcji.

W rozdziale tym autor przeanalizował problem identyfikacji źródeł sygnałów ulokowanych w wielowarstwowych sieciach złożonych, które są przecież wyraźnie obecne w świecie realnym. Należy mocno podkreślić, że ten istotny problem został po raz pierwszy postawiony i zbadany w przedstawionej mi do oceny rozprawie. Na schematycznym

rysunku 5.1 problem ten został zilustrowany dla dwóch kolejnych chwil czasu, umożliwiając wprowadzenie modelu SI z dobrze określoną macierzą szybkości zarażania węzłów.

Autor pokazał, w ramach modelu SI, w jaki sposób jakość wykrywania źródła (`precision') zależy od liczby warstw, gęstości obserwatorów oraz szybkości/wskaźnika infekcji (patrz wymowne rysunki 5.3-5.8). Mianowicie, wykrył on dwa wyraźnie różne zakresy parametrów w zależności od tego czy szybkość infekowania międzywarstwowego jest duża czy mała. Dla dużej szybkości infekowania międzywarstwowego zaobserwował on bardzo interesujące zjawisko interferencji konstruktywnej/synergii obserwacji pochodzących z różnych warstw sieci. Podnosi to dokładność detekcji powyżej progu wyznaczonego przez odpowiadającą jej sieć jednowarstwową o tej samej gęstości czujników i tej samej wielkości sieci. Widać to świetnie na rysunkach 5.5 i 5.6.

Z drugiej strony, jeżeli ta szybkość jest mała to ma miejsce wzajemne zakłócanie się sygnałów płynących od czujników znajdujących się w różnych warstwach. Zmniejsza to dokładność.detekcji poniżej wspomnianego powyżej progu.

W uzupełnieniu, autor rozwinął w tym rozdziale metodę (którą można nazwać heurystycznym filtrem) poprawiającą dokładność detekcji źródeł poprzez odrzucenie zakłócających obserwacji.

Pragnę podkreślić, że zawarte w recenzji uwagi krytyczne w najmniejszym stopniu nie podważają mojej nadzwyczaj pozytywnej opinii o rozprawie.

W podsumowaniu mogę z pełnym przekonaniem powiedzieć, że ta bardzo dobra rozprawa doktorska spełnia wszelkie ustawowe wymagania stawiane tego typu pracom. Dlatego wnoszę o jej przyjęcie do kolejnego etapu procedury doktorskiej. Ponadto wnioskuję o uznanie rozprawy za wyróżniającą ze względu na ogromny walor użytkowy a także inne bardzo ważne elementy - zawarłem je w przedłożonej recenzji.

R. Kutner.