Politechnika Wrocławska

Katedra Inteligencji Obliczeniowej

Wrocław, 23.03.2015

dr hab. inż. Przemysław Kazienko, prof. nadzw.
Katedra Inteligencji Obliczeniowej, Politechnika Wrocławska
Wyb. Wyspiańskiego 27, 50-370 Wrocław
Email: kazienko@pwr.wroc.pl, tel: 71 3203609
WWW: http://www.ii.pwr.wroc.pl/~kazienko/

RECENZJA

Rozprawy doktorskiej mgr inż. Roberta Alberta Kłopotka
pt. *„Invasive Analysis of Social Networks"*

Rozprawę napisano pod kierunkiem prof. dr hab. inż. Krzysztofa Trojanowskiego i ukończono w 2014 roku.

Recenzja została wykonana na zlecenie Instytutu Podstaw Informatyki Polskiej Akademii Nauk w Warszawie zgodnie z pismem datowanym na 05.11.2014.

I. Przedmiot i problematyka rozprawy

Rozprawa doktorska mgr inż. Roberta Kłopotka poświęcona jest zagadnieniom związanym z modelowaniem sieci złożonych (*complex networks*) reprezentowanych przez grafy dwudzielne. Zasadniczym kierunkiem badań opisanych w pracy było zaproponowanie metod estymacji parametrów modelu *Cold Start User Item Model* (CSUIM), który został stworzony przez dr Chojnackiego. Jej głównym celem było więc opracowanie algorytmów i metod spełniających wymagania (i) wystarczającej dokładności oraz (ii) sensownej złożoności obliczeniowej dla przypadków grafów dwudzielnych odzwierciedlających zjawiska rzeczywiste. Należy przy tym zauważyć, że modelowanie z wykorzystaniem grafów dwudzielnych po przyjęciu pewnych założeń można stosować do rozwiązywania zadziwiająco dużej klasy problemów. A zatem, podjęta tematyka badań jest ważna poznawczo i posiada szerokie zastosowanie w praktyce.

Zarysowany w pracy zakres zawiera wiele ogólnych i szczegółowych problemów naukowych, które wymagały zastosowania odpowiedniego aparatu badawczego, zarówno teoretycznego jak i aplikacyjnego. Na szczególne podkreślenie zasługuje, zdaniem recenzenta, owo ukierunkowanie pracy na praktyczną możliwość zastosowania rozważanych metod teoretycznych. Metody teoretyczne są tutaj reprezentowane głównie przez modele parametrycznych grafów dwudzielnych, zaś praktyczny aspekt jest odzwierciedlony przez eksperymentalne badania nad wspomnianymi własnościami technicznymi, które jednak także wymagały zastosowania odpowiednich metod badawczych.

Wartym podkreślenia jest również interdyscyplinarny charakter recenzowanej rozprawy. Łączy ona różne dziedziny nauki, tj. matematykę (teoria grafów, rachunek prawdopodobieństwa), fizykę (sieci i systemy złożone) oraz informatykę (testowanie systemów informatycznych, algorytmy). Z tego powodu

Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław

tel: 71 320 35 16
fax: 71 321 10 18,
email: kio@pwr.edu.pl

www.kio.pwr.edu.pl

wymagała ona skutecznego i efektywnego połączenia wiedzy z różnych dziedzin, w tym także integracji wiedzy teoretycznej z praktyczną. Pomimo owego charakteru interdyscyplinarnego, praca mieści się w dziedzinie informatyki.

II. Zawartość i układ rozprawy

Recenzowana rozprawa doktorska liczy 135 stron (plus 7 stron z podziękowaniami i autoreferatem) i składa się z 5 jasno określonych rozdziałów, spisu oznaczeń oraz wykazu rysunków, tabel, algorytmów, jak również cytowanych prac. Została napisana poprawnym językiem angielskim.

W rozdziale pierwszym praca doktorska mgr inż. Kłopotka rozpoczyna się przedstawieniem motywacji podjętych badań ze szczególnym podkreśleniem ich istotności w analizie sieci społecznych. Zaprezentowano argumentację dla potrzeby modelowania sieci społecznych za pomocą grafów oraz metod opisujących ich dynamikę. Następnie umieszczono hipotezę badawczą bronioną w pracy, która stanowi, że model CSUIM może być użyty do analizy ilościowej grafów dwudzielnych odzwierciedlających rzeczywistość poprzez estymację jego parametrów na podstawie migawki sieci. Hipoteza rozważana jest dwutorowo: (i) poprzez zaproponowanie metod estymacji parametrów modelu, które pozwolą na identyfikację z wystarczającą precyzją oraz (ii) zaproponowanie metod wychwycenia konieczności adjustacji parametrów modelu w analizie sieci czasowych. W pracy zaproponowano metodologię prowadzenia badań polegającą na sprawdzaniu zgodności estymowanych parametrów modelu CSUIM z rzeczywistymi parametrami, które zostały użyte do wygenerowania sieci z wykorzystaniem modelu CSUIM. Oznacza to wprost, że sprawdzeniu podlegała skuteczność metody estymacji jedynie dla przypadku stacjonarnego (założenia modelu i realizacje sieci pochodzą z tego samego rozkładu) i dla syntetycznych danych. Jest to prawidłowe ustanowienie metody badawczej, ale bez szerokiego testowania metod dla danych rzeczywistych pozostawiłoby niedosyt w zakresie postawionej hipotezy. Rozdział pierwszy kończy się wymieniem oryginalnych dokonań oraz przeglądem treści pracy.

W rozdziale drugim omówione są podstawowe modele sieci uni- oraz bimodalnych z uwzględnieniem modeli dedykowanych dla sieci społecznych. Opis jest dość oszczędny, a wymienione wybrane modele sieci nie dają obrazu istniejącej różnorodności i bogactwa wiedzy na ten temat. Nie wspomniano choćby o modelach unimodalnych grafów losowych Erdos-Renyi-Gilbert, rodzinie wykładniczych grafów losowych, rodzinie modeli konfiguracyjnych, modelowaniu blokowym czy stochastycznych modelach blokowych. Rozdział zawiera przegląd istniejących podejść związanych z generowaniem grafów dwudzielnych bazujących na podstawowych miarach strukturalnych, zwłaszcza na stopniu węzła i współczynniku gronowania / grupowania (*clustering coefficient*). Rozdział ma charakter dość skondensowanego, acz uporządkowanego i momentami krytycznego przeglądu literaturowego. Autor zwrócił uwagę na odbiegający charakter grafów dwudzielnych, dla których nie ma zdefiniowanych niektórych miar i pojęć. Został przywołany współczynnik gronowania dla grafów dwudzielnych, wykorzystywany przez Autora w dalszych badaniach. Rozdział kończy się przedstawieniem modelu

Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław

tel: 71 320 35 16
fax: 71 321 10 18,
email: kio@pwr.edu.pl

www.kio.pwr.edu.pl

Cold Start User-Item Model, dla którego w dalszej części pracy proponowane są metody estymacji parametrów.

W rozdziale trzecim Autor przedstawia analityczną metodę estymacji parametów modelu CSUIM. Jej osiągnięcie jest zasadniczo poprawne (patrz uwagi szczegółowe), jednakże użyteczność dość ograniczona. Jest to spowodowane jego złożonością obliczeniową oraz zależnością od parametru k w modelowaniu prawdopodobieństwa określającego na ile dany węzeł posiada stopień mniejszy od k. Dalej pokazane są wyniki dokładności metody analitycznej zrealizowanej z dodatkowymi założeniami, pozwalającymi na efektywne jej wykorzystanie. Dla siedmiu sieci syntetycznych wygenerowanych przez model CSUIM z różnymi układami parametrów sprawdzany jest błąd relatywny estymowanych parametrów oraz podstawowych metryk opisujących grafy, tj. modularność, ilość grup, ilość wierzchołków typu *user*, ilość wierzchołków typu *item*, średni współczynnik gronowania dla wierzchołków typu *user* oraz *item*. Wydaje się, że zabrakło nieco rozważań na temat takiego doboru własności grafów, które miałby sprawdzać podobieństwo grafu wzorcowego z odtworzonym przy użyciu estymowanych parametrów modelu. Badanie ich mogłoby być potencjalnie użyteczne; w szczególności dotyczy to np. ścieżek czy średnicy grafu (patrz niezacytowane książki Bollobas B: *Random graphs*, Second Edition. Cambridge University Press, Cambridge, 2011 oraz książka dwóch Polaków – Janson S., Łuczak T., Ruciński A.: *Random Graphs*. John Willey & Sons, Inc, 2000) a także rozkładów struktur (a nie tylko pojedynczych miar, uwzględnianych przez Autora), patrz np. prace Snijdersa i Pattison (Snijders, T.A.B., Pattison, P., Robins, G.L., Handcock, M.: *New specifications for exponential random graph models*. Sociological Methodology, 2006, pp. 99-153) lub prace dotyczące motywów Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U.: *Network motifs: simple building blocks of complex networks*. Science, 298, 2004, pp. 824–827 oraz aktualny artykuł z Polski: Kotorowicz M., Kozitsky Y: *Motif based hierarchical random graphs: structural properties and critical points of an Ising model*. Condensed Matter Physics 2011, Vol. 14, No 1, 13801: 1-18. Sekcję poświęconą wynikom można by uzupełnić o szerszą analizę otrzymanych wyników, włączając w to analizę istotności statystycznej różnic wyników przy różnych układach parametrów.

Następnie Autor przedstawia metodę estymacji parametrów modelu CSUIM stosując podejście uczenia maszynowego. W tym celu określa 16 atrybutów opisujących wygenerowaną sieć, z których 13 używa do uczenia modeli regresji liniowej, maszyny wektorów nośnych (SVM), sieci neuronowej oraz procesu gaussowskiego. Wnioskowane są dwa parametry modelu: prawdopodobieństwo sposobu dołączania nowego węzła typu *item* (preferencyjny lub losowy) oraz udział ilości krawędzi dołączanych w sposób preferencyjny. Na potrzeby treningu modeli uczenia maszynowego zostały przygotowane syntetyczne zbiory danych (sieci) otrzymane przy stałych parametrach modelu CSUIM, prócz jednego modelowanego, dla których następuje liniowy przegląd wartości z dopuszczalnego zakresu. W pracy nie podano informacji o sposobie realizacji uczenia, a zwłaszcza o sposobie podziału na zbiór uczący-testowy. Dopiero w następnym zestawie eksperymentów poświęconych analizie wpływu wielkości próbki uczącej oraz rozmiaru grafu na jakość predykcji Autor uściśla, że eksperymenty będą realizowane wg typowej procedury walidacji krzyżowej. Otrzymane wyniki badań potwierdzają skuteczność i użyteczność zaproponowanego podejścia szacowania parametrów modelu CSUIM

Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław

tel: 71 320 35 16
fax: 71 321 10 18,
email: kio@pwr.edu.pl

www.kio.pwr.edu.pl

podczas modelowania miar sieciowych, a w mniejszym stopniu dla ustalania parametrów modelu.

W dalszej części rozdziału znajduje się interesująca dyskusja na temat modeli opisujących rozkład stopni węzłów w sieciach. Rozważana jest możliwość otrzymania analitycznego modelu dla przypadku mieszanego dołączania preferencyjnego i losowego. Słusznie stwierdzono, że dotychczas nie został ten problem rozwiązany a istniejące propozycje mają zastosowania tylko do przypadków skrajnych w sieciach unimodalnych. Również trafnie zauważono, że początkowa liczba krawędzi ma istotne znaczenie w formowaniu się sieci, na początku jej istnienia, ale traci znaczenie po uzyskaniu wystarczająco dużego rozmiaru. Interesujące są też zamieszczone w pracy wyniki obserwacji właściwości modelu CSUIM z zakresie prawie-liniowej zależności logarytmu stopnia węzła od wielkości parametru α (β) (odpowiadającego prawdopodobieństwu przyłączania nowych węzłów w sposób preferencyjny). Ponadto, zasadnie zauważono, że parametr γ modelu CSUIM odpowiadający za wielkość populacji wierzchołków uczestniczących w tzw. mechanizmie odbijania (*bouncing*) jest zależny do modularności. Można dodać, że jest to stwierdzenie tym bardziej właściwe im wyższy poziom homofilii w grafie.

W podrozdziale 3.7 zebrano wcześniejsze rozważania na temat wyznaczania parametrów modelu CSUIM i zaproponowano cztery algorytmy realizujące to zadanie na zasadzie połączenia podejścia regresji i symulacji. Rozdział trzeci kończy się na zaprezentowaniu wyników eksperymentów i konkluzjami na temat otrzymanych wyników. Z trzech zaproponowanych podejść najlepszym, wg Autora, okazuje się podejście działające na zasadzie połączenia regresji i symulacji. Nie jest to jednak poparte dogłębną analizą wyników, a jedynie obserwacją wartości bezwzględnych uzyskiwanych błędów.

Nie sposób tutaj skomentować stosunkowo enigmatycznej struktury rozdziału trzeciego, która utrudnia poruszanie się po jego dość rozbudowanej zawartości.

Kolejny, czwarty rozdział, jest poświęcony zagadnieniu odkrywania zmian parametrów generujących w modelu CSUIM. Rozważany był scenariusz jednoczesnej zmiany tylko jednego z parametrów w trakcie rozrostu sieci. W celu efektywniejszego odkrywania zmian parametrów zaproponowano ulepszenie obliczania wykładnika potęgowego dla metody z rozdziału 3.5, które znacznie skraca czas obliczeń oraz poprawia jakość estymacji. Następnie dla parametrów α, β, γ, *du* i *dv* prezentowane są wyniki eksperymentalnej obserwacji zmian estymowanych parametrów po wprowadzeniu zaburzenia w połowie czasu generowania sieci. Przedstawione wyniki zostały opatrzone komentarzem potwierdzającym, że istnieje możliwość obserwowania zmian estymowanych parametrów po zaistnieniu zaburzenia w generowanej sieci przez model CSUIM. Można by się jednak tutaj pokusić o nieco idące dalej wnioski i zaproponować automatyczną metodę identyfikacji zaburzeń stacjonarności rozkładu parametrów, które adoptowałyby znane z literatury uczenia maszynowego rozwiązania dla detekcji „dryftu" czyli zmian w charakterystyce danych (*concept drift*).

W rozdziale piątym przedstawiono podsumowanie przeprowadzonych i opisanych wcześniej prac oraz ogólnie nakreślono możliwości dalszych badań naukowych.

Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław

tel: 71 320 35 16
fax: 71 321 10 18,
email: kio@pwr.edu.pl

www.kio.pwr.edu.pl

Rozprawa zawiera 113 cytowanych pozycji bibliograficznych, w tym 3 publikacje, których doktorant jest autorem lub współautorem.

Pod koniec umieszczono również bardzo użyteczny spis oznaczeń.

III. Oryginalne osiągnięcia

Oryginalne osiągnięcia zawarte w rozprawie mieszą się w rozdziałach 3 i 4. Należą do nich w szczególności:

1. przedstawiona w podrozdziale 3.1.1 analiza teoretyczna sposobu wyznaczania parametrów modelu CSUIM,
2. zaproponowanie nowego podejścia do wyznaczania parametrów modelu CSUIM, z wykorzystaniem standardowych metod uczenia nadzorowanego (rozdział 3.2)
3. zaproponowanie sposobu wyznaczania parametrów modelu CSUIM w postaci czterech algorytmów działających na zasadzie połączenia podejścia regresji i symulacji pkt. 3.7.
4. zaprezentowanie sposobu detekcji zmiany parametrów generujących w modelu CSUIM, rozdz. 4.

Należy zauważyć, że kierunek badań zawartych w rozprawie jest ważny zarówno z poznawczego jak i z praktycznego punktu widzenia. Uzyskane przez Autora rezultaty wyraźnie wskazują, że charakterystyka grafów (danych źródłowych) wykorzystywanych przez systemy rekomendacyjne ma istotny a czasami nawet zasadniczy wpływ na ich parametry techniczne, co jest ważnym osiągnięciem rozprawy.

IV. Uwagi krytyczne i dyskusyjne

Należy zauważyć, że poniższe uwagi nie umniejszają w zasadniczy sposób dokonań Autora wymienionych w poprzednim punkcie.

Uwagi ogólne

1. Zarówno opis jak i same badania zaproponowanych przez Autora metod pozostawiają pewien niedosyt w zakresie analizy porównawczej. Autor mógł się pokusić o nieparametryczną ocenę istotności statystycznej uzyskiwanych wyników dla poszczególnych podejść. Ponadto eksperymenty dla każdej z badanych metod lepiej, aby były przeprowadzone na wspólnych danych, zarówno syntetycznych jak i rzeczywistych. Kolekcja użytych zbiorów danych rzeczywistych mogłaby być znacznie bardziej rozbudowana.
2. Tytuł pracy, brzmiący „*Invasive Analysis of Social Networks*", jest dość niejednoznaczny w obliczu zawartości doktoratu. W pracy trudno znaleźć precyzyjne opisy 'inwazyjnych' sposobów analizy sieci społecznych. W szczególności należy zwrócić uwagę, że Autor skupia się na tylko jednym rodzaju sieci, mianowicie na tych, które są odzwierciedlane poprzez grafy dwudzielne. Należy także zauważyć, że w istocie rozwiązania i sieci

Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław

tel: 71 320 35 16
fax: 71 321 10 18,
email: kio@pwr.edu.pl

www.kio.pwr.edu.pl

rozważane w pracy są bardziej ogólne. Sieci takie określa się mianem sieci złożonych (*complex networks*), zaś to, że jednym z *modów* sieci są użytkownicy, nie ma istotnego znaczenia dla proponowanych rozwiązań.

3. W pracy zabrakło rozważań na temat sposobu doboru własności grafów, które miałby sprawdzać podobieństwo grafu wzorcowego z grafem odtworzonym przy użyciu estymowanych parametrów modelu. Badanie ich mogłoby być potencjalnie użyteczne. Ponadto porównanie mogłoby wybiegać poza zastosowanie zagregowanych postaci miar sieciowych (głównie średnie) i użycie empirycznych rozkładów wielkości opisujących grafy. Przykładowo, można by rozpatrzyć podobieństwo rozkładów stopnia węzła dwóch grafów (patrz dywergencja Kullbacka-Leiblera), zamiast ich wartości średnich.
4. W pracy można by dołączyć omówienie potencjalnych możliwości rozwiązań zadania estymacji parametrów modelu CSUIM, które byłoby traktowane jako zadanie wielokryterialnej optymalizacji. Rozważanie mogłoby być uzupełnione o wskazanie innych, bardziej wyrafinowanych, metod uczenia maszynowego, które potrafią modelować dane ustrukturalizowane. Dzięki temu można by pokusić się o wnioskowanie wszystkich parametrów modelu jednocześnie lub *prawie*-jednocześnie.
5. Recenzent odniósł wrażenie, że przedłożona praca zawiera fragmenty o bardzo wysokiej jakości merytorycznej i prezentacyjnej. Jednakże fragmenty te przeplatane są innymi częściami, które wprowadzają utrudnienia w śledzeniu toku rozumowania oraz nawigowaniu po pracy. Na uwagę zasługuje jednak fakt, że praca była pisana w języku angielskim, co może powodować dodatkowe utrudnienia w odbiorze.
6. Przedstawiona w pracy możliwość obserwowania zmian estymowanych parametrów po zaistnieniu zaburzenia w generowanej sieci modelu CSUIM w istocie nie pokrywa zakresu zakładanego w celach pracy. Można bowiem rozważać zaproponowanie automatycznej metody identyfikacji zaburzeń stacjonarności rozkładu parametrów, które adoptowałyby znane z literatury uczenia maszynowego rozwiązania dla detekcji dryftu tj. zmian charakterystyki (*concept drift*).

Uwagi szczegółowe

1. Opis parametru δ w modelu *Cold Start User-Item Model* (str. 16) jest myląco związany przez parametr numeru iteracji t. W odczytywanych intencjach wyrażenia chodziło raczej o pokazanie, że obowiązuje ono dla każdej iteracji.
2. Na str. 18 wyrażenie $E(|U(t)|)=m+\delta t$ oraz następujące po nim wyrażenie liczności wartości oczekiwanej $V(t)$ w połączeniu z ilością $2m$ inicjalną wierzchołków w przywoływanym modelu CSUIM działa jedynie przy równej liczbie wierzchołków typu *user* i typu *item*. Dalsza dyskusja na temat uproszczeń zależności przy dużej liczbie iteracji jest prawdziwa jedynie dla $m<<t$. Nie wynika to z treści pracy.

Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław

tel: 71 320 35 16
fax: 71 321 10 18,
email: kio@pwr.edu.pl

www.kio.pwr.edu.pl

3. Na str. 13 oznaczenie sąsiedztwa $N_S(n)$ wierzchołka n przedstawione w opisie słownym jest sprzeczne w porównaniu z następującym w kolejnym zdaniu zapisem formalnym.
4. We wzorze 3.1 na str. 21 nie wprowadzono oznaczenia η. Nie jest też on parametrem rozważanego modelu CSUIM. Widnieje jedynie odniesienie w liście użytych oznaczeń.
5. W sekcji 3.1.1 poświęconej analitycznej estymacji parametrów modelu CSUIM występują uchybienia w postaci użytych symboli. We wzorze 3.3 występuje wcześniej nie wprowadzone oznaczenie I, które powinno raczej być V, czyli zbiorem wierzchołków typu *item* zgodnie z wprowadzeniem w pkt. 2.4.

V. Podsumowanie i ocena rozprawy

Podsumowując, należy stwierdzić, że rozprawa pomimo pewnych zastrzeżeń stanowi istotny wkład w dziedzinę modelowania i generowania grafów losowych o zadanych parametrach możliwie bliskich grafom będącym odwzorowaniem rzeczywistych sieci społecznych. Autor uzyskał wartościowe i oryginalne z naukowego punktu widzenia wyniki zawarte zwłaszcza w sposobie wyznaczania parametrów modelu CSUIM jak i zaproponowanych algorytmów.

Doktorant posiada akceptowalny, choć nierozległy, dorobek naukowy wyrażony przez 3 anglojęzyczne artykuły i referaty konferencyjne zacytowane w rozprawie. W uznanej bazie *Scopus* zarejestrowanych jest obecnie 2 prace doktoranta, w tym jedna nierelewantna z tematyką i niezacytowana w rozprawie. W żadnej z prac nie występuje Promotor jako współautor.

W związku z powyższym stwierdzam, ze opiniowana rozprawa doktorska mgr inż. Roberta Kłopotka spełnia wymagania stawiane w obowiązujących przepisach ustawy o stopniu naukowym doktora i wnoszę o dopuszczenie jej Autora do publicznej obrony.

Kwiecień P.

Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław

tel: 71 320 35 16
fax: 71 321 10 18,
email: kio@pwr.edu.pl

www.kio.pwr.edu.pl