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. (ε, κ)-Randomized Anonymization

Akihito Yamamoto[1], Eizen Kimura[2], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo, [2]Medical school of Ehime University

As the amount of biomedical and healthcare data increases, data mining for medicine becomes more and more important for health improvement. At the same time, privacy concerns in data utilization have also been growing. The key concepts for privacy protection are *k*-anonymity and differential privacy, but *k*-anonymity alone cannot protect personal presence information, and differential privacy alone would leak the identity. To promote data sharing throughout the world, universal methods to release the entire data while satisfying both concepts are required, but such a method does not yet exist. Therefore, we propose a novel privacy-preserving method, (ε, *k*)-Randomized Anonymization. In this study, we first present two methods that compose the Randomized Anonymization method [1]. They perform *k*-anonymization and randomized response in sequence and have adequate randomness and high privacy guarantees, respectively. Then, we show the algorithm for (ε, *k*)-Randomized Anonymization, which can provide highly accurate outputs with both *k*-anonymity and differential privacy. In addition, we describe the analysis procedures for each method using an inverse matrix and expectation-maximization (EM) algorithm. In the experiments, we used real data to evaluate our methods' anonymity, privacy level, and accuracy. Furthermore, we show several examples of analysis results to demonstrate high utility of the proposed methods.

b. Differentially Private SNPs Ranking Publication of GWAS

i) Compressive Mechanism-based Method with Haar Wavelet Transform

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

To promote the use of personal genome information in medicine, it is important to analyze the relationship between diseases and the human genomes. Therefore, statistical analysis using genomic data is often conducted, but there is a privacy concern with respect to releasing the statistics as they are. Existing

methods to address this problem using the concept of differential privacy cannot provide accurate outputs under strong privacy guarantees, making them less practical. In this study, for the first time, we investigate the application of a compressive mechanism to genomic statistical data and propose two approaches [2]. The first is to apply the normal compressive mechanism to the statistics vector along with an algorithm to determine the number of nonzero entries in a sparse representation. The second is to alter the mechanism based on the data, aiming to release significant single nucleotide polymorphisms with a high probability. In this algorithm, we apply the compressive mechanism with the input as a sparse vector for significant data and the Laplace mechanism for nonsignificant data. By using the Haar wavelet transform for the compressive mechanism, we can determine the number of nonzero elements and the amount of noise. In addition, we give theoretical guarantees that our proposed methods achieve ε-differential privacy. We evaluated our methods in terms of accuracy and rank error compared with the Laplace and exponential mechanisms. The results show that our second method in particular can guarantee high privacy assurance as well as utility.

ii) Enhancement via a Joint Permutate-and-Flip

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

Owing to an increase in the amount of biomedical and healthcare data, privacy concerns regarding the use of genomic data have become well-recognized. Specifically, it is essential to develop personalized medicine to extract significant loci associated with diseases through large-scale genomic statistical analyses while protecting privacy. Although there are several differentially private methods for this purpose, they are too computationally complex to achieve high accuracy, and there is room for improvement in terms of the output error. In this study, we propose a novel mechanism, Joint Permute-and-Flip, that can provide higher-quality outputs than state-of-the-art techniques for top-K selection [8]. We also present an efficient algorithm that can perform Joint Permute-and-Flip in $O(m \log m)$ time when the dataset contains m elements, making it applicable even to large-scale analyses involving 106 elements. Additionally, we propose new score functions suitable for genomic statistical analysis that can be expressed as a single equation and achieve high accuracy. This is expected to facilitate the construction of accurate and efficient scores for a wider variety of genome statistics. Experimental results demonstrate that our Joint Permute-and-Flip method outperforms existing methods in terms of both accuracy and rank error and requires only half the run time of the exponential mechanism.

c. Differentially Private Publication of GWAS Statistics

i) Publication via Local Differential Privacy

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

As the amount of personal genomic information and privacy concerns in data publication have been growing, several studies have pointed out that the presence information of a particular individual could be revealed from the statistics obtained in large-scale genomic analyses. Existing methods for releasing genome statistics under differential privacy do not prevent the leakage of personal information by untrusted data collectors. In addition, the existing studies for statistical tests using a contingency table had restrictions on the number of cases and controls. Moreover, the methods for correcting for population stratification cannot protect genotype information. Thus, developing a more general and stronger method is desired. In this study, we present privacy-preserving methods for releasing key genome statistics [6]. Our methods enhance the randomized response technique and guarantee individuals' privacy, even when untrusted data collectors exist. Moreover, our methods do not require any restrictions on the contingency tables, and they also guarantee the privacy of targeted genotype information for the analyses to correct for population stratification. The experimental results indicate that our methods can achieve comparable high accuracy to existing methods while preserving privacy more strictly from any data collectors. Furthermore, for statistical analysis using a contingency table, we consider the case where different privacy budgets are assigned to each of the row and column information, and present optimal methods in terms of privacy assurance for the entire table that outperform the existing method. Overall, this study is the first step toward genomic statistical analysis under local differential privacy.

ii) Publication 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 recent increase in the medical data and health awareness, the use of genomic data to promote personalized medicine has been widely considered. Simultaneously, privacy concerns have arisen with the publication of statistics obtained from large-scale genomic statistical analysis such as GWAS. All existing differentially private methods for GWAS statistics protect privacy by adding noise based on global sensitivity, considering the worst-case scenario of possi-

ble datasets. However, the amount of noise required in practical cases is considerably smaller, and these methods do not achieve the desired accuracy in private statistics. In this study, we propose a privacy-preserving method for publishing much more accurate statistics using smooth sensitivity, which generates tailored noise for each dataset [7]. We first introduce a more rigorous theorem on the properties of the noise distribution than was known previously and propose a new ϵ-differentially private method for publishing GWAS statistics. We also provide theoretical proof of the privacy guarantee. Thereafter, we present novel theorems for computing the smooth sensitivity significantly faster than conventional approaches. This enables the application of smooth sensitivity to GWAS statistics, which would otherwise be impossible because of the exceedingly high computational complexity. Based on these theorems, we performed detailed analyses of key GWAS statistics and developed efficient algorithms to obtain their smooth sensitivities. Experimental results demonstrate that our proposed methods achieve at least 3 times higher accuracy than existing global sensitivity-based methods. Furthermore, the execution time is sufficiently short, and the accuracy increases when the dataset becomes larger, suggesting that our methods are suitable for the publication of statistics in large-scale analysis. Because our method is expected to be applicable to other general statistics, this study is an important step toward highly accurate statistical analysis using smooth sensitivity.

d. Differential Private Publication of Graph Properties

i) Reducing Communication Cost for Publishing Differentially Private Subgraph Counting

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 suggest the use of hash functions to cut down the communication costs when counting subgraphs under edge local differential privacy [5]. While various algorithms exist for computing graph statistics, including the count of subgraphs, under the edge local differential privacy, many suffer with high communication costs, making them less efficient for large graphs. Though data compression is a typical approach in differential privacy, its application in local differential privacy requires a form of compression that every node can reproduce. In our study, we introduce linear congruence hashing. With a sampling rate of s, our method can cut communication costs by a factor of s^2, albeit at the cost of increasing variance in the published graph statistic by a factor of s. The experimental results indicate that, when matched for communication costs, our method achieves a reduction in the l_2-error for triangle counts by up to 1000 times compared to the performance of leading algorithms.

ii) Hardness of Bounding Influence via Graph Modification

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 minimally modifying graphs and digraphs by way of exclusively deleting vertices, exclusively deleting edges, or exclusively adding new edges, with or without connectivity constraints for the resulting graph or digraph, to ensure that centrality-based influence scores of all vertices satisfy either a specified lowerbound or upperbound [3]. Here, we classify the hardness of exactly or approximately solving this problem for: (1) all vertex- and edge-deletion cases for betweenness, harmonic, degree, and in-degree centralities; (2) all vertex deletion cases for eigenvector, Katz, and PageRank centralities; (3) all edge-deletion cases for eigenvector, Katz, and PageRank centralities under a connectivity and weak-connectivity constraint; and (4) a set of edge-addition cases for harmonic, degree, and in-degree centralities. We show that some of our results, in particular those for eigenvector, Katz, and PageRank centralities, hold for planar or planar subcubic classes of graphs and digraphs. Finally, under a variety of constraints, we establish that no polynomial time constant factor approximation algorithm can exist for computing the cardinality of a minimum set of vertices or minimum set of edges whose deletion ensures a lowerbound betweenness centrality score, or a lower- or upperbound eigenvector, Katz, or PageRank centrality score unless $P = NP$.

iii) Hardness of Counting Proper Connected Colorings

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

A k-proper connected 2-coloring for a graph is an edge bipartition which ensures the existence of at least k vertex disjoint simple alternating paths (i.e., paths where no two adjacent edges belong to the same partition) between all pairs of vertices. In this work, for every positive integer k, we show that exactly counting such colorings is #P-hard under many-one counting reductions, as well as #P-complete under many-one counting reductions when $k = 1$. Furthermore, for every positive integer k and every $\varepsilon > 0$, we show that the worst case asymptotic running

time for any algorithm approximating the number of k-proper connected 2-colorings for an order n graph within a multiplicative factor of $1 + \varepsilon$ must be at least the worst case asymptotic running time of an algorithm approximating an n-variable instance of #3-SAT within the same multiplicative factor. Here, assuming the Exponential Time Hypothesis (ETH), for every positive integer k and every $\varepsilon > 0$, we are able to rule out the existence of a $2^{o(nk)}/\varepsilon^2$ algorithm for approximating the number of k-proper connected 2-colorings of an order n graph within a factor of $1 + \varepsilon$. In addition, for every positive integer k, we rule out the existence of a $2^{o(nk)}$ time algorithm for finding a k-proper connected 2-coloring of an order n graph under the ETH, or for exactly counting such colorings assuming the moderated Counting Exponential Time Hypothesis (#ETH) [9].

2. Development of Artificial Intelligence Technologies for Biomedical Research

a. Feature Selection for Cancer Classification

Zixuan Wang[1], Yi Zhou[2], Tatsuya Takagi[3], Jiangning Song[4], Yu-Shi Tian[3] and Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo, [2]Beijing International Center for Mathematical Research, Peking University, [3]Graduate School of Pharmaceutical Sciences, Osaka University, [4]Biomedicine Discovery Institute and Monash Data Futures Institute, Monash University

Microarray data have been widely utilized for cancer classification. The main characteristic of microarray data is "large p and small n" in that data contain a small number of subjects but a large number of genes. It may affect the validity of the classification. Thus, there is a pressing demand of techniques able to select genes relevant to cancer classification. This study proposed a novel feature (gene) selection method, Iso-GA, for cancer classification [10]. Iso-GA hybrids the manifold learning algorithm, Isomap, in the genetic algorithm (GA) to account for the latent nonlinear structure of the gene expression in the microarray data. The Davies–Bouldin index is adopted to evaluate the candidate solutions in Isomap and to avoid the classifier dependency problem. Additionally, a probability-based framework is introduced to reduce the possibility of genes being randomly selected by GA. The performance of Iso-GA was evaluated on eight benchmark microarray datasets of cancers. Iso-GA outperformed other benchmarking gene selection methods, leading to good classification accuracy with fewer critical genes selected. The proposed Iso-GA method can effectively select fewer but critical genes from microarray data to achieve competitive classification performance.

b. KEGG for Taxonomy-based Analysis of Pathways and Genomes

Minoru Kanehisa[1], Miho Furumichi[1], Yoko Sato[2], Masayuki Kawashima[3] and Mari Ishiguro-Watanabe[4]: [1]Institute for Chemical Research, Kyoto University, [2]Digital Lab Division, Fujitsu Limited, [3]Network Support Co. Ltd, [4]Division of Medical Data Informatics, Institute of Medical Science, The University of Tokyo

KEGG is a manually curated database resource integrating various biological objects categorized into systems, genomic, chemical and health information. Each object (database entry) is identified by the KEGG identifier (kid), which generally takes the form of a prefix followed by a five-digit number, and can be retrieved by appending /entry/kid in the URL. The KEGG pathway map viewer, the Brite hierarchy viewer and the newly released KEGG genome browser can be launched by appending /pathway/kid, /brite/kid and /genome/kid, respectively, in the URL. Together with an improved annotation procedure for KO (KEGG Orthology) assignment, an increasing number of eukaryotic genomes have been included in KEGG for better representation of organisms in the taxonomic tree. Multiple taxonomy files are generated for classification of KEGG organisms and viruses, and the Brite hierarchy viewer is used for taxonomy mapping, a variant of Brite mapping in the new KEGG Mapper suite. The taxonomy mapping enables analysis of, for example, how functional links of genes in the pathway and physical links of genes on the chromosome are conserved among organism groups.

Publications

1. Yamamoto, A., Kimura, E. and Shibuya, T. (ε, k)-Randomized Anonymization: ε-Differentially Private Data Sharing with k-Anonymity, *Proc. HEALTHINF*. 287-297, 2023.
2. Yamamoto, A., Shibuya, T. Privacy-Preserving Statistical Analysis of Genomic Data using Compressive Mechanism with Haar Wavelet Transform, *J. Comput. Biol.* 30(2):176-188, 2023.
3. Barish, R., Shibuya, T., Hardness of bounding influence via graph modification, *LNCS*. 13878:129-143, 2023.
4. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes, *NAR*. 51(D1), D587–D592, 2023.
5. Hillebrand, Q., Suppakitpaisarn, V., and Shibuya,

T., Unbiased locally private estimator for polynomials of Laplacian variables, *Proc. KDD*. 741–751, 2023.
6. Yamamoto, A. and Shibuya, T., Privacy-Preserving Genomic Statistical Analysis Under Local Differential Privacy, *LNCS*:13942:40-48, 2023.
7. Yamamoto, A. and Shibuya, T., Privacy-Preserving Publication of GWAS Statistics using Smooth Sensitivity, *Proc. PST*. 1-12. 2023.
8. Yamamoto, A. and Shibuya, T., A Joint Permute-and-Flip and Its Enhancement for Large-Scale Genomic Statistical Analysis, *Proc. IEEE ICDMW/ TrustKDD*. In press.
9. Barish, R., Shibuya, T., The Fine-Grained Complexity of Approximately Counting Proper Connected Colorings. *LNCS*. 123-136. 14462. 2023.
10. Wang, Z., Zhou, Y., Takagi, T., Song, J., Tian, Y-S., Shibuya, T., Genetic Algorithm-Based Feature Selection with Manifold Learning for Cancer Classification using Microarray Data, *BMC Bioinformatics*, 24(139), 2023.
11. Barish, R.: On the number of *k*-proper connected edge and vertex colorings of graphs. *Thai. J. Math*. In press.