Poster: A Privacy Risk Arising from Communication with TV: Consideration from Attribute Inference for Users

Takahiro Higuchi*, Naoto Yanai*, Kensuke Ueda†, Yasunori Ishihara‡and, Toru Fujiwara*
*Osaka University
†Mitsubishi Electric Corporation
‡Nanzan University

***Abstract*—TV is able to provide users with various services. However, sensitive information about users may be revealed via communication between the users and TV. In this paper, we investigate privacy risks through communication with TV. In particular, by utilizing the k-nearest neighbor algorithm with the Mahalanobis' distance, we infer ages and genders of users from their viewed TV programs. Besides, we also discuss the available data source, which is suitable for the inference, via a questionnaire survey about actually viewed TV shows with about 1,100 users. Furthermore, we conduct an empirical study about the inference. As a result, the age of the users for five ages and the gender of the users can be discriminated with up to 0.541 accuracy and 0.756 accuracy, respectively.**

I. Introduction

TV has been widely used since the TV can produce attractive services for users. However, sensitive information about users, e.g., taste and gender, may be revealed via communication between the users and TV. For example, while one of the services is to enable a user to operate TV from anywhere in his/her house, information such as a channel and a time stamp is sent in communication on that service. We found TVs which were sending these information without encryption, and confirmed that the viewed TV programs by users can be identified via the information.

In Japan, most people have viewed TV programs via terrestrial broadcasts. The terrestrial broadcasts consist of about ten channels, and their TV programs are broadcasted according to a schedule day and time. Since even metadata of smartphones [1] or room climate data [2] may violate privacy of users, viewing information of TV may be sensitive information.

In this work, we investigate privacy risks arising from communication with TVs. More specifically, we address inference about ages and genders via information of the viewed TV programs by users. In doing so, data, whose variance is large between ages and small in a single age, is necessary. Hence, we also discuss the available data source, which is suitable for the inference via a questionnaire survey about actually viewed TV programs with about 1,100 users. Furthermore we conduct an empirical study about the inference.

II. Inference Algorithm about Ages

We utilize the k-nearest neighbor (k-NN) algorithm as an inference algorithm for ages. Here, we define inference of ages as identifying an age of a user from information of his/her viewed TV programs. In doing so, statistical data with respect to viewing information for each age is crucial. More precisely, as statistical data for each age, a dataset, whose variance between ages is large but that in a single age is small, is desirable. We then utilize the Mahalanobis' distance for the k-NN algorithm. whereas an effect of a direction with a large variance is small for the Mahalanobis' distance, a difference from an average value on a direction with a small variance affect the Mahalanobis' distance strongly. Therefore, the Mahalanobis' distance is suitable for the dataset described above. We infer ages of users among such a dataset by utilizing the k-NN algorithm described above. We also infer genders of users in a similar way.

III. Generation of Statistical Data

To the best of our knowledge, a statistical dataset described in Section II has been unreleased. We hence conduct a questionnaire survey to create statistical data. The questionnaire survey was conducted by targeting males and females with the ages of 18 or over via a smart-phone application named *SmartAnswer* (Japanese only. https://smartanswer.colopl-research.jp/) at 2018/10/20. The main question is that the respondents choose the actually viewed TV programs, including content-sharing services such as YouTube, from the given list. The list consists of 102 broadcast TV programs in Japan during 2018/09/20-2018/10/20, and these TV programs are classified into four categories, i.e., dramas, animations, sports, and variety shows. In this work, based on attribute information provided by Google, ages are classified into five classes of ”18-24”, ”25-34”, ”35-44”, ”45-54”, and ”55 and over”, and genders are two classes of ”male” and ”female”. We have obtained about 110 answers for each gender and each age, i.e., totally 1,101 answers as the questionnaire results.

NDSS '16, 21-24 February 2016, San Diego, CA, USA
Copyright 2016 Internet Society, ISBN 1-891562-41-X
http://dx.doi.org/10.14722/ndss.2016.23xxx

Next, a list of TV programs viewed by some of ages is chosen from the questionnaire results. Firstly, TV programs, such that the number of viewers, i.e., the respondents, is many, are picked up of the 102 TV programs. Secondly, respondents, such that the number of their viewed TV programs in the picked-up programs are many, are picked up of the all respondents. Thirdly, for all the programs picked up in the first step, a total viewing rate for all of the respondents picked up in the second step and a viewing rate for each age of the respondents picked up in the second step are computed, respectively. Finally, a feature value for each program picked up in the first step is computed, and then let TV programs whose feature values are large be *TV programs suitable for inference of ages*. Here, a feature value is computed by sum of squared deviations between the total viewing rate obtained in the third step and the viewing rate for each age, where only the positive deviations are utilized. Similarly, *TV programs suitable for inference of genders* can be also computed.

Finally, for each respondent, vectors in a space of dimensions, i.e., the number of the chosen TV programs, are generated from information of the viewed TV programs as follows: for each TV program, set 1 if the respondent viewed the program and 0 otherwise. Moreover, variance-covariance matrices and mean vectors are generated as statistical data.

IV. Evaluation of Inference Algorithm

We infer ages with the statistical data generated in the previous section, and then evaluate accuracy of the inference. Here, accuracy is a value whereby the maximized value is 1, and a higher value means a better accuracy. Similarly, we also infer genders and evaluate its accuracy.

1) Evaluation of Inference about Ages: In inference about ages, 18 TV programs are chosen as TV programs suitable for inference of ages. First, the respondents are classified into test data, data for statistical data, and removed data. For the data for statistical data, a set of the respondents for each age is dually partitioned in a random way. On the other hand, the removed data is respondents who do not view the chosen TV programs. The removed data is excluded from the following experiments, and the remaining 500 respondents are utilized for the experiments.

Next, the Mahalanobis distances between the test data and each dually partitioned set for any ages are computed. As a result, k ages are obtained from the nearest Mahalanobis distances, and then an attribute as the inference result is determined by majority voting of the k attributes.

Finally, to evaluate the accuracy, we score as follows: 1 point if the result is identical to an actual age; 0.5 point if the result is anteroposterior ages of an actual age, e.g., ”18-24” or ”35-44” for ”25-34”and, 0 point otherwise.

For all combinations with respect to all test data, the above steps are executed and then a summation of scores for the test data is divided by the number of the test data. Let the division result be accuracy of inference about ages. We evaluate the accuracy described above with a cross-validation method by utilizing 50 respondents for each age, i.e., totally 250 respondents, as test data and about 250 respondents as data for statistical data. In particular, 10 respondents for each age, i.e., totally 50 respondents, are randomly chosen as test data, and the remaining 200 test data is utilized as a part of data for statistical data. That is, by utilizing 450 respondents as data for statistical data, ages for the 50 respondents are inferred. The above steps are iterated five times by changing 50 respondents chosen from each age, and a mean value of the resultant accuracies is computed. Likewise, we also generate statistical data by each of all the subsets except for empty sets, and infer ages. We also evaluate the accuracies.

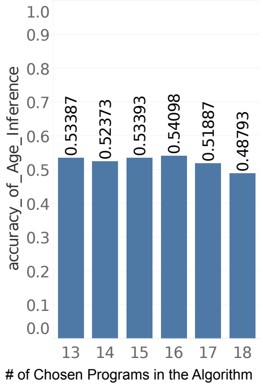


Fig. 1. Accuracy_of_Age

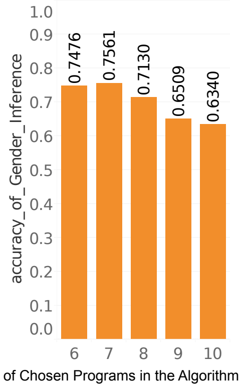


Fig. 2. Accuracy_of_Gender

We show the result in Fig. 1. In the results, the best accuracy is 0.541, which is obtained with 16 programs.

2) Evaluation of Inference about Genders: We also evaluate inference about genders in a similar way to Section IV-1. In inference with respect to genders, 10 TV programs are chosen as TV programs suitable for inference of genders. In this section, we score as follows: 1 point if the result is identical to an actual gender; and, 0 point otherwise. We show the result in Fig. 2. In the results, the best accuracy is 0.756, which is obtained with 7 programs.

V. Conclusion

In this work, we showed that ages and genders of users can be inferred by the k-NN algorithm with the Mahalanobis' distances from information of the viewed TV programs. We also discussed how to generate statistical data suitable for inference, and evaluated accuracies of the inference. We plan to conduct an experiment utilizing actual packets sent from TVs as test data.

Acknowledgment: This research was supported by Mitsubishi Electric Cybersecurity Research Alliance Laboratories. We would like to appreciate Manami Inoue from COLOPL,Inc for her variable comments.

References

[1] J. Mayer, P. Mutchler, and J. C. Mitchell, “Evaluating the Privacy Properties of Telephone Metada”, Proc. of NAS 2016, pp.5536-5541, 2016.

[2] P. Morgner, and C. Müller, and M. Ring, and B. Eskofier, and C. Riess, and F. Armknecht, and Z. Benenson, “Privacy Implications of Room Climate Data”, Proc. of ESORICS 2017, LNCS 10493, pp 324-343, 2017.

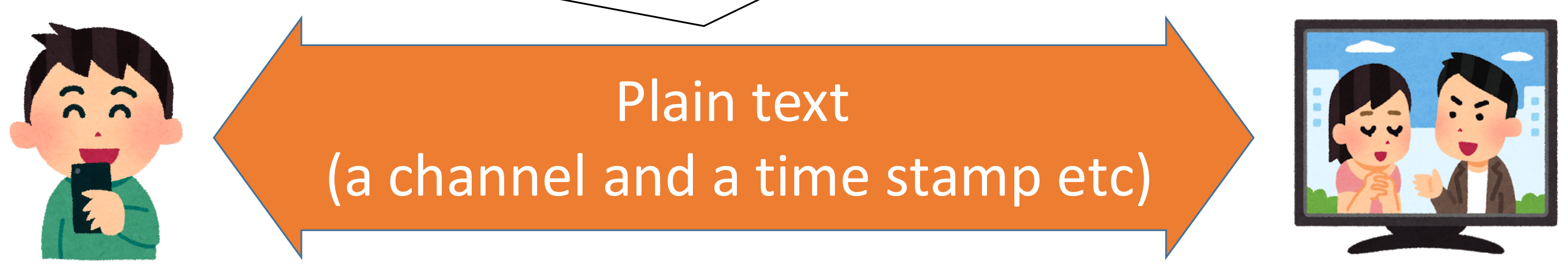
A Privacy Risk Arising from Communication with TV: Consideration from Attribute Inference for Users

Takahiro Higuchi, Osaka University
Naoto Yanai, Osaka University
Kensuke Ueda, Mitsubishi Electric Corporation
Yasunori Ishihara, Nanzan University
Toru Fujiwara, Osaka University

A Privacy Risk Arising from Communication with TV

- Sensitive information about users may be revealed via communication between the users and the TVs.

TV programs viewed by users can be identified via the plain text.
（We actually have found such a TV. ）



Plain text
(a channel and a time stamp etc)

In Japan, terrestrial broadcasts consist of about ten channels, and their TV programs are broadcasted according to a schedule day and time.

Goal

- Investigate whether gender and age of users can be inferred via viewed TV programs
  - Propose an Inference method, and evaluate it

Inference Algorithm about Ages

- Algorithm : the k-nearest neighbor algorithm with Mahalanobis distance
- Required Data : statistical data with respect to viewing information for each age
  - A dataset whose variance between ages is large, but that in a single age is small

Such a dataset has been unreleased, so we also discuss how to generate a suitable dataset.

We also infer genders of users in a similar way.

Generation of Statistical Data

1. Conduct a Questionnaire
   - Question : From the given list, choose TV programs which respondents actually viewed

     The list consists of 102 TV programs in total of 4 genres (drama, animations, sports, and variety shows) broadcast in Japan during 2018/9/20-10/20.

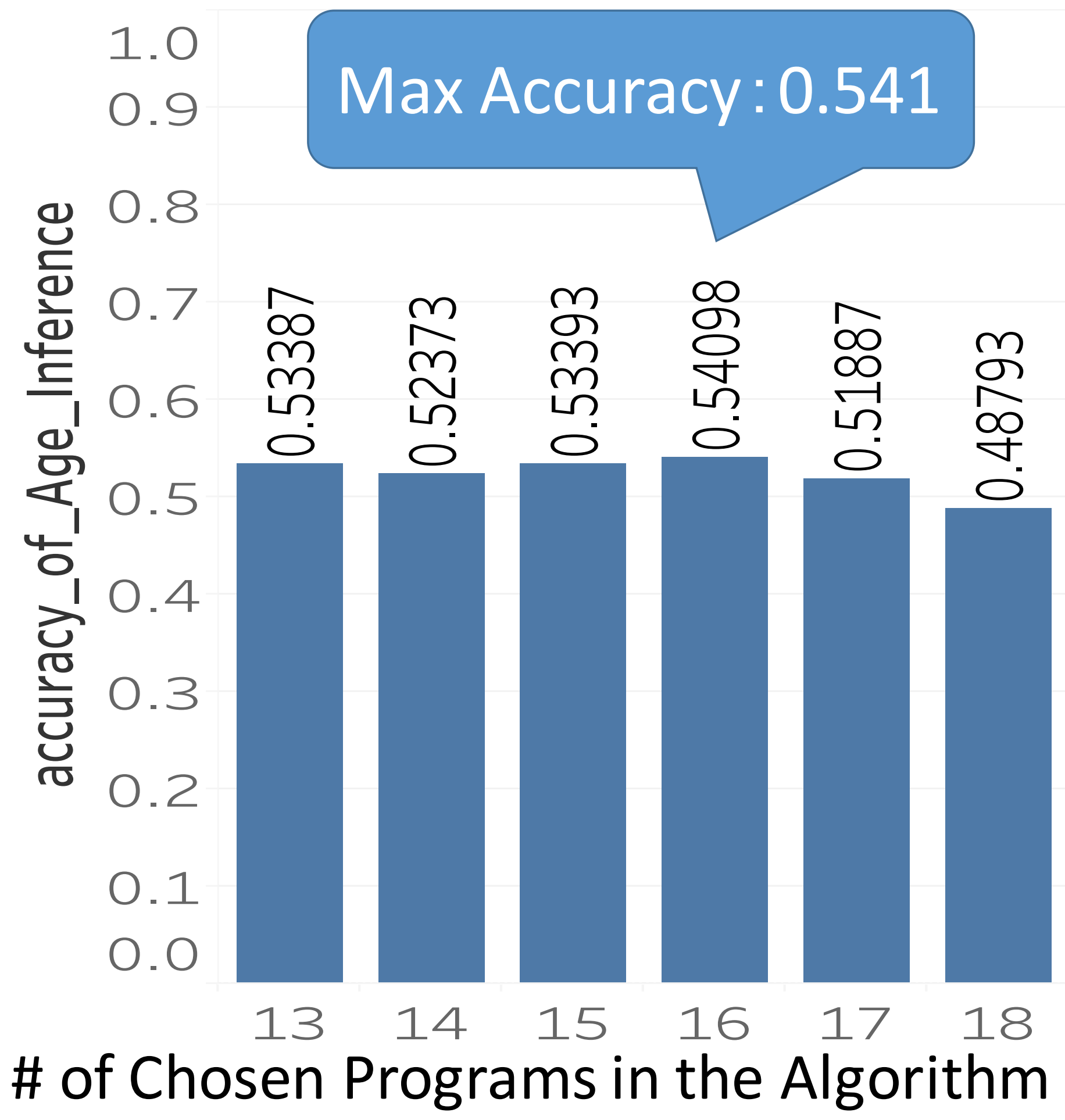
   - Questionnaire Result : About 110 answers for five genders, i.e., “18-24”, “25-34”, “35-44”, “45-54”, and “55 and over”) and two genders, i.e., “male” and “female”, that is totally 1101 answers

     In this work, classify ages (and genders) into one of the classes.
2. Choose a list of TV programs viewed by some of ages from the result
3. Generate variance-covariance matrices and mean vectors as statistical data

Evaluation of Inference Algorithm about Ages

1. Classify the respondents into test data and data for statistical data
2. Infer ages of test data, and score as follows
   - 1 point : if the result is identical to an actual age
   - 0.5 point : if the result is anteroposterior ages of an actual age
   - 0 point : otherwise
3. Evaluate the following division result as accuracy
   - Divide a summation of scores for the test data by the number of the test data

18 TV programs are chosen as TV programs viewed by some of ages.
We execute the above steps with statistical data generated by each of all the subsets except for empty set.

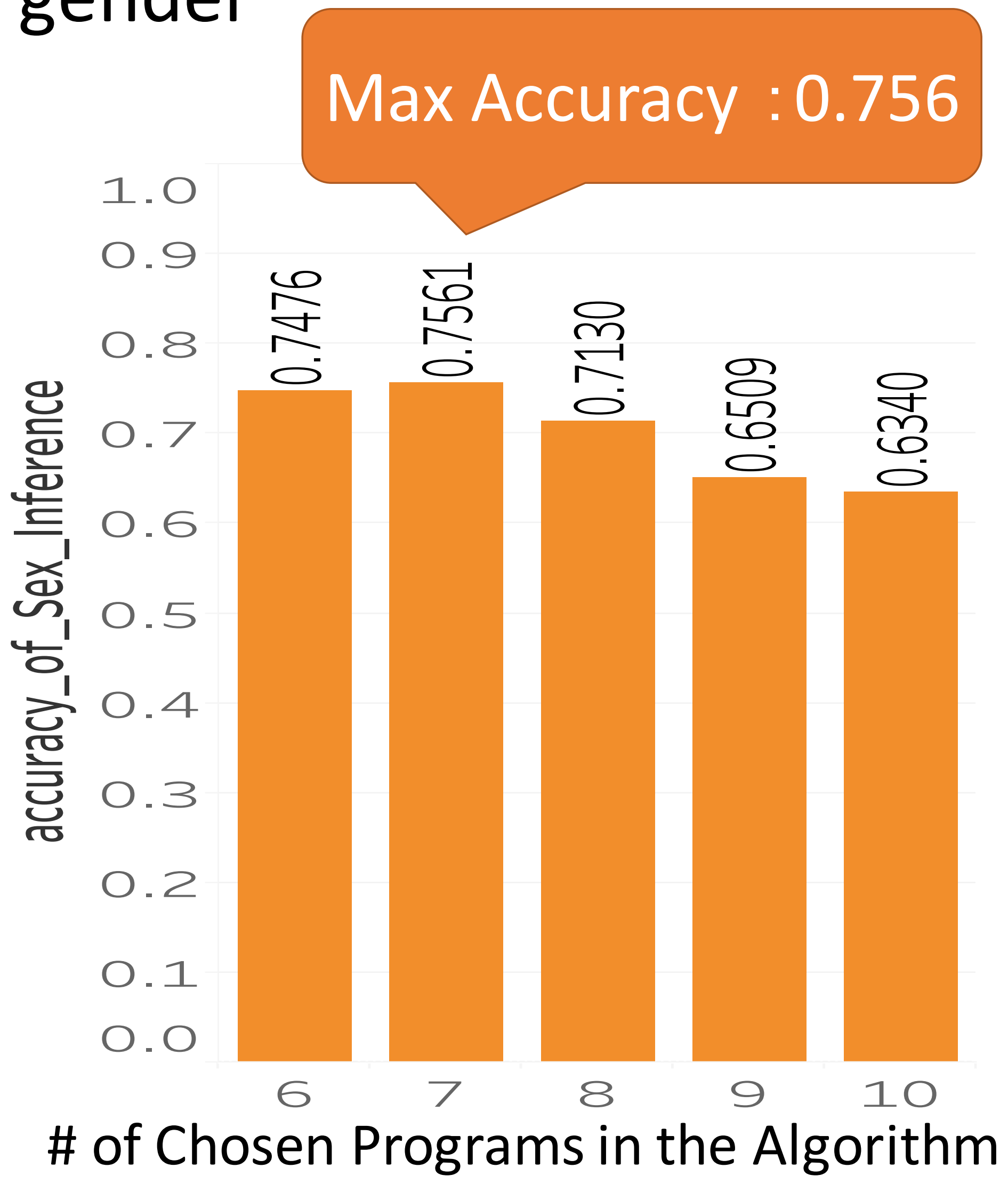


Max Accuracy : 0.541

Evaluation of Inference Algorithm about Genders

1. Classify the respondents into test data and data for statistical data
2. Infer genders of test data, and score as follows
   - 1 point : if the result is identical to an actual gender
   - 0 point : otherwise
3. Evaluate the following division result as accuracy
   - Divide a summation of scores for the test data by the number of the test data

10 TV programs are chosen as TV programs viewed by some of ages.
We execute the above steps with statistical data generated by each of all the subsets except for empty set.



Max Accuracy : 0.756