Prof. dr hab. Tomasz Łuczak
Wydział Matematyki i Informatyki
Uniwersytetu im. Adama Mickiewicza
w Poznaniu

Poznań, dnia 9 czerwca 2019 roku

Recenzja rozprawy doktorskiej Piotra Wygockiego „On nearest neighbors"

Rozprawa doktorska mgra Piotra Wygockiego składa się z dwóch, luźno ze sobą powiązanych, części. Pierwsza z nich dotyczy ważnego *problemu znajdowania najbliższych sąsiadów*, gdy, mając dany ustalony na początku skończony zbiór punktów $X \subset \mathbb{R}^d$, dla każdego zadanego punktu $q \in \mathbb{R}^d$ należy znaleźć punkt X leżący najbliżej q. Autor skupił się na pewnej szczególnej wersji tego problemu, dla której znalazł algorytmy typu Las Vegas pozwalające na jego stosunkowo szybkie rozstrzygnięcie. Zaproponowane przez niego procedury w zmyślny sposób łączą znane metody redukcji wymiaru i „kwantyzacji" przestrzeni z konstrukcją pewnych specjalnych funkcji skrótu, a analiza ich działania, miejscami bardzo techniczna i skomplikowana, wymagała zastosowania niebanalnych narzędzi rachunku prawdopodobieństwa. Wysiłek ten jednak w pełni się opłacił – w chwili powstania podane w pracy algorytmy były najbardziej wydajne ze wszystkich znanych procedur, zarówno jeśli chodzi o czas wstępnego przetwarzania danych początkowych opisujących zbiór X jak i czas potrzebny do ustalenia punktu najbliższego zadanemu punktowi q, a w niektórych przypadkach ich złożoność była bliska optymalnej. Nie dziwi zatem, że znaczna część opisanych w rozprawie wyników została zaprezentowana na dwóch dobrych konferencjach informatycznych, a następnie stała się podstawą dwóch publikacji doktoranta, przy czym współautorami pierwszej byli Andrzej Pacuk, Piotr Sankowski i Karol Węgrzycki, a druga została napisana wespół z Piotrem Sankowskim. Zawartość merytoryczną tej części pracy oceniam wysoko, zarówno jeśli chodzi o znaczenie poruszanego zagadnienia, zaawansowanie zastosowanych metod, jak i wreszcie wagę osiągniętych wyników.

Druga część pracy poświęcona jest problemowi rozpowszechniania informacji w sieciach typu „małych światów". Autor bada parametry sieci Twitter istotne dla szybkości rozchodzenia się informacji, wskazując na pewne istotne rozbieżności uzyskanych przez siebie wyników z przewidywaniami najbardziej rozpowszechnionych teoretycznych modeli rozchodzenia się informacji. Następnie podaje serię stosunkowo prostych twierdzeń dotyczących rozprzestrzeniania się informacji w pewnym modelu teoretycznym, opartym na grafach losowych, którego własności, jego zdaniem, dobrze przybliżają własności sieci „rzeczywistych" takich jak analizowana wcześniej sieć Twitter.

Aczkolwiek trudno nie mieć uznania dla wysiłku autora włożonego w przebadanie tak ogromnej struktury jaką jest Twitter, poziom drugiej części rozprawy jest, moim zdaniem, nie tak wysoki jak części pierwszej. Owocem badań sieci Twitter są wykresy, które, zdaniem autora, pokazują poważne odchylenia jej własności od przewidywań teoretycznych. Przyznam, że wyglądają one przekonywująco, niemniej zabrakło mi tutaj, na przykład, jakiejkolwiek analizy statystycznej. Jeśli chodzi o rozważania teoretyczne, analiza modelu $\overline{SIR}$ opartego na skierowanym grafie losowym zaproponowana przez autora sprowadza się do analizy struktury tegoż grafu losowego z przeskalowanym parametrem p. Ten model grafu rozpatrywany był przez A.Baraka i P.Erdősa (*SIAM J. Algebraic Discrete Methods 5 (1984) 508–514*) i od tego czasu jego własności zostały dobrze zbadane. Na przykład, wzór rekurencyjny na prawdopodobieństwo $p_{n,k}$, kluczowy dla rozważań autora, pojawia się w pracy K.Simona (*Theoretical Computer Science 58 (1988), 325–346*) jako wzór (11) w Lemacie 3.2.

Przejdę teraz do oceny redakcji recenzowanej rozprawy. Należy z uznaniem podkreślić, że autor jasno zdefiniował poruszany problem, wyczerpująco opisał uzyskane przez siebie i swoich współautorów wyniki i czytelnie przedstawił ogólną ideę dowodu poprawności proponowanych algorytmów. Niestety, zaprezentowana w rozprawie analiza poszczególnych etapów działania algorytmu pełna jest luk, niedopowiedzeń i niezrozumiałych kroków. Wymownym tego przykładem jest Lemat 3.3.8, którego pełną treść przytaczam poniżej:

> *For any random vector* $v \in \mathbb{R}^d$ *of independent random variables such that* $\mathbb{P}[v_i = \pm 1] = \pm\frac{1}{2}$, *for every* $p \in [1, \infty]$ *and* $||x||_2 = 1$ *and* $\kappa_{c,p} < \frac{1}{\sqrt{2}}$, *it holds:*
>
> $$\mathbb{P}[|\langle v, x\rangle| \leqslant \alpha] \leqslant 1 - \frac{(1 - \sqrt{2}\alpha)^2}{2}.$$

Fakt, że w założeniu występują p i $\kappa_{c,p}$, natomiast nieobecne jest występujące w tezie lematu α, bynajmniej nie pomoga czytelnikowi w podążaniu za rozumowaniem autora. W tym miejscu nietrudno jest błąd poprawić (p w założeniu jest zbędne, a $\alpha = \kappa_{c,p}$), w innych wypadkach zadanie to nie jest już takie łatwe. Na przykład w tezie Lematu 3.3.5 prawdopodobieństwo $\mathbb{P}[\langle v, x\rangle \leqslant \alpha]$ szacowane jest z góry przez β_α, dla „α spełniającego warunek $\Phi(\alpha)$ dla $||x||_2 = 1$". Niestety, w rozprawie nie mogłem się doszukać ani sformułowania warunku $\Phi(\alpha)$, ani definicji stałej β_α, zapewne występującej w tym warunku. Takich przypadków niezdefiniowanych stałych pojawiających się ni stąd ni zowąd, czy niezrozumiałych i niepotrzebnych założeń, jest w pracy więcej. Podobne uchybienia nie powinny pojawić się w rozprawie doktorskiej. Autor winien zwrócić szczególną uwagę by unikać podobnych błędów w przyszłości, bowiem nie dość, że utrudniają czy wręcz uniemożliwiają czytanie pracy, to mogą budzić podejrzenia, że nie do końca poprawnie rozumie on matematyczną zawartość formułowanych twierdzeń.

Niemniej, ze względu na wagę wyników zawartych w pierwszej części rozprawy uważam, że mimo istotnych usterek redakcyjnych, przedstawiona praca spełnia ustawowe i zwyczajowe warunki stawiane rozprawom doktorskim w dziedzinie nauk matematycznych w zakresie informatyki i wnoszę o dopuszczenie mgra Piotra Wygockiego do dalszych etapów przewodu doktorskiego.