Human Genome Center

Division of Medical Data Informatics

医療データ情報学分野

Professor	Tetsuo Shibuya, Ph.D.	教　授	博士(理学)	渋　谷　哲　朗
Assistant Professor	Robert Daniel Barish, Ph.D.	助　教	博士(学術)	ロバート ダニエル バリッシュ

The objective of Division of Medical Data Informatics is to develop fundamental data informatics technologies for medical data, including algorithm theory, big data technologies, artificial intelligence, data mining, and privacy preserving technologies. Medical data, especially genome data are increasing exponentially from basics to clinical research in medical science. Our aim is to innovate medical science with novel data informatics solutions.

1. Development of Privacy Preserving Technologies for Medical Data

a. Privacy-Optimized Randomized Response for Sharing Multi-Attribute Data

Akihito Yamamoto[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

With the increasing amount of data in society, privacy concerns in data sharing have become widely recognized. Particularly, protecting personal attribute information is essential for a wide range of aims from crowdsourcing to realizing personalized medicine. Although various differentially private methods based on randomized response have been proposed for single attribute information or specific analysis purposes such as frequency estimation, there is a lack of studies on the mechanism for sharing individuals' multiple categorical information itself. The existing randomized response for sharing multi-attribute data uses the Kronecker product to perturb each attribute information in turn according to the respective privacy level but achieves only a weak privacy level for the entire dataset. Therefore, in this study, we propose a privacy-optimized randomized response that guarantees the strongest privacy in sharing multi-attribute data. Furthermore, we present an efficient heuristic algorithm for constructing a near-optimal mechanism [1]. The time complexity of our algorithm is $O(k^2)$, where k is the number of attributes, and it can be performed in about 1 second even for large datasets with $k = 1{,}000$. The experimental results demonstrate that both of our methods provide significantly stronger privacy guarantees for the entire dataset than the existing method. Overall, this study is an important step toward trustworthy sharing and analysis of multi-attribute data. In addition, we show an analysis example using genome statistics to confirm the high utility of our method, along with supplemental materials.

b. Generalization and Enhancement of Piecewise Mechanism for Collecting Multidimensional Data

Akihito Yamamoto[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

As the amount of data in society increases, the importance of collecting and storing data while protecting privacy also increases. In particular, protecting personal numeric data is essential for crowdsourcing and big data analytics. Although various methods have been proposed for specific analysis purposes such as mean estimation, methods for storing numeric values themselves are still lacking. Furthermore, no method that can flexibly collect all data information with multiple attributes exists. Therefore, this study first generalizes the piecewise mechanism (PM), the state-of-the-art method in collecting a single numeric value, and proposes a new mechanism that achieves a truly smaller variance of the collected private values than the original one. Subsequently, we enhance our generalized PM for collecting multidimensional numeric data while considering a situation in which each attribute information has its own privacy level [2]. The proposed mechanism is optimal in terms of privacy guarantees for the entire dataset, and is highly advisable for collecting all information with high privacy assurance. We further evaluate our mechanism both theoretically and experimentally and show that it outperforms existing methods. We measure the accuracy of the collected private values using real census data as well, demonstrating the utility of our mechanism.

c. Differentially Private Selection using Smooth Sensitivity

Akihito Yamamoto[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

With the growing volume of data in society, the need for privacy protection in data analysis also rises. In particular, private selection tasks, wherein the most important information is retrieved under differential privacy are emphasized in a wide range of contexts, including machine learning and medical statistical analysis. However, existing mechanisms use global sensitivity, which may add larger amount of perturbation than is necessary. Therefore, this study proposes a novel mechanism for differentially private selection using the concept of smooth sensitivity and presents theoretical proofs of strict privacy guarantees. Simultaneously, given that the current state-of-the-art algorithm using smooth sensitivity is still of limited use, and that the theoretical analysis of the basic properties of the noise distributions are not yet rigorous, we present fundamental theorems to improve upon them. Furthermore, new theorems are proposed for efficient noise generation [3]. Experiments demonstrate that the proposed mechanism can provide higher accuracy than the existing global sensitivity-based methods. Finally, we show key directions for further theoretical development.

d. Cycle Counting under Local Differential Privacy for Degeneracy-bounded Graphs

Quentin Hillebrand[1], Vorapong Suppakitpaisarn[2], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo, [2]Graduate School of Information Science and Technology, The University of Tokyo

We propose an algorithm for counting the number of cycles under local differential privacy for degeneracy-bounded input graphs. Numerous studies have focused on counting the number of triangles under the privacy notion, demonstrating that the expected ℓ_2-error of these algorithms is $\Omega(n^{1.5})$, where n is the number of nodes in the graph. When parameterized by the number of cycles of length four (C_4), the best existing triangle counting algorithm has an error of $\mathrm{O}(\mathrm{n}^{1.5}+\mathrm{C}_4{}^{0.5}) = \mathrm{O}(n^2)$. In this study, we introduce an algorithm with an expected ℓ_2-error of $O(\delta^{1.5}n^{0.5} + \delta^{0.5}d_{\max}{}^{0.5}n^{0.5})$, where δ is the degeneracy and $d_{\max}$ is the maximum degree of the graph. For degeneracy-bounded graphs ($\delta \in \Theta(1)$) commonly found in practical social networks, our algorithm achieves an expected ℓ_2-error of $O(d_{\max}{}^{0.5}n^{0.5}) = O(n)$. Our algorithm's core idea is a precise count of triangles following a preprocessing step that approximately sorts the degree of all nodes. This approach can be extended to approximate the number of cycles of length k, maintaining a similar ℓ_2-error, namely $O(\delta^{(k-2)/2}\, d_{\max}{}^{0.5}n^{(k-2)/2} + \delta^{k/2}n^{(k-2)/2})$ or $O(d_{\max}{}^{0.5}n^{(k-2)/2}) = O(n^{(k-1)/2})$ for degeneracy-bounded graphs.

2. Development of Biomedical Database Technologies

a. KEGG: Biological Systems Database as a Model of the Real World

Mari Ishiguro-Watanabe[1], Minoru Kanehisa[2]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo, [2]Institute for Chemical Research, Kyoto University.

KEGG is a database resource for representation and analysis of biological systems [5]. Pathway maps are the primary dataset in KEGG representing systemic functions of the cell and the organism in terms of molecular interaction and reaction networks. The KEGG Orthology (KO) system is a mechanism for linking genes and proteins to pathway maps and other molecular networks. Each KO is a generic gene identifier and each pathway map is created as a network of KO nodes. This architecture enables KEGG pathway mapping to uncover systemic features from KO assigned genomes and metagenomes. Additional roles of KOs include characterization of conserved genes and conserved units of genes in organism groups, which can be done by taxonomy mapping. A new tool has been developed for identifying conserved gene orders in chromosomes, in which gene orders are treated as sequences of KOs. Furthermore, a new dataset called VOG (virus ortholog group) is computationally generated from virus proteins and expanded to proteins of cellular organisms, allowing gene orders to be compared as VOG sequences as well. Together with these datasets and analysis tools, new types of pathway maps are being developed to present a global view of biological processes involving multiple organism groups.

b. Fair Selection of Clearing Schemes for Kidney Exchange Markets

Robert Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

For the Kidney Exchange Problem (KEP), one has a barter exchange market represented by a digraph with vertices corresponding to either immunologically incompatible donor-acceptor pairs, non-directed donors, cadavers, or unpaired recipients, and directed edges corresponding to possible kidney exchanges. The objective is then to solve the associated clearing problem of finding an above threshold weight partition of the network into vertex-disjoint transplant cycles and/or paths. In this study, with a primary motivation being the broad applicability of the KEP model to barter exchange markets of indivisible goods, we conduct a theoretical investigation of the problem of uniformly, and in this sense "fairly", sampling witnesses for a formalization of the KEP we denote KEP-(Lc, Lp, ϒ), where we have cycle and path vertex-wise length constraints Lc and Lp, respectively, and where we require that the sum of all edge weights in a partition is at least $\Upsilon \in \mathrm{N}_0$ [6]. Here, for KEP- (∞,∞, 0), we provide an $O(4^g \cdot n^4 \cdot m)$ time uniform sampling scheme (assuming access to an idealized coin flipping oracle) for networks on n vertices and m edges admitting bimodal embeddings (i.e., embeddings where each set of edges oriented away from a given vertex occur contiguously in a rotational ordering of edges incident to the vertex) on genus ≤ g surfaces, as well as a Fully Polynomial-time Almost Uniform Sampling (FPAUS) scheme for arbitrary genus digraphs. Subsequently, taking inspiration from recent rapid experimental advances in using boson sampling (respectively, Guassian boson sampling) to approximate the permanents (respectively, hafnians) of complex matrices, we reduce the uniform sampling problem for KEP-(Lc, Lp, ϒ) on networks with n vertices and m edges to calculating at most $O(n \cdot m)$ permanents of hollow Hermitian {−1, 0, 1} matrices. However, we moderate this latter finding by ruling out (unless NP = RP) a Fully Polynomial-time Randomized Approximation Scheme (FPRAS) for the permanent of such matrices.

c. String Editing under Pattern Constraints

Robert Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

We introduce the novel Nearest Pattern Constrained String (NPCS) problem of finding a minimum set Q of character mutation, insertion, and deletion edit operations sufficient to modify a string χ to contain all contiguous substrings in a pattern set P and no contiguous substrings in a forbidden pattern set F. Letting Σ be the alphabet of allowed characters, and letting η and ϒ be the longest string length and sum of all string lengths in $P \cup F$, respectively, we show that NPCS is fixedparameter tractable in $|P|$ with time complexity $O(2^{|P|} \cdot \Upsilon \cdot |\Sigma| \cdot (|P| + \eta) (|\chi|+1))$. Additionally, we consider a generalization of the NPCS problem in which we allow for constraints based on the membership of substrings in regular languages. In particular, we introduce a problem we denote String Editing under Substring in Language Constraints (StrEdit-SILC), where provided a wildcardfree string $\chi \in \Sigma^*$, a finite set of regular languages $R = \{L1, L\,2, \ldots\}$, and a regular language L_F , the objective is to find a minimum cost set of mutation, insertion, and deletion edit operations Q that suffice to convert the input string χ into a string $\chi' \in \Sigma^*$, where no substring has membership in L_F, and $\forall Li \in R$, there exists a substring in Li. Here, letting Ψ and ϖ be the sum of all regular expression lengths and longest regular expression length for languages in $R \cup \{L_F\}$, respectively, and letting $C_{mid} \in \mathbb{N}$ be the maximum cost of an edit operation, we show that StrEdit-SILC is fixed-parameter tractable with respect to Ψ, having time complexity $O(2^{\Psi} \cdot |\chi| \cdot (\varpi \cdot |\Sigma|+C_{mid}))$. However, we also show that StrEdit-SILC is MAX-SNP-hard and otherwise difficult to approximate under stringent constraints.

d. Affine Optimal *k*-proper Connected Edge Colorings

Robert Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

We introduce *affine optimal k-proper connected edge colorings* as a variation on Fujita's notion of *optimal k-proper connected colorings* with applications to the frequency assignment problem [8]. Here, for a simple undirected graph G with edge set E_G , such a coloring corresponds to a decomposition of E_G into color classes $C_1, C_2, \ldots, C_n$, with associated weights $w_1, w_2, \ldots, w_n$, minimizing a specified affine function $A{:=}\Sigma w_i \cdot |C_i|$, while also ensuring the existence of k vertex disjoint *proper paths* (i.e., simple paths with no two adjacent edges in the same color class) between all pairs of vertices. In this context, we define $\zeta^k(A, G)$ as the minimum possible value of A under a k-proper connectivity requirement. For any fixed number of color classes, we show that computing $\zeta^k(A', G)$ is treewidth fixed parameter tractable. However, we also show that determining $\zeta^k(A', G)$ with the affine function $A' := |C_2|$ is NP-hard for 2-connected planar graphs in the case where $k = 1$, cubic 3-connected planar graphs for $k = 2$, and k-connected graphs $\forall k \geq 3$. We also show that no fully polynomial-time randomized approximation scheme can exist for approximating $\zeta^k(A', G)$ under any of the aforementioned constraints unless $NP = RP$.

e. Proper Colorability of Segment Intersection Graphs

Robert Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

We consider the vertex proper coloring problem for highly restricted instances of geometric intersection graphs of line segments embedded in the plane [9]. Provided a graph in the class UNIT-PURE-k-DIR, corresponding to intersection graphs of unit length segments lying in at most k directions with all parallel segments disjoint, and provided explicit coordinates for segments whose intersections induce the graph, we show for $k = 4$ that it is NP-complete to decide if a proper 3-coloring exists, and moreover, $\#P$-complete under many-one counting reductions to determine the number of such colorings. In addition, under the more relaxed constraint that segments have at most two distinct lengths, we show these same hardness results hold for finding and counting proper (k−1)-colorings for every $k \geq 5$. More generally, we establish that the problem of proper 3-coloring an arbitrary graph with m edges can be reduced in $O(m^2)$ time to the problem of proper 3-coloring a UNIT-PURE-4-DIR graph. This can then be shown to imply that no $2^{o(\sqrt{n})}$ time algorithm can exist for proper 3-coloring PURE-4-DIR graphs under the Exponential Time Hypothesis (ETH), and by a slightly more elaborate construction, that no $2^{o(\sqrt{n})}$ time algorithm can exist for counting the such colorings under the Counting Exponential Time Hypothesis (#ETH). Finally, we prove an NP-hardness result for the optimization problem of finding a maximum order proper 3-colorable induced subgraph of a UNIT-PURE-4-DIR graph.

f. Counting on Rainbow k-Connections

Robert Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

For an undirected graph imbued with an edge coloring, a rainbow path (resp. proper path) between a pair of vertices corresponds to a simple path in which no two edges (resp. no two adjacent edges) are of the same color. In this context, we refer to such an edge coloring as a rainbow k-connected w-coloring (resp. k-proper connected w-coloring) if at most w colors are used to ensure the existence of at least k internally vertex disjoint rainbow paths (resp. k internally vertex disjoint proper paths) between all pairs of vertices. At present, while there have been extensive efforts to characterize the complexity of finding rainbow 1-connected colorings, we remark that very little appears to known for cases where k >1. In this work, we first show that the problems of counting rainbow k-connected w-colorings and counting k-proper connected w-colorings are both linear time treewidth Fixed Parameter Tractable (FPT) for every k>0 and w>0. Subsequently, and in the other direction, we extend prior NP-completeness results for deciding the existence of a rainbow 1-connected w-coloring for every w>1, in particular, showing that the problem remains NP-complete for every k>0 and w>1. This yields as a corollary that no Fully Polynomial-time Randomized Approximation Scheme (FPRAS) can exist for approximately counting such colorings in any of these cases (unless NP = RP). Next, concerning counting hardness, we give the first #P-completeness result we are aware of for rainbow connected colorings, proving that counting rainbow k-connected 2-colorings is #P-complete for every k>0 [10].

g. Counting 2-Factors of 4-Regular Bipartite Graphs

Robert Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

We prove that counting 2-factors (i.e., spanning 2-regular subgraphs or vertex disjoint cycle covers) of 4-regular bipartite graphs is #P-complete under many-one counting (i.e., weakly parsimonious) reductions. This resolves a missing case in a proof by Felsner et al. that counting 2-factors of k-regular bipartite graphs is #P-complete for cases k >5 and k=3. Due to a bijective correspondence, it establishes the same hardness result for counting the Eulerian orientations of 4-regular bipartite graphs [11].

h. Recognition and Proper Coloring of Unit Segment Intersection Graphs

Robert Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

In this work, we concern ourselves with the fine-grained complexity of recognition and proper coloring problems on highly restricted classes of geometric intersection graphs of “thin” objects (i.e., objects with unbounded aspect ratios) [12]. As a point of motivation, we remark that there has been significant interest in finding algorithmic lower bounds for classic decision and optimization problems on these types of graphs, as they appear to escape the net of known planar or geometric separator theorems for “fat” objects (i.e., objects with bounded aspect ratios). In particular, letting n be the order of a geometric intersection graph, and assuming a geometric ply bound, per what is known as the "square root phenomenon", these separator theorems often imply the existence of $2^{o(\sqrt{n})}$ algorithms for problems ranging from finding proper colorings to finding Hamiltonian cycles. However, in contrast, it is known for instance that no $2^{o(\sqrt{n})}$ time algorithm can exist under the Exponential Time Hypothesis (ETH) for proper 6-coloring intersection graphs of line segments embedded in the plane. We begin by establishing algorithmic lower bounds for proper *k*-coloring and recognition problems of intersection graphs of line segments embedded in the plane under the most stringent constraints possible that allow either problem to be non-trivial. In particular, we consider the class UNIT PURE-*k*-DIR of unit segment geometric intersection graphs, in which segments are constrained to lie in at most *k* directions in the plane, and no two parallel segments are permitted to intersect.

Here, under the ETH, we show for every $k \geq 3$ that no $2^{o(\sqrt{n/k})}$ time algorithm can exist for either recognizing or proper *k*-coloring UNIT-PURE-k-DIR graphs of order n. In addition, for every $k \geq 4$, we establish the same algorithmic lower bound under the ETH for the problem of proper $(k-1)$-coloring UNIT-PURE-k-DIR graphs when provided a list of segment coordinates specified using $O(n \cdot k)$ bits witnessing graph class membership. As a consequence of our approach, we are also able to show that the problem of properly 3-coloring an arbitrary graph on m edges can be reduced in $O(m^2)$ time to the problem of properly $(k - 1)$-coloring a UNIT-PURE-k-DIR graph. Finally, we consider a slightly less constrained class of geometric intersection graphs of lines (of unbounded length) in which line-line intersections must occur on any one of $(r = 3)$ parallel planes in R3. In this context, for every $k \geq 3$, we show that no $2^{o(n/k)}$ time algorithm can exist for proper *k*-coloring these gr*aphs unless the ETH is false.*

i. Polyhedral roll-connected colorings of partial tiling

Robert Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

We consider the problem of coloring the faces of an edge-to-edge partial tiling *T* such that a specified face-colored polyhedron *P* "rolling" over this tiling – where each time a face of the polyhedron is congruent with a polygonal tile in *T*, both the face and the tile must have the same coloration – is able to reach any tile from any other tile [13]. Here, for *P* corresponding to any Platonic solid, we show that the existence of such a coloring, using at most $w \geq 1$ distinct colors, can be decided in $O(T)$ time. On the other hand, when we require at least two internally disjoint manners of rolling from any starting location to any ending location, we show for *P* corresponding to the cube and for $w = 3$ that deciding the existence of and counting such colorings becomes NP-hard and #P-hard, respectively.

3. Survey on Cutting-Edge Quantum Machine Learning Technologies

Yaswita Gujju[1], Atsushi Matsuo[2], Rudy Raymond[3,4,5]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo, [2]IBM Quantum, IBM Research, [3]Department of Computer Science, The University of Tokyo, [4]Global Technology and Applied Research, J. P. Morgan Chase & Co., [5]Quantum Computing Center, Keio University

The past decade has witnessed significant advancements in quantum hardware, encompassing improvements in speed, qubit quantity, and quantum volume—a metric defining the maximum size of a quantum circuit effectively implementable on near-term quantum devices. This progress has led to a surge in quantum machine learning (QML) applications on real hardware, aiming to achieve quantum advantage over classical approaches. Our survey [14] focuses on selected supervised and unsupervised learning applications executed on quantum hardware, specifically tailored for real-world scenarios. The exploration includes a thorough analysis of current QML implementation limitations on quantum hardware, covering techniques like encoding, ansatz structure, error mitigation, and gradient methods to address these challenges. Furthermore, the survey evaluates the performance of QML implementations in comparison to classical counterparts. In conclusion, we discuss existing bottlenecks related to applying QML on real quantum devices and propose potential solutions to overcome these challenges in the future.

Publications

1. Akito Yamamoto and Tetsuo Shibuya, "Privacy-Optimized Randomized Response for Sharing Multi-Attribute Data", *Proc. 29th IEEE Symposium on Computers and Communications*, 2024, pp.1-8.
2. Akito Yamamoto and Tetsuo Shibuya, "Generalization and Enhancement of Piecewise Mechanism for Collecting Multidimensional Data", *Proc. 17th IEEE International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*, in press.

3. Akito Yamamoto and Tetsuo Shibuya, "Differentially Private Selection using Smooth Sensitivity", *Proc. 43rd IEEE International Performance, Computing and Communications Conference*, in press.
4. Quentin Hillebrand, Vorapong Suppakitpaisarn, and Tetsuo Shibuya, "Cycle Counting under Local Differential Privacy for Degeneracy-bounded Graphs", *Proc. the 42nd International Symposium on Theoretical Aspects of Computer Science*, in press.
5. Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuua, Mari Ishiguro-Watanabe, "KEGG: biological systems database as a model of the real world", *Nucleic Acids Research*, 2024, gkae909.
6. Robert Barish, Tetsuo Shibuya, "Fair Selection of Clearing Schemes for Kidney Exchange Markets", *Proc. 17th International Conference on Combinatorial Optimization and Applications*, in press.
7. Robert Barish, Tetsuo Shibuya, "String editing under pattern constraints", Theoretical Computer Science, *Theoretical Computer Science*, 1022, 2024, 114889, ISSN 0304-3975.
8. Robert Barish, Tetsuo Shibuya, "Affine optimal k-proper connected edge colorings", *Optimization Letters*, S11590-024-021, Springer, 2024.
9. Robert Barish, Tetsuo Shibuya, "Proper colorability of segment intersection graphs", *Journal of Combinatorial Optimization*, 47(4), 2024.
10. Robert Barish, Tetsuo Shibuya, "Counting on rainbow *k*-connections", In: Chen, X., Li, B. (eds) Theory and Applications of Models of Computation (TAMC 2024), *Lecture Notes in Computer Science*, vol. 14637, Springer, 2024, pp 272–283.
11. Robert Barish, Tetsuo Shibuya, "Counting 2-factors of 4-regular bipartite graphs is #P-complete", *Lecture Notes in Computer Science*, Springer, in press.
12. Robert D. Barish and Tetsuo Shibuya. "Recognition and Proper Coloring of Unit Segment Intersection Graphs". *Proc. 19th Scandinavian Symposium and Workshops on Algorithm Theory*. Leibniz International Proceedings in Informatics (LIPIcs), 294, , 2024, pp. 5:1-5:19.
13. Robert Barish, Tetsuo Shibuya, "Polyhedral roll-connected colorings of partial tilings", *Proc. Canadian Conference on Computational Geometry*, July 17-19, 2024, pp. 317-324.
14. Yaswitha Gujju, Atsushi Matsuo, Rudy Raymond, "Quantum Machine Learning on Near-Term Quantum Devices: Current State of Supervised and Unsupervised Techniques for Real-World Applications", *Phys. Rev. Applied,* 21, 067001.