IAOS Satellite Meeting on Statistics for
the Information Society
August 30 and 31, 2001, Tokyo, Japan

An Advanced Statistical Information Disclosure over the Internet

INOUE, Tatsuki, *Waseda University*
ASAHI, Yumi and *Rikkyo University*
YAMAGUCHI, Kazunori *Rikkyo University, Japan*

Abstract

We illustrate the outline of an online system for disclosing the microdata set via the Internet without showing the raw original data. And we discuss the security of protecting personal information in the system and list up some approaches by using the current Internet Technologies. This is one of the examples of advanced order-made summarization system for the microdata statistics in the E-governmental services.

1 Introduction

By conventional statistical software, a dataset is read into the software by our hands, and researchers analyze the dataset, and then we obtain the statistical information from the resources. Whereas the government of Japan holds a microdata set that is not utilized for extracting information directly without fiduciary obligation of government employees in Japan. We get the useful information by using order-made summarization of the microdata in governmental service instead of the restriction, but any various statistical methods are not able to apply in this service.

In the recent study, some useful perturbation techniques for the disclosing microdata set e.g. global/local recording, suppression, additional noise, are discussed by many researchers including Willenborg and Waal (1996) and Takemura (1999). The techniques are effective in controlling disclosure risk of microdata set through the transforming observed values into broader intervals or suppressing the values before opening the dataset to the public. So the disclosed dataset is already processed information.

On the other hand, there are many services for disclosing some statistical information on the Internet. And several online statistical applications by using Java technology or CGI system on WWW browser are developing with widening the Internet (Inoue et. al. 2001). In these situations, online statistical software for microdata set naturally attracts us, i.e. the online statistical applications directly access the original of microdata set and the public users get only processed information via the Internet.

One of the reasons why the microdata is so restricted to open to public is the protection of answer's private information under Statistical Law in Japan. It protects a person from suffering a financial loss for the disclosure. For a nongovernmental enterprise, the financial loss would be heavy damage if the other business rivals acquired the knowledge through the disclosed own data. But it is potentially useful for the people to disclose microdata that are held by the government if the system ensures people's privacy.

In this paper, we introduce an online system for summarizing a microdata set by any users via the Internet without showing all numbers of individual observations. This is an example of advanced order-made summarization system for the microdata statistics in the electronically governmental services.

2 Data Representation System with the faculties of analyzing undisclosed data on Web

It seems inconsistent statement on analyzing for closed data set in the public space. But, it is available to create a system that serves some calculation results or graphics from a dataset to users and not provide the

whole of its resources.

It assumes that a microdata set is fitted into a rectangular sheet with individuals and variables except variables of directly identifiable each individual, e.g. the annuity holder's number or the driving license number. Observations consist of categorical or quantitative numerical numbers.

A "masked data analyzing system" that shows us statisticalresults from the closed data via analyzing. The system dose not open the entire source data to the users via the Internet. The system gives us the variables name and its description without numerical data.

There are two approaches to build the masked data analyzing system. One approach is to add masking-function in general statistical software by each software vender. The other is building up analyzing system without facility of data source printing by each data provider. The example of the former type is that the statistical software venders distribute masking-function such as add-in modules, and a data owner provides closed data with highly encryption, the encryption can only be decoded by the software vender's add-in module. Thus, the users who have statistical system with the special modules can analyze the crypt data with masking system. Add-in type has advantage that the general statistical software has various methods in the packages. However, there is less secure than another type. The encryption data design for decoding i.e. it is available to decode by using some algorithms and distributing to general public users. This situation is more chance to leak the entire data. Therefore, we should not send microdata on line even if involving masked facility but only send processed statistical information. So, the data providers make the masked data analyzing system as follows:

1) The masked data analyzing system is server side calculation system.
2) It shows variable information and available methods for representation.
3) Users can access it via Web interface.
4) Users can select variables and methods, and then Server calculates statistical information and returns the result to users.

However, their facilities are not enough to protect answerer's privacy from any third party. We should concern the following four additional security requirements for the identification problems:

5) We safeguard gaining access to the computer from hackers.
6) We avoid reverting to the original dataset by gathering calculation results.
7) The calculations results are guaranteed against deliberate falsification of data & programs by intrusion into the computer or on communication paths.
8) We devise a countermeasure to continue services against a dangerously overload or DoS (Denial of Services) attack.

2.1 An example of masked data analyzing system

In this section, we introduce a system as an example with graphical output by MDS (Multi-Dimensional Scaling) statistical method, which satisfied (1)-(4). (5)-(8) are discussed in the next session. MDS is to determine a space in which each concept can be mapped so that similar objects are close and dissimilar objects are far away. The MDS algorithm seeks to find the fewest dimensions in which such a space is possible.

In this example, variables names in a data set are selected by users on the Web browser for making a (dis)similarity matrix among the objects. The matrix is calculated by a Perl scripts through the CGI system on the server with the users input parameters, and then the obtained matrix is given as a parameter to MDS package on the scripts. The MDS package is distributed by NetLib on the Internet and the package is already installed in the server. The X-Y coordinates of the plot points of the objects are extracted from the output of the MDS calculations on the scripts. Finally, the scripts save the coordinates on the server's disk

space and print out an URL of a plotting program with the coordinates location as a parameter to the users. The users get graphics image of arrangement of selected objects on the browser as the result of the exchange of data between a server and a client. The plotting program is written by JAVA language for providing interactively interface to users on the web browser. Fig.1 illustrates the flowchart of these processes.

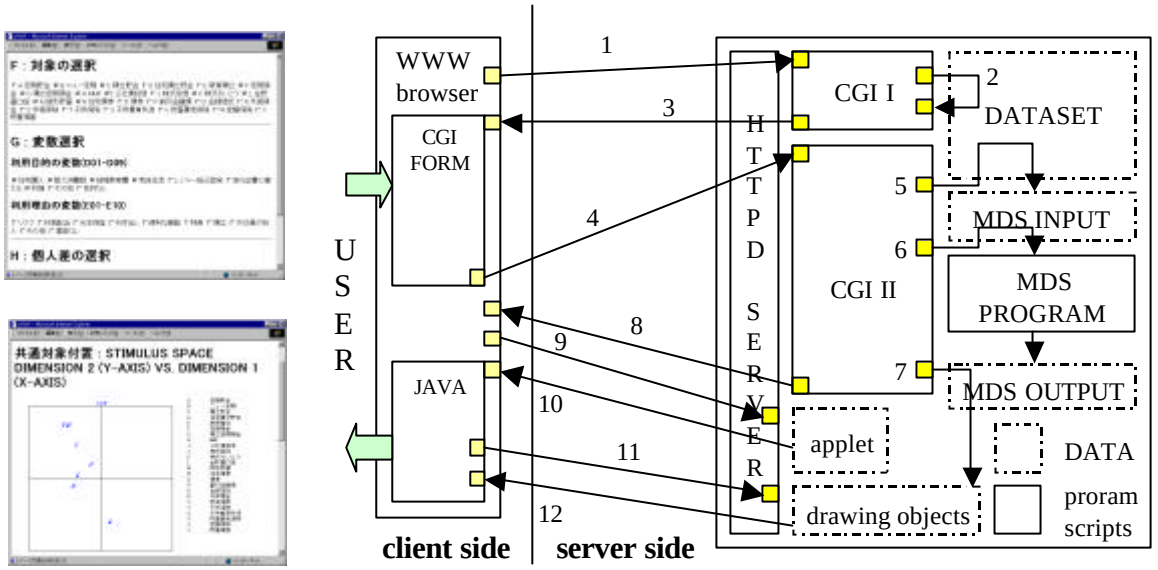

Fig 1. The flowchart of an example masked analyzing system.

3 Laying on security for preventing leakage of the information

Some ideas of additional security mechanism for that online system are described here. There is no absolute guarantee of protection against hacking server. But, it is vital that the servers keep secure against the intruders. Previous example system described in the previous section is all in a standalone server. The standalone server is more sensitive than functionally separated servers i.e. consist of the following three machines:

a) HTTPD server for providing the disclosure system's interface to users,
b) Database server for storing the microdata set,
c) Calculation server for getting statistical information from the database.

HTTPD server has two network interfaces. One interface is connected with the Internet and the other is connected with a private network. A database server and a calculation server are on the private network and they are not access directly from the Internet (Fig. 2). The database server b) is read from only the calculation server c).

The MDS package is already installed in this system, but any packages or procedures for getting statistical information are available if the packages are able to install into and run on the server c). The utilization of online distributed package for calculation engine is useful for detection of the modifications in the server too. The calculation server gets a target package from each different two-mirror site individually. And then the server makes a comparison of each package for checking the program modification against an intrusion into the server. This mechanism is one of the assurance about security requirements 7) in the section 2. And a standard is needed to use the variety of opened statistical packages on the Internet. The standard is of benefit to us both statistical packages distributors and its users.

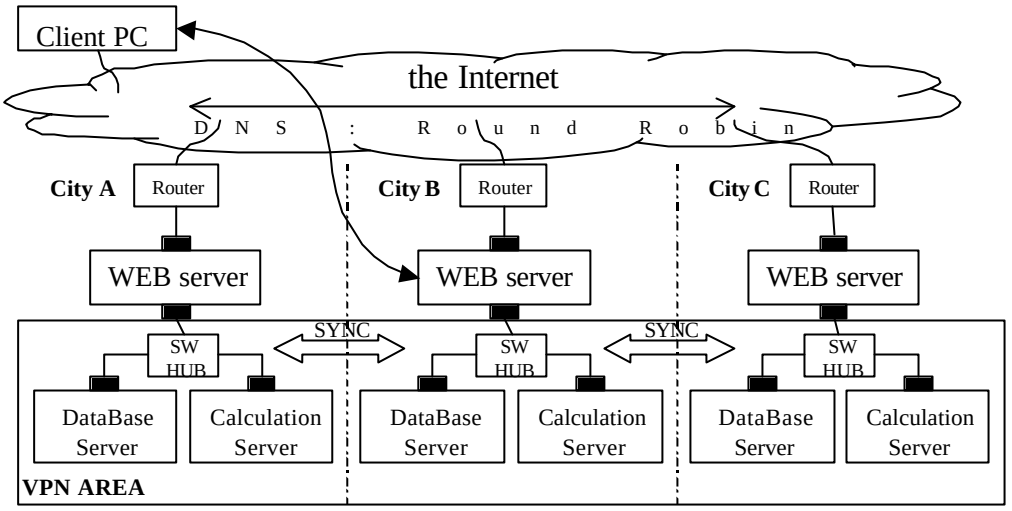

Fig. 2 Secure distributed system for the masked data analyzing via the Internet.

Relying on different two or many Internet servers greatly enhances the system's security against 7) and 8). Mirroring machines as a receptionist of this system are allocated to several municipalities with the server b) and c). Each server on the private network is maintained and connected via VPN (Virtual Private Network) bonding technology for synchronizing dataset or coordinating calculation. A User posts a statistical calculation to two different sites then these servers answer the result of the calculation to the user individually. Users are able to check whether answers are same or not by themselves. Such users own checking improves reliability and safety on the masked data analyzing system. From the dispersal over the Internet, it is expected for an effectiveness of a load balancing against 8).

Through the result of calculation, a reconstruction of original data or an identification of individuals should be prevented. For example, users submit a calculation to the system for getting a minimum of observations if the system provides summarization methods of a microdata set. The result number is observed value, but if the value frequently occurs in the dataset, the disclosure of this value to users is complicated in identification of individuals. However, if the number is unique observation in the dataset, the disclosure causes identification problems by the retrieval of information with the unique key. We consider the countermeasures for leaking detection of such a unique key, which simultaneously calculate the statistical information and the risk of identification on the calculation server. When the risk of identification exceeds a permissible level, the result of the calculation is processed for hiding the key or the server aborts the calculation immediately. As yet the permissible level and the risk function against identification have been studied. And we have considered the possibilities of some modification of the theory for controlling disclosure risk of microdata set.

References

Inoue et. al. (2001). A prototype of Data Representation System, Proceedings of the ISM Symposium 2001 –Statistical software in the Internet age-, ISM.

Takemura A. (1999). Local recording by maximum weight matching for disclosure control of microdata sets, ITME Discussion Paper No. 11, Faculty of Economics, University of Tokyo.

Willenborg, L. and de Waal, T. (1996). Statistical Disclosure Control in Practice. Lecture Notes in Statistics 111, Springer, New York.