"Selected Topics on Randomized Algorithms" — an abstract

Dominik Bojko

This doctoral thesis gathers results from different projects related to probability theory, stochastic processes and algorithmic. It is divided into three main branches: Leader Election Algorithms, Reservoir Sampling Algorithms and Differential Privacy of Probabilistic Counters.

First chapter gathers outline, notations and descriptions of basic concepts utilized in the dissertation.

The second one is devoted to Leader Election algorithms. Firstly, we present motivation behind this concept and very extensive descriptions of possible general arrangements of such the procedures and theirs succinct history. Next we introduce urn model, justify its usability and provide a basic block of randomized Leader Election algorithm consistent with this framework.

Further we consider two main branches of arrangements, namely procedures designated for a fixed number n of devices given a priori and for any number of entities in the network, bounded by its capacity N.

A first substantial contribution is for the first approach. For any given finite set of possible identifiers, we attain the optimal distribution with respect to probability of successful choice of a leader. We utilize several uncommon tricks and the Lagrange's multipliers method to do so. Calculation of the probability mass function of the optimal solution for big n occurs to be time-consuming, hence we also provide precise approximations of optimal distribution via method proposed by S.Stević (which approximates the recurrence sequences), together with Banach's fixed point theorem. The latter part of this section reveals an astonishing fact about necessary and sufficient memory supplies (and simultaneously bounds on runtime of procedure) to guarantee correctness of Leader Election in urn model for a given number of users with probability at least $1 - \varepsilon$. Namely, the neccessary number of bits is $-\lceil \lg(\varepsilon) \rceil$, but sometimes it is needed to use one bit more, which is the sufficient number of bits.

A next part is a foundation of the rest of the first chapter. It introduces Uniform Leader Election algorithm in urn model. Further we provide a novel, yet very natural result by techniques characteristic for difference equations theory. Namely, we provide a monotonicity of the success probability with respect to the number of competitors. Next we provide a novel, general and optimal lower bound on the time complexity of Leader Election algorithms which should be reliable for any number of devices in a network with bounded size, namely $-\lceil \lg(\varepsilon) \rceil$ bits. We compare it with results for Uniform Leader Election and the analogous remarks for a size of the network given a priori. It worth to mention, that the proof of this fact utilizes Cauchy—Schwarz theorem.

Afterwards we consider descripiton of known algorithms, theirs analysis and propositions of rectifications. In particular we present the state-of-the-art of universal Leader Election algorithms and show its flagrant weaknesses.

Next section generalizes the concept of Leader Green Election algorithm according to geometric distribution for bounded networks. It bases on the algorithm of P.Jacquet. We refine some crucial ideas of LGE algorithm and produce more flexible solution, called Geometric Green Leader Election (shortly GeoGLE). We provide a theorem with extensive formulation, which precisely tunes all the parameters of reliable GeoGLE. A very technical proof of this crucial contribution is postponed to appendix. It utilizes W-Lambert multi-function, several results for MaxGeo counter, a technique from the proof of Poisson's theorem, Leibniz' theorem, generalized Weierstrass' Product Inequality and many subtle inequalities. Further we provide a broad discussion about GeoGLE solution and its effectiveness, together with implementation's details. We show that GeoGLE is the best amongst all known algorithms in the considered class of procedures. We also present a result of a very dilate Monte Carlo simulations of this algorithm, which affirms the quality of the precise choice of parameters.

The latter crucial contribution in this chapter considers a mixture of Uniform Leader Election and GeoGLE algorithms. We show that the reliability of our adaptation of such the solution is monotonic with respect to the number of devices n. It occurs to be almost as good in terms of a total runtime as GeoGLE algorithm, but there remain a lot of space for its rectifications, what is additionally affirmed by simulations. Our solution utilizes some of the theorems and method from the proof for GeoGLE algorithm, together with one clever trick and several standard considerations. A summary and comparison of all the considered Leader Election algorithms are provided in the latter section of the chapter.

Third chapter is devoted to a particular domain of Big Data analysis, namely to Reservoir Sampling algorithms. Introduction describes the essential ideas behind this kind of procedures, including theirs short history, starting from the famous Algorithm R. Next we introduce a concept of sliding window algorithms and provide a general Reservoir Sampling algorithm in a sliding window of the fixed discrete size assumed a priori. A fundamental algorithm bases on devil's staircase Markov chains. Further we present a basic algorithms and a bunch of artful tricks, which improve the efficiency of standard coupling techniques applied to our solution. Finally we show how to extend this procedure to a wide class of distributions, which includes a uniform one. We also analyze several examples of such the generalizations and theirs properties.

The latter section of this chapter provides a Reservoir Sampling algorithm with update probabilities $\min\left(1, \frac{g}{n^\alpha}\right)$, for some constansts g and α and the number of data items already explored n. We itemize diverse laws of convergence, together with asymptotical expected values of the pointers of the data sampled by our algorithm, with respect to cases dependent on g and α. Next 6 consecutive subsections provide analysis for each of the specified cases ($\alpha = 0$, $\alpha \in (0,1)$, $\alpha = 1$, $\alpha \in (1,2)$, $\alpha = 2$, $\alpha > 2$). Each of the cases utilizes different approaches and notions: Fatou's lemma, weak convergence, generalized harmonic numbers, Riemann's ζ function, generalized Weierstrass' Product Inequality, just to mention a few. In result we receive a variety of limiting distributions (e.g. geometric ($\alpha = 0$), exponential ($\alpha \in (0,1)$), Beta ($\alpha = 1$)). Next we present several ideas of applications of this solution with a prominent example involving bitcoin capitalization. In appendix we provide a technical approaches which let us obtain more precise results for expected values of the distributions of pointers for some particular cases of the algorithm. They utilize e.g. Z-transform or formal power series with rational exponents.

The last chapter is devoted to differential privacy of probabilistic counters. First we provide an intuition and definition of a differential privacy and a motivation to analyze probabilistic counters in terms of this property.We introduce Morris and MaxGeo counters and related probabilistic counters like e.g. HyperLogLog. We prove that under certain assumptions about the minimal number of requests, both probabilistic counters evoked before are differentially private by inherent randomization. We show that some of the differential privacy parameters tend to 0 as the number of users grow, what improves the quality of the mechanisms. Gory fragments of the proof for Morris counter' privacy are postponed to appendix (sincerely, a lion's share of this very careful reasoning). We demonstrate how probabilistic counters can be used for constructing a differentially private data aggregation protocol. We also compare our contribution with state-of-the-art Laplace method based solution, which utilizes extra randomization and significantly more memory supplies than Probabilistic Counters. We recall papers related to main topics of this chapter and show some directions of possible future work and applications.

Bojko

„Wybrane zagadnienia z zakresu algorytmów zrandomizowanych" — streszczenie

Dominik Bojko

Niniejsza praca doktorska zbiera wyniki z różnych projektów związanych z teorią prawdopodobieństwa, procesami stochastycznymi i algorytmiką. Jest podzielona na trzy główne gałęzie: Algorytmy Wyboru Lidera, Algorytmy Próbkowania do Rezerwuaru oraz Prywatność Różnicową Liczników Probabilistycznych.

Rozdział pierwszy zawiera zarys pracy, notację i opisy podstawowych pojęć wykorzystywanych w rozprawie.

Drugi poświęcona jest algorytmom Wyboru Lidera. W pierwszej kolejności przedstawiamy motywację tej koncepcji oraz bardzo obszerne opisy możliwych ogólnych założeń takich procedur i ich zwięzłą historię. Następnie prezentujemy model urnowy, uzasadniamy jego użyteczność i podajemy podstawowy blok zrandomizowanego algorytmu Wyboru Lidera zgodny z tymi ramami.

Dalej rozważamy dwie główne aranżacje, a mianowicie procedury przeznaczone dla określonej z góry liczby urządzeń n oraz dla dowolnej liczby podmiotów w sieci, ograniczonej jej pojemnością N.

Pierwszy istotny wkład dotyczy pierwszego z tych podejść. Dla dowolnego skończonego zbioru możliwych identyfikatorów uzyskujemy rozkład optymalny ze względu na prawdopodobieństwo pomyślnego Wyboru Lidera. W tym celu wykorzystujemy kilka nietypowych sztuczek i metodę mnożników Lagrange'a. Obliczenie rozkładu rozwiązania optymalnego dla dużych n okazuje się czasochłonne, stąd też podajemy dokładne przybliżenia optymalnego rozkładu metodą zaproponowaną przez S.Stevića (aproksymującą ciągi rekurencyjne), wraz z twierdzeniem Banacha o punkcie stałym. Ostatnia część tej sekcji ujawnia zdumiewający fakt dotyczący niezbędnych i wystarczających zasobów pamięci (a jednocześnie ograniczeń na czas wykonania procedury), gwarantujący poprawność algorytmu Wyboru Lidera w modelu urnowym dla określonej liczby użytkowników z prawdopodobieństwem co najmniej $1-\varepsilon$. Dokładniej, potrzeba co najmniej $-\lceil \lg(\varepsilon) \rceil$ bitów, a czasem jeden bit więcej, co jest już wystarczającą pamięcią.

Kolejna część jest podstawą reszty pierwszego rozdziału. Wprowadza algorytm Jednostajnego Wyboru Lidera w modelu urnowym. Ponadto dostarcza nowego, ale bardzo naturalny wyniku przy pomocy technik charakterystycznych dla teorii równań różnicowych. Mianowicie zapewniamy monotoniczność prawdopodobieństwa sukcesu w odniesieniu do liczby konkurentów. Następnie dostarczamy nowe, ogólne i optymalne dolne ograniczenie złożoności czasowej algorytmów Wyboru Lidera, które powinny być niezawodne dla dowolnej liczby urządzeń w sieci o ograniczonym rozmiarze, a dokładniej $-\lceil \lg(\varepsilon) \rceil$ bitów. Porównujemy je z wynikami dla Jednostajnego Wyboru Lidera i analogicznych uwag dla przypadku rozmiaru sieci podanej a priori. Warto wspomnieć, że dowód tego faktu wykorzystuje twierdzenie Cauchy'ego–Schwarza.

Następnie rozważamy opis znanych algorytmów, ich analizę i propozycje poprawek. W szczególności przedstawiamy najnowocześniejsze uniwersalne algorytmy Wyboru Lidera i pokazujemy ich rażące słabości.

Następna sekcja uogólnia koncepcję algorytmu Eko-Wyboru Lidera zgodnie z rozkładem geometrycznym dla sieci ograniczonych. Opiera się na algorytmie Philipa Jacqueta. Dopracowujemy kilka kluczowych pomysłów algorytmu EWL i tworzymy bardziej elastyczne rozwiązanie, zwane Geometrycznym Eko-Wyborem Lidera (w skrócie GeoEWL). Proponujemy twierdzenie o rozbudowanym sformułowaniu, które precyzyjnie dostraja wszystkie parametry niezawodnego GeoEWLa. Bardzo techniczny dowód tego kluczowego wkładu do pracy został przestawiony do załącznika. Wykorzystuje multi-funkcję W-Lamberta, kilka wyników dla licznika MaxGeo, technikę z dowodu twierdzenia Poissona, twierdzenie Leibniza, uogólnione nierówności iloczynowe Weierstrassa i wiele subtelnych nierówności. Dalej prezentujemy szeroką dyskusję na temat rozwiązania GeoEWL i jego skuteczności wraz ze szczegółami implementacyjnymi. Pokazujemy, że GeoEWL jest najlepszym spośród wszystkich dotychczas znanych algorytmów w rozważanej klasie procedur. Przedstawiamy również wynik bardzo głębokich symulacji metodą Monte Carlo dla tego algorytmu, który potwierdza jakość precyzyjnie dobranych parametrów.

Ostatni kluczowy wkład w tym rozdziale dotyczy kombinacji algorytmów Jednostajnego Wyboru Lidera i GeoEWLa. Pokazujemy, że niezawodność naszej adaptacji takiego rozwiązania jest monotoniczna w stosunku do liczby urządzeń n. Okazuje się, że to rozwiązanie pod względem całkowitego czasu działania jest niemal tak dobre, jak algorytm GeoEWL, ale wciąż pozostaje wiele przestrzeni na jego doprecyzowanie, co dodatkowo potwierdzają symulacje. Nasze rozwiązanie wykorzystuje niektóre twierdzenia i metody z dowodu dla algorytmu GeoEWL, wraz z jedną sprytną sztuczką i kilkoma standardowymi rozważaniami. Podsumowanie i porównanie wszystkich rozważanych w pracy algorytmów Wyboru Lidera znajdują się w ostatniej części rozdziału.

Rozdział trzeci poświęcony jest konkretnej dziedzinie analizy Big Data, a mianowicie algorytmom Próbkowania do Rezerwuaru. Wstęp opisuje podstawowe idee stojące za tego rodzaju procedurami, w tym ich krótką historię, począwszy od słynnego Algorytmu R. Następnie przedstawiamy pojęcie procedur w oknie przesuwnym oraz ogólny algorytm Próbkowania do Rezerwuaru w oknie przesuwnym o ustalonym, dyskretnym rozmiarze założonym a priori.

Podstawowy protokół opiera się o diabelskie schody Markowa. Dalej przedstawiamy podstawowe algorytmy oraz szereg pomysłowych sztuczek, które poprawiają efektywność standardowych technik couplingowych zastosowanych początkowo w naszym rozwiązaniu. Na koniec pokazujemy, jak rozszerzyć naszą procedurę na szeroką klasę rozkładów, w szczególności na rozkład jednostajny. Analizujemy również kilka przykładów takich uogólnień wraz z ich własnościami.

W drugiej części tego rozdziału przedstawiono algorytm Próbkowania do Rezerwuaru z prawdopodobieństwami aktualizacji $\min\left(1, \frac{g}{n^\alpha}\right)$, dla pewnych stałych g and α oraz liczby danych do tej pory odkrytych n. Wyszczególniamy różne prawa zbieżności, wraz z asymptotycznymi wartościami oczekiwanymi liczników kontrolujących próbkowane dane dla różnych przypadków, zależnych od g i α. Kolejne 6 podrozdziałów pokazują analizę dla każdego z wyszczególnionych przypadków ($\alpha = 0$, $\alpha \in (0,1)$, $\alpha = 1$, $\alpha \in (1,2)$, $\alpha = 2$, $\alpha > 2$). Każdy z nich wykorzystuje różne podejścia i pojęcia: lemat Fatou, słabą zbieżność, uogólnione liczby harmoniczne, funkcję ζ Riemanna, uogólnione nierówności iloczynowe Weierstrassa, żeby wymienić tylko kilka nich. W rezultacie otrzymujemy różne rozkłady graniczne (np. geometryczny ($\alpha = 0$), wykładniczy ($\alpha \in (0,1)$), Beta ($\alpha = 1$)). Następnie przedstawiamy kilka pomysłów na zastosowania tego rozwiązania z wybitnym przykładem kapitalizacji bitcoinów. W załączniku przedstawiamy techniczne rozwiązania, które pozwalają uzyskać dokładniejsze wyniki dla oczekiwanych wartości oczekiwanych rozkładów liczników dla niektórych szczególnych przypadków algorytmu. Wykorzystują m.in. Z-transformatę, czy też formalne szeregi potęgowe z wykładnikami wymiernymi.

Ostatni rozdział poświęcony jest Prywatności Różnicowej Liczników Probabilistycznych. Najpierw podajemy intuicję i definicję Prywatności Różnicowej i motywację stającą za analizą Liczników Probabilistycznych pod kątem tej własności.Przedstawiamy liczniki Morrisa i MaxGeo oraz powiązane z nimi Liczniki Probabilistyczne, jak przykładowo HyperLogLog. Udowadniamy, że przy pewnych założeniach dotyczących minimalnej liczby zapytań, oba wspomniane wcześniej Liczniki Probabilistyczne są Prywatne Różnicowo poprzez ich nieodłączną randomizację. Pokazujemy, że niektóre parametry Prywatności Różnicowej zbiegają do 0 wraz ze wzrostem liczby użytkowników, co poprawia jakość tych mechanizmów. Brutalne fragmenty dowodu dla Prywatności licznika Morrisa zostały przełożone do załącznika (a dokładniej lwia część tego bardzo ostrożnego rozumowania). Przypominamy również niektóre aplikacje licznika MaxGeo, które dzięki naszemu rozwiązaniu mogą zostać w prosty sposób przekształcone na mechanizmy Prywatne Różnicowo. Pokazujemy, w jaki sposób Liczniki Probabilistyczne mogą być używane do skonstruowania Prywatnych Różnicowo protokołów agregacji danych. Porównujemy również nasz wkład z najlepszym znanym rozwiązaniem, opartym na metodzie Laplace'a, które wykorzystuje dodatkową randomizację i znacznie więcej zasobów pamięci aniżeli Liczniki Probabilistyczne. Wspominamy artykuły związane z głównymi tematami tego rozdziału i wskazujemy kierunki możliwych przyszłych prac i zastosowań.

Bojko