prof. dr hab. Krzysztof Diks
Instytut Informatyki
Uniwersytet Warszawski
diks@mimuw.edu.pl

Warszawa, 22.04.2022

# Recenzja rozprawy doktorskiej Pana Dominika Bojko *Selected Topics on Randomized Algorithms*

Pan Dominik Bojko złożył pracę doktorską w dziedzinie nauk ścisłych i przyrodniczych w dyscyplinie matematyka. Jej główne wyniki dotyczą ścisłej, matematycznej analizy wybranych algorytmów probabilistycznych w trzech różnych obszarach zastosowań. Jako informatyk w swojej recenzji skupiam się na informatycznej wadze osiągniętych wyników a mniej na ich matematycznym aspekcie.

W swojej rozprawie autor zajmuje się trzema ważnymi problemami algorytmicznymi:

- wybór lidera (ang. *leader election*),
- próbkowanie (w czasie rzeczywistym) do rezerwuaru (ang. *reservoir sampling*),
- prywatność różnicowa (ang. *differential privacy*).

Każdy z tych problemów ze względu na wagę swoich zastosowań był i jest intensywnie badany. Z natury tych problemów wynika, że najlepszymi metodami ich rozwiązywania są metody probabilistyczne a rolą badacza jest wykazanie skuteczności zaproponowanych rozwiązań. Proponowane algorytmy są zazwyczaj proste i intuicyjne, natomiast wyzwaniem staje się precyzyjna analiza ich skuteczności. Tak jest właśnie w przypadku ocenianej rozprawy doktorskiej. Autor proponuje (lub adaptuje istniejące) algorytmy rozwiązywania wspomnianych wyżej problemów, a następnie metodami matematycznymi (probabilistycznymi) dokonuje analizy ich skuteczności pod kątem wybranych kryteriów, potwierdzając jednocześnie eksperymentalnie wyniki otrzymane analitycznie.

## Wybór lidera

Problem wyboru lidera jest jednym z podstawowych w systemach rozproszonych, gdy grupa niezależnych, ale współpracujących ze sobą agentów musi wybrać koordynatora swoich działań. Zadanie staje się nietrywialne, gdy agenci się nie znają a środowisko, w którym działają może się zmieniać dynamicznie lub jest zawodne – część agentów znika, pojawiają się nowi, komunikacja jest obarczona błędami. Modele obliczeń rozproszonych są często wieloparametrowe a dobór parametrów ma istotny wpływ na przyjęte rozwiązania. Jednym z najważniejszych parametrów jest schemat komunikacji agentów. Jest on zadany za pomocą sieci bezpośrednich połączeń (grafu). W najprostszej sytuacji każdy może komunikować się z

każdym (graf pełny). Agenci w sieci mogą komunikować się albo wymieniając bezpośrednio między sobą komunikaty po łączach, albo przesyłając komunikaty jednocześnie wszystkim w swoim otoczeniu. W tym drugim przypadku, gdy wiele komunikatów jest wysyłanych jednocześnie, mamy do czynienia z kolizjami, które nie zawsze mogą być wykrywane. Ważną cechą obliczeń jest synchroniczność lub jej brak. Obliczenia synchroniczne odbywają się w rundach, z których każda składa się z wysłania/otrzymania komunikatu i obliczeń lokalnych przygotowujących następną rundę. W sieciach anonimowych liczba uczestniczących w obliczeniach agentów nie jest znana. W przypadku sieci quasi-anonimowych znamy górne ograniczenie na liczbę agentów, zaś w przypadku sieci nie anonimowych liczba agentów jest ściśle określona. Dodatkowym ograniczeniem jest postać komunikatów, które agenci wymieniają między sobą. W najprostszym modelu komunikaty są bitowe, gdzie jedynce odpowiada sygnał dźwiękowy, tzw. bip. Mówimy w tym przypadku o modelu sygnału dźwiękowego (ang. *beep model*).

W recenzowanej dysertacji wyboru lidera dokonuje się w pełnej, nie anonimowej lub quasi-anonimowej sieci, w komunikacji synchronicznej z użyciem sygnałów dźwiękowych (model typu bip), bez wykrywania kolizji. Z punktu widzenia jakości proponowanych rozwiązań interesują nas algorytmy, w których każdy uczestnik obliczeń może zostać liderem na równych prawach, w każdych powtarzanych obliczeniach. Mówimy wówczas o powtarzalnej uczciwości. Większość omawianych w rozprawie algorytmów jest typu *Atlantic City* z kontrolowanym prawdopodobieństwem sukcesu.

Przyjęty w rozprawie model obliczeń uważam za w pełni uzasadniony. Obejmuje on podstawowe problemy związane z wyborem lidera a ich rozwiązanie, jak też związana z tym analiza, pozwala na lepsze zrozumienie natury problemu.

Podstawą algorytmów rozważanych przez doktoranta jest tzw. model urnowy. W tym modelu przyjmujemy, że mamy pewną liczbę urn ponumerowanych kolejnymi liczbami naturalnymi. W jednej rundzie każdy agent wrzuca piłeczkę ze swoim identyfikatorem do losowo wybranej urny, przy zadanym, takim samym dla wszystkich rozkładzie prawdopodobieństwa wyboru. Liderem zostaje agent, którego piłeczka znajdzie się w urnie z największym numerem. Oczywiście runda nie musi skończyć się sukcesem, gdy do urny z największym numerem trafi więcej niż jedna piłeczka. W takim przypadku można powtórzyć rundę, a w celu przyśpieszenia obliczeń w kolejnej rundzie udział mogą wziąć tylko agenci, których piłeczki trafiły do urny z największym numerem. Uwaga, w pracy pomija się komunikację wyników obliczeń w rundzie, co z punktu widzenia zastosowań jest niezwykle ważne i wpływa na szybkość całego procesu wyboru lidera. Z punktu widzenia skuteczności powyższego algorytmu niewątpliwie ważne są liczba uczestniczących w procesie agentów i liczba dostępnych urn.

W rozprawie podstawowym modelem w analizie jest model sygnału dźwiękowego. Parametrem tego modelu jest rozkład prawdopodobieństwa $D$ nad zbiorem liczb naturalnych, jego parametry $\Theta$ oraz liczba urn $L$. Każdy agent losowo generuje numer urny z zadanym rozkładem. Jeśli ten numer jest większy od $L$, to wybiera $L$ jak numer urny, do której chce wrzucić swój identyfikator. Następnie bierze binarną reprezentację wylosowanej liczby o długości $K = \lceil \log L \rceil$. Teraz algorytm wyboru lidera odbywa się w $K$ rundach ponumerowanych od $K$-1 do 0. Na początku żaden z agentów „nie śpi". W rundzie o numerze $i$ wszyscy nieuśpieni agenci, dla których w wylosowanych numerach na $i$-tej pozycji w zapisie binarnym jest

jedynka, wysyłają sygnał dźwiękowy. Agenci, którzy nie wysłali sygnału, ale go usłyszały usypiają. Liderem zostaje agent, który nie zostaje uśpiony. Przedmiotem analizy algorytmu jest dobór parametrów rozkładu tak, żeby skutecznie (z kontrolowanym prawdopodobieństwem) wybierać lidera, ale robić to oszczędnie – przy możliwie minimalnej liczby rund $K$. Innym ciekawym parametrem jakości algorytmu jest liczba wysłanych komunikatów, co jest szczególnie ważne w aspekcie tzw. energooszczędnych (zielonych) obliczeń.

W podrozdziale 2.5 autor analizuje nie anonimowy, urnowy algorytm wyboru lidera przy znanych $n$ (liczba agentów) i $L$. Interesuje go rozkład prawdopodobieństwa, który maksymalizuje prawdopodobieństwo sukcesu wyboru lidera. Rozkład ten został w pełni scharakteryzowany w twierdzeniach 2.5.1 i 2.5.3. Formuły wyznaczające rozkład są numerycznie skomplikowane oraz obliczeniowo trudne i obarczone błędami zaokrągleń. (Swoją drogą autor mógłby pokusić się o precyzyjną analizę złożoności obliczania optymalnego rozkładu.) Żeby ominąć problem obliczeniowej trudności wyznaczania optymalnego rozkładu autor zaproponował metodę obliczania rozkładu przybliżonego, choć jego dokładna analiza, w szczególności błędu przybliżenia, została przeniesiona do dalszych badań. W zamian autor potwierdził jakość zaproponowanych rozwiązań eksperymentalnie. Wartościowym wynikiem w tej części jest podanie dokładnej wartości $K$ (długość zapisu binarnego $L$ – liczby urn) dla osiągnięcia sukcesu przy zadanym prawdopodobieństwie. Parametr $K$ odpowiada zarówno za czas działania algorytmu (w modelu sygnału dźwiękowego), jak też za liczbę wymienianych komunikatów.

W podrozdziale 2.6 autor analizuje algorytm wyboru lidera w sieciach quasi-anonimowych, gdy dane jest ograniczenie górne $N$ na liczbę agentów. Główny wynik tego rozdziału to podanie optymalnej wartości parametru $K$ (a tym samy także liczby urn $L$), dla wyboru lidera kończącego się sukcesem z zadanym prawdopodobieństwem, przy czym do losowania urn wykorzystuje się rozkład jednostajny.

Podrozdział 2.7 jest poświęcony znanym, podobnym algorytmom i ich alternatywnym wersjom zaproponowanym przez autora.

Podrozdziały 2.8 i 2.9 dotyczą wyboru lidera w modelu sieci quasi-anonimowej, przy założeniu geometrycznego rozkładu prawdopodobieństwa wyboru urn. Uzasadnieniem takie postępowania są obliczenia energooszczędne zaproponowane przez Jacquet. Głównym wynikiem rozdziału 2.8 jest podanie metody wyznaczania parametru rozkładu $p$ oraz liczby urn $L$ tak, żeby osiągnąć sukces z zadanym prawdopodobieństwem. Autor zadbał także o to, żeby pokazać jak efektywnie obliczeniowo wyznaczać optymalny rozkład, jak też podał prostszy sposób wyznaczania rozkładu przybliżonego. W rozdziale 2.9 autor proponuje dwufazowy wybór lidera. W pierwszej fazie stosowany jest rozkład geometryczny tak, żeby szybko ograniczyć zbiór potencjalnych kandydatów na lidera. W drugiej fazie stosuje się algorytm z rozkładem jednostajnym. Główny wynik drugiego rozdziału, najbardziej zaawansowany technicznie, to wyznaczenie parametrów algorytmu tak, żeby osiągnąć sukces z zadanym prawdopodobieństwem.

W podsumowaniu rozdziału autor podał wyniki eksperymentalne potwierdzające jakość przyjętych rozwiązań.

Wyniki rozdziału 2. potwierdzają, że autor dogłębnie i sprawnie technicznie potrafi dokonać analizy algorytmów wyboru lidera w bardzo naturalnym modelu. Autor wykazuje optymalność swoich rozwiązań zarówno analitycznie, jak i eksperymentalnie. Jego praca pozwala lepiej zrozumieć naturę problemu wyboru lidera w ujęciu probabilistycznym.

## Próbkowanie danych do rezerwuaru

Ta część rozprawy dotyczy ważnego aspektu analizy danych wielkiego rozmiaru, z których chcemy pobrać reprezentatywną losową próbkę. Zakładamy przy tym, że przetwarzanie jest strumieniowe i próbkowanie odbywa się w czasie rzeczywistym a po analizie każdej kolejnej danej mamy losową próbkę z tego, co dotychczas widzieliśmy. W niektórych zastosowania dane w próbce muszą pochodzić z ograniczonej grupy ostatnio przetwarzanych danych zwanej oknem i żądamy, żeby własność losowości odnosiła się do próbki z okna. Dodatkowym ograniczeniem jest to, że rozmiar pamięci, którą dysponujemy w stosunku do ogromu przetwarzanych danych, jest pomijający (stały, niezależny od wielkości danych). Przechowywane dane nie mogą pochodzić spoza aktualnego okna, a wszystkich danych z okna tez nie jesteśmy w stanie zapamiętać. W rozprawie Autor skupia się na próbkowaniu do jednoelementowego rezerwuaru z przesuwającego się okna, przy wykorzystaniu pamięci stałego rozmiaru niezależnego od wielkości okna.. Proponuje alternatywną metodę próbkowania dla rozkładu jednostajnego do tej zaproponowanej przez Bravermanna, Ostrovskyego i Zaniolo. Metoda jest na tyle ogólna, że umożliwia próbkowanie z zadanym, „dopuszczalnym" rozkładem. W pracy podano wystarczające warunki dla rozkładu dopuszczalnego i przedstawiono kilka praktycznie stosowalnych rozkładów. Przeprowadzono także dyskusję ograniczeń dla klas dopuszczalnych rozkładów. Główny wynik tego rozdziału to zastosowanie łańcucha Markowa typu „szatańskie schody" (ang. *devil's staircaise*) do generowania próbek okienkowych z różnymi rozkładami prawdopodobieństwa dla pozycji w próbce. Główny pomysł polega na symulacji ruchów na „szatańskich schodach". Niestety bezpośrednie zastosowanie takiego podejścia nie pozwala na uzyskanie próbki z rozkładem jednostajny. W podrozdziale 3.2.7 autor wykazuje jednak, że symulacja dwóch schodów umożliwia uzyskanie próbki z rozkładem jednostajnym. Uważam, że algorytmy zaprezentowane w tym rozdziale są nietrywialne i formują najlepszą część pracy. Potwierdzeniem wyników analitycznych są wyniki eksperymentalne. Pozostaje analiza złożoności obliczeniowej zaproponowanych algorytmów. Autor sygnalizuje ten problem, ale pomija go w swojej dysertacji.

Ostatnia część rozdziału 3. jest poświęcona próbkowaniu z zadanym ciągiem prawdopodobieństw określającym z jakim prawdopodobieństwem $n$-ty element z wejściowego ciągu danych ma być brany do próbki. Autor analizuje w jaki sposób dobierać prawdopodobieństwa, żeby w próbce pojawiały się elementy blisko interesujących nas pozycji w ciągu wejściowym. Uzasadnieniem takiego podejścia są dobrze dobrane przykłady

Rozdział 3 jest dla recenzenta najbardziej interesujący ze względu na ważne, współczesne zastosowania do przetwarzania wielkich wolumenów danych w czasie rzeczywistym. Rozdział ten jest też najbardziej zaawansowany algorytmicznie.

## Prywatność różnicowa

Zróżnicowana prywatność umożliwia gromadzenie i udostępnianie zagregowanych informacji użytkowników, przy jednoczesnym zachowaniu prywatności pojedynczych użytkowników. W tym celu dodaje się losowe szumy do zagregowanych danych tak, żeby danych pojedynczego użytkownika nie można było zidentyfikować (prawdopodobieństwo takiego faktu jest niezwykle małe i możemy nim sterować parametrami zaszumiania). W swojej rozprawie autor wykazał przydatność liczników probabilistycznych do agregacji wyników ankiety przy zachowaniu prywatności różnicowej. Wkładem autora jest wykazanie, że w pewnych sytuacjach liczniki probabilistyczne Morrisa i MaxGeo mogą być konkurencyjne do powszechnienie stosowanego zaszumiania Laplace'a. Obserwacja autora rozprawy jest świeża a analiza jakości prywatności różnicowej liczników probabilistycznych zaawansowana technicznie.

## Podsumowanie

**Przedstawiona do recenzji praca spełnia wszystkie warunki bardzo dobrej rozprawy doktorskiej.** Problemy badawcze, którymi zajmuje się autor są w jądrze zainteresowań badaczy w obszarach, których dotyczą. Wybór lidera jest naturalnym problemem badanym od lat w środowisku obliczeń rozproszonych, natomiast próbkowanie do rezerwuaru i agregacja danych przy zachowaniu prywatności różnicowej to gorąca tematyka związana z przetwarzaniem i udostępnianiem (w zagregowanej postaci) wielkich wolumenów danych. Główne wyniki pracy to precyzyjna analiza probabilistyczna istniejących rozwiązań i autorskich propozycji ich usprawnień. Wyniki analityczne są często potwierdzane eksperymentalnie. Autor wykazał się biegłością w stosowaniu metod probabilistycznych do analizy losowych procesów kombinatorycznych. Praca jest dobrze zredagowane i ustrukturyzowana. Zawiera wszystkie niezbędne definicje, tak że przy jej czytaniu nie ma potrzeby do sięgania do źródeł zewnętrznych. Dyskusja zaproponowanych rozwiązań pozwala docenić jakość wyników na tle innych rozwiązań w tematyce, której dotyczą. **Nie mam wątpliwości, że rozprawa *Selected Topics on Randomized Algorithms* może być podstawą nadania panu Dominikowi Bojko stopnia naukowego doktora w dyscyplinie matematyka.**

Signature Not Verified
Dokument podpisany przez KRZYSZTOF DIKS
Data: 2022.04.26 06:32:06 CEST

Krzysztof Diks