Prof. dr hab. inż. Marian Chudy
Wydział Cybernetyki
Wojskowej Akademii Technicznej
Ul. S. Kaliskiego 2
00-908 Warszawa

Warszawa 12-01-2022

RECENZJA

Osiągnięcia naukowego dr. inż. Mariusza Kamoli p.t: „Metody analizy i odkrywania struktury sieci społecznych i technologicznych" oraz Jego aktywności naukowej przedłożonych do oceny w postępowaniu habilitacyjnym prowadzonym na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej.

1. Uwagi ogólne.

Przedstawione do oceny **osiągnięcie naukowe** składa się z 10 publikacji, w tym
- jednej monografii (H1), której współautorem jest Piotr Arabas, PWN,
- pięciu artykułów (H2-H5, H8), w tym cztery (H2, H3, H5, H8) w czasopismach posiadających Impact Factor, przy czym tylko jeden (H8) nie jest indeksowany w bazie Web of Science,
- trzech artykułów konferencyjnych (H6, H9, H10), których jeden jest indeksowany w bazie Web of Science,
- jednego rozdziału w monografii (H7).

Współautorem w publikacjach H4 i H5 jest Piotr Arabas, współautorem w H8 jest Barbara Laskowska, współautorami H9 są Ewa Niewiadomska-Szynkiewicz i Bartłomiej Piech.
Wspólnym wątkiem rozważań w przedstawionych publikacjach są sieci złożone, a wyniki uzyskane przez Autora dotyczą:
1. analizy sieci złożonych i badania ich własności
2. rekonstrukcji sieci złożonych w oparcie o dostępne dane.

Prace nie zawierają formalnej definicji sieci złożonej.
Podział na wątki badawcze w ramach wyżej wymienionych nurtów badawczych (zaproponowany przez Habilitanta) jest moim zdaniem adekwatny do zawartości merytorycznej publikacji.
Przedstawione osiągnięcie naukowe wykorzystuje zarówno istniejące metody badawcze jak i przedkłada propozycję nowych- autorskich metod.

2. Zawartość i ocena dorobku w zakresie pierwszego nurtu badawczego.

(Analizy sieci złożonych i badania ich własności)

Tego wątku badawczego dotyczą publikacje H1, H2, H3 oraz H4, H5 i H7.
W H1 Autor prezentuje metodę oceny jakości grupowania aglomeracyjnego w sieci dwudzielnej. Rozważane są cztery rodzaje projekcji oraz cztery rodzaje miar odległości między węzłami. Ocena jakości grupowania jest dwukryterialna. Każdej kombinacji projekcji i odległości odpowiada punkt w przestrzeni dwuwymiarowej, którego składowe to wartość korelacji współwystąpienia dla projekcji na węzły- aktorzy oraz wartość korelacji współwystąpienia dla projekcji na węzły- filmy. Taka prezentacja wyników umożliwia poszukiwanie zbioru rozwiązań Pareto-optymalnych.

W pracy H2 zawarta jest procedura porównywania zgodności wyników grupowania w sieci dwudzielnej. Autor nawiązuje do tzw. prawa Conwaya, które wskazuje na podobieństwo struktury modułów oprogramowania w systemie informatycznych do struktury organizacyjnej wytwórcy tego oprogramowania.
Próbę weryfikacji tej zależności przeprowadzono na zbiorze dużych i zakończonych sukcesem projektów oraz ich wykonawców. Wykonano dwa zadania badawcze.
Pierwszym i głównym rezultatem badań jest opracowanie algorytmu wyznaczania optymalnej liczby grup w procesie grupowania aglomeracyjnego zarówno dla modułów oprogramowania jak i ich wykonawców (programistów).
Autor wprowadza dwie funkcje zależne od liczby kroków grupowania:
- $J_{TP}(G,k)$ oznaczającą sumę odległości par węzłów sieci G, które znajdują się w tej samej grupie w k-tym kroku grupowania,
- $J_{TN}(G,k)$ oznaczającą sumę dopełnień odległości (od średnicy sieci G) par węzłów, które znalazły się w różnych grupach w k-tym kroku grupowania.

Suma tych funkcji oznacza łączny koszt grupowania. Krok, w którym ta suma osiąga maksimum wyznacza wyjściowy podział sieci.
Wyniki działania algorytmu wskazują, że uzyskane od tego momentu wartości optymalne liczby grup są dla modułów i ich wykonawców różne.
Należy w tym miejscu zauważyć, że te wyniki są rezultatem przyjętych przez Autora postaci funkcji.
Jakie wyniki uzyskano by gdyby te funkcje miały inną interpretację?
Drugim rezultatem prac badawczych Autora (w kontekście zależności Conwaya) jest określenie stopnia podobieństwa strukturalnego dwóch sieci w kontekście grupowania aglomeracyjnego. Dla zadanej liczby N grup modułów i N grup programistów Autor wykonuje N! sprawdzeń wartości sumy aktywności programistów na rzecz modułów w ramach grup utożsamianych ze sobą i analogicznej sumy dla grup nieutożsamionych.
Przedstawione zobrazowanie (kolumny macierzy to grupy modułów, wiersze to grupy programistów) pokazuje części wspólne i rozłączne.
Przedstawiona procedura ma charakter przybliżony i nie wyznacza jednoznacznie tego podobieństwa.
W ramach wzbogacania wiedzy o sieci Autor w H3 proponuje metodykę oceny wrażliwości wybranych indeksów istotności węzłów na błędy w szacowaniu wag połączeń między nimi.
Założono, że wagi mogą przyjmować wartości całkowite od 1 do 10 i odzwierciedlają siłę powiązań między węzłami. Autor rozpatruje trzy indeksy istotności określające podatność węzła- usługi v_k na awarie innych węzłów-usług:

- Page Rank, który agreguje istotność wszystkich usług wpływających na węzeł v_k,
- Reach Centrality, oznaczający odsetek wszystkich usług, których awaria może dotyczyć v_k
- Maximum Input, uwzględniający tylko usługi sąsiadów o największym wpływie.

Badanie wrażliwości Autor dokonuje poprzez zaburzenie wartości wag oddziaływaniem zmiennej losowej o dyskretnym rozkładzie jednostajnym.
Wyniki pomiarów traktowane jako trzy niezależne składowe tworzą podzbiór wektorów w przestrzeni E^3, co dalej pozwala wyznaczać rozwiązania Pareto-optymalne.

Ważną charakterystyką sieci jest jej odporność na awarie łączy i węzłów sieci. Do analizy w H4 przyjęto model sieci bazujący na grafie, w którym węzłami są tzw. systemy autonomiczne (SA). Analiza dotyczy polskiej części sieci Internet.
Istnienie trasy między parą $\{v_i, v_j\}$ AS-ów Autor definiuje jako istnienie co najmniej jednego AS, który jest dostępny zarówno z v_i jak i v_j.

Zaproponowano wskaźnik $u_E(v_i)$ podatności strukturalnej węzła v_i na awarię zależny od liczby utraconych połączeń do pozostałych AS ze względu na awarie pojedynczych **łączy** oraz wskaźnik podatności $u_V(v_i)$ zależny od awarii poszczególnych **węzłów**.
Dla łączy i węzłów zdefiniowano wskaźniki **istotności**;

- $S(e_j)$ dla łącza oznaczający liczbę par AS, które utracą łączącą je trasę w wyniku awarii tego łącza,
- $S(v_i)$ dla węzła oznaczający liczbę par AS, które utracą łączącą je trasę w wyniku awarii tego węzła.

Uwzględniając położenie geograficzne AS ustalono, że atak na sieć w promieniu większym niż 200m powoduje gwałtowny wzrost braku połączeń między AS.
Uzyskane wyniki nie są skonfrontowane z innymi metodami wyznaczania podatności sieci na awarie łączy i węzłów.
Innym rodzajem sieci złożonej rozpatrywanym przez Autora w H5 jest sieć bazująca na kolokacji słów w języku polskim. W takiej sieci **węzłami są słowa**, natomiast **łącza** są wyznaczane jako **kolokacje słów.**
Wytypowano czternaście cech kolokacji i zbadano, które z tych cech mogą mieć wpływ na wartość wagi kolokacji. Uzyskane wyniki mogą być wykorzystane do proponowania kolokacji, które będą charakterystyczne na tle natłoku informacji nieistotnych.
Kolejnym zagadnieniem rozważanym przez Autora jest ocena zmian struktury i niektórych właściwości sieci złożonej (H6). Punktem wyjścia były wnioski zawarte w pracy [18]. Według autorów [18] obecność w sieci triad przechodnich świadczy o stabilności grupy i zachowaniu równowagi informacyjnej, natomiast duża liczba triad nieprzechodnich jest zaprzeczeniem tych cech. Idąc za tym wskazaniem Habilitant analizuje dynamikę przechodzenia nieprzechodniej triady typu 210 (brak jedynego odwzajemniania relacji między dwoma węzłami) do przechodniej typu 300. Autor pokazał, że ten proces domykania jest podobny do reakcji chemicznej drugiego rzędu. Ta analogia pozwala na dogłębne badanie procesu domykania, co znalazło zastosowanie w odniesieniu do terminarza ferii zimowych w polskiej wersji H6.
W ramach badań dynamiki sieci Autor (H7) przedstawia swój model dyfuzji informacji wśród użytkowników serwisu „social news". Według tego modelu łączne rozprzestrzeniane się wiadomości jest podobne do odpowiedzi członu inercyjnego na skok jednostkowy, poprzedzone w niektórych przypadkach fazą liniową. Badania przeprowadzono w oparciu o polski serwis informacyjny.

3. Zawartość i ocena dorobku w zakresie drugigo nurtu badawczego.

(Rekonstrukcja sieci złożonych w oparcie o dostępne dane)

Wyróżnione jako 2.1. i 2.2. wątki badawcze sąsiadują w wielu częściach przedstawionych do oceny publikacji. Takie sąsiedztwo obserwujemy w H2, gdzie Autor proponuje metodę

rekonstrukcji sieci modułów oprogramowania i powiązań merytorycznych programistów. Danymi wejściowymi były repozytoria projektów o otwartym kodzie źródłowym.
Sieć powiązań miedzy programistami powstała w oparciu o graf dwudzielny programista-aktywność. Odtworzone sieci modułów i programistów wskazują na istnienie wyjątkowych programistów, którzy uczestniczą w tworzeniu większości plików oraz istnienie szczególnych modułów, które są tworzone przez większość programistów. Wyniki te są pretekstem do dyskusji o zakresie obowiązywania tzw. prawa Conwaya.
Badanie podobieństwa dokumentów może odbywać się poprzez rekonstrukcję dwudzielnej sieci podobieństw, w której węzły typu A reprezentują dokumenty a węzły typu B są reprezentowane przez słowa bądź kolokacje słów. Ponieważ kolokacji słów może być wiele istnieje konieczność ich ograniczenia do zbioru najistotniejszych. W pracy H1 przedstawiono algorytm wyboru N znaczących kolokacji, które mogą być punktem wyjścia do wyboru przez ekspertów kolokacji najistotniejszych.
Osobnym zagadnieniem jest rekonstrukcja sieci podobieństwa utworów muzycznych. Obiektem podstawowym jest motyw muzyczny, który składa się z wyróżnionych n-gramów czyli n-znakowych fragmentów tekstu muzycznego zwanych realizacjami motywu. Podobieństwo motywów obliczane jest z wykorzystaniem indeksu Jaccarda realizacji motywów. Podobieństwo dwóch utworów obliczane jest jako maksimum podobieństwa par motywów pochodzących z porównywanych utworów. Liczba n jest parametrem algorytmu. Tak zdefiniowane wielkości pozwalają zbudować sieć utworów muzycznych.
Nie zbadano wrażliwości sieci podobieństwa na zmianę parametru n.
Rekonstrukcja sieci dokonywana jest zwykle przy różnych ograniczeniach zasobowych.
Przykładem takiego ograniczenia np. w sieci powiązań użytkowników serwisu Instagram jest liczba danych o profilach użytkowników udostępnianych w jednostce czasu. Na bazie tego serwisu w H6 przedstawiono algorytm rekonstrukcji podgrafu o dużej gęstości.
Do aktualnie pozyskanych węzłów dołączanych jest N najbardziej usieciowionych węzłów, co wymaga wspomagającego ten etap iteracji posortowania nowoodkrytych węzłów.
Jest to algorytm heurystyczny, którego jakość jest trudna do zweryfikowania.
W pracy H9 Autor rekonstruuje sieć powiązań miedzy użytkownikami stacjonarnej sieci telefonicznej w oparciu o zaszyfrowane dane pochodzące od operatora. Modelem użytkowym jest sieć dwudzielna z ograniczeniem do liczby r najczęściej wybieranych numerów docelowych przez abonamenta. Liczba r jest zatem parametrem algorytmu. Skutki zmian wielkości r zostały częściowo przeanalizowane tylko w odniesieniu do aspektu bezskalowości.
Zagadnienie transformacji sieci z danymi rzeczywistymi do sieci z anonimowymi danymi ale z zachowaniem podstawowych cech umożliwiających dokonanie pożądanych analiz jest rozważane w pracy H10. Zagadnienie to dotyczy miedzy innymi takiego przekształcenia sieci ulic aby zachowując anonimowość kierowców pojazdów można było analizować ich zachowanie.
Ważnym aspektem transformacji sieci oryginalnej do wynikowej jest zapewnienie niewielomianowej deanonimizacji węzłów-skrzyżowań, co utrudniłoby dotarcie do danych rzeczywistych.
Przestawiony algorytm jest konstrukcją heurystyczną z utrudnioną możliwością oceny jakości.

4. Liczbowe charakterystyki dorobku naukowego.

Sumaryczna wartość IF publikacji wchodzących w skład **osiągnięcia naukowego** wynosi 11,518, łączny IF wszystkich publikacji wynosi 20,579, liczba punktów MNiSW za te publikacje jest równa 904, indeks cytowań wynosi 3 (H=3) według Web of Science, H=5 według Scopus, H=6 wwedług Google Scholar.

Szczegółowe dane liczbowe dorobku **po uzyskaniu stopnia doktora** przedstawiają poniższe tabele.

	Liczba prac
Prace z listy JCR uwzględnione w wykazie MNiSW, autor	4
Prace z listy JCR uwzględnione w wykazie MNiSW, współautor	5
Prace spoza listy JCR uwzględnione w wykazie MNiSW międzynarodowe	7
Prace spoza listy JCR uwzględnione w wykazie MNiSW krajowe	9
Materiały konferencyjne, międzynarodowe	10
Materiały konferencyjne, krajowe	2
Rozdziały w monografiach naukowych	6
Monografie, współautorstwo	2
Razem	45

Informacja o liczbie cytowań.

	Liczba publikacji	Liczba cytowań	Liczba cytowań bez autocytowań
Web of Science	29	65	57
Scopus	28	110	93
Google Scholar	43	160	135

Powyższe charakterystyki oceniam pozytywnie.

5. Inne osiągnięcia naukowe.

Zainteresowania naukowe Habilitanta są dość szerokie i mają swoje odzwierciedlenie w licznych publikacjach.

Do głównych wątków dodatkowych realizowanych równolegle z zawartymi w zgłoszonym osiągnięciu naukowym można zaliczyć:

- badania w zakresie inżynierii ruchu w sieci pakietowej (12 publikacji),
- analizy danych i sterowanie obiektami przemysłowymi (3 publikacje),
- badania i rozwój potencjału rynku domenowego (4 publikacje),
- analizy danych heterogenicznych (1 praca),
- opracowanie narzędzi i algorytmów przetwarzania równoległego i rozproszonego (3publikacje).

6. Osiągnięcia związane z działalnością naukową.

6.1. Osiągnięcia dydaktyczne.

Habilitant prowadzi zajęcia dydaktyczne na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej na wszystkich etapach kształcenia
Na studiach I stopnia uczestniczy w realizacji czterech przedmiotów, na studiach II stopnia w sześciu, a na studiach podyplomowych w jednym. Habilitant był promotorem 18 prac magisterskich oraz 23 inżynierskich.

6.2. Współpraca naukowa.

Habilitant współpracował z Uniwersytetem w Genui w ramach projektu Econet.
Głównym podmiotem współpracy jest NASK.
Habilitant uczestniczył łącznie w realizacji 15 projektów badawczych o tematyce zbliżonej do tematyki osiągnięcia naukowego.

6.3.Staże naukowe.

Habilitant odbył staż naukowy w Universitat-Gesamthochschule Siegen w 1999r.
oraz dwukrotnie staż w firmie Knoledge-Support System Ltd. W Manchesterze w latach 2000-2001.

6.4.Działalność popularyzująca naukę.

Do tej kategorii działalności można zaliczyć:
- prezentację techniki głębokiego uczenia maszynowego w ramach konferencji „Perspektywy dla rozwoju Internetu Rzeczy" w 2020 r.
- prezentacja pracowni sterowania siecią na konferencji Secure w 20214 .
- prezentację wyszukiwarki wolnych domen na konferencja NASK w 2013 i 2014 r.
- artykuł „Tomografia sieciowa" w Biuletynie NASK.

7. Ocena końcowa.

Przedstawione osiągnięcie naukowe jest zbiorem wielu różnych autorskich metod i algorytmów dotyczących analizy i rekonstrukcji siei złożonych. Sieci te odnoszą się do wielu praktycznych problemów. Zakres tematyczny osiągnięcia naukowego oraz związanych z nim prac jest bardzo obszerny, co jest zarówno zaletą jak i wadą dorobku naukowego.
Zdecydowana większość przedstawionych metod i algorytmów ma charter heurystyczny bez analitycznych ocen jakości. Ubogo prezentowane są również zestawienia porównawcze z innymi wynikami opisanymi w literaturze.
Należy jednak podkreślić, że rozpatrywane zagadnienia mają skomplikowaną naturę, a dodatkowo ich precyzyjny opis jest utrudniony przez ograniczony dostęp do danych.
Tematyka rozważań jest aktualna i z praktycznego punktu widzenia przydatna.
Moim zdaniem dorobek naukowy Habilitanta spełnia wymagania sformułowane w art. 219 ustawy „Prawo o szkolnictwie wyższym i nauce" z dnia 20-07-2018 r..
Rekomenduję przejście do kolejnych etapów postępowania habilitacyjnego.