Capturing human/social activities through massive web data

Psychological Reluctance to Personal Information Disclosure and Anti-Leak Technology

Today, so many people want their personal information undisclosed that it is necessary to ensure its anonymity for preventing personal identification. However, excessive anonymization leads to a deterioration of information quality and impedes effective use of information. How is it possible to bolster active use of information while protecting it?

A dilemma between anonymization and use of statistical data

What studies are you each undertaking in the domain of human and social sciences as part of the Transdisciplinary Research Integration Project?

Echizen My research seeks to make academic use of statistical data, including personal survey data that were collected with the assurance that they would be used only within the university or research institution, beyond organizational boundaries while maintaining anonymity to a certain degree.

For example, if medical data, which typically includes the patient’s name, address, age, disease, and medication, were made publicly available without change, the patient could be identified. To prevent this, it is necessary to blur the patient’s attributes by deleting the name and address and by generalizing some of the information; for example, “Tokyo” could be generalized to “Japan,” and “age 32” could be generalized to “thirties.” Since more than one person would usually have the same general attribute, individuals could not be identified.

However, this generalization approach to ensuring anonymity impairs the value and accuracy of data. In other words, there is a trade-off between the level of data anonymity and the academic utility of the data. In this environment, the new Statistics Act came into force in Japan in 2009 to pave the way for academic use of questionnaire information. This means that higher priority was placed on academic utility than on data anonymity. Data anonymity has traditionally been emphasized on the assumption that anonymized data could be made freely available. Nowadays, the degree of anonymization has been lowered to increase data utility while more emphasis has been placed on measures to prevent data leaks.

Kobayashi In fact, many past statistical data published for secondary analysis were hard to use. For instance, they had address information merely at the prefectural level and hampered comparative analyses between urban and rural areas. Among the data governed by the Statistics Act, social survey data, such as the information that I use, are mostly personally unidentifiable. I do not think it is necessary to further anonymize individual data, unlike medical information. In fact, we recode any observations with extremely rare frequency as missing values. This is called top-coding. A considerable number of methodologies for boosting anonymity have already been established.

So, my research focuses on how to encourage people to offer their personal data. Among other things, life logs on human behaviors and communications are now gathering the interest of those in academic and other circles. However, people are so worried about privacy that many are reluctant to offer them. I am exploring what constitutes their psychological reluctance and how to persuade them to offer socially valuable data.

Isao Echizen
Associate Professor, Digital Content and Media Sciences Research Division, NII

Tetsuro Kobayashi
Associate Professor, Information and Society Research Division, NII
Data leak prevention technologies and ideas for lowering psychological reluctance

Please give more details about your respective studies.

Echizen I have developed a method for identifying the source of leaked anonymized data that associates individual anonymization processes with user identification data. It is called “fingerprinting of anonymized data.” (Figure) This approach capitalizes on the multiplicity of data anonymization processes, which means that there are many different processes for achieving the same degree of anonymization. Suppose that there are data consisting solely of birth date and gender. User A is provided with a data file containing “1971” and “male” while User B is provided with one containing “August 10, 1971” and “gender unknown.” Our method prepares, for each user, a set of data generated by an anonymization process that varies with the user that has the same level of anonymization as every other prepared set. In the event of data leakage, the association between the anonymization process and the user ID is used to help identify the person responsible. Moreover, awareness of this identification method among users should make them more careful about data management. That is, the anonymization processes themselves deter data leakage.

Application of this method to social networking services (SNSs) and blogs would enable the source of a privacy leak to be identified from an analysis of the text containing the leaked information. In this application, not only would the anonymization process used vary with the user, but the degree of anonymization would vary with the group.

Kobayashi I conducted an experiment in which smartphone users were asked to disclose their life logs from their smartphones. The life logs were specifically classified into three categories: (1) locative information, (2) web browsing history; and (3) voice, SNS, and Gmail communications. In this event, three different levels of compensation were set for separate categories: 1,000 yen for category (1), 5,000 yen for (2) and 10,000 yen for (3). I introduced the conditions that combined the categories of life logs offered and compensation levels (i.e. $2^2*2^3 = 24$ conditions) to see the effect of each factor on the life-log disclosure. As expected, the experiment confirmed that the life-log disclosure was facilitated with the monetary compensation. On the other hand, it also found out that nearly 30% of the subjects disagreed to make any disclose irrespective of the level of compensation. Many showed strong reluctance to offer information about voice and other communications histories even if the content of communications was not to be recorded. On the other hand, their level of reluctance was lower with respect to GPS data. The subject may not have been very aware of the risk involved in offering locative data. Therefore, I think that it may be possible to gradually widen the scope of available data from which induces relatively minor reluctance, like GPS data, to remove the initial psychological barriers, and then move on to the types of information that people feel more reluctance to offer.

There is several major hurdles to the collection of personal information and it must never be leaked even after collection. Meanwhile, if everyone offers it its overall quality will improve and its utility as “public goods” will increase as well. So, it will be necessary to conduct further studies on the design of the incentive to facilitate information disclosure.

Echizen I agree. It is tough to deal with two mutually conflicting factors, namely data anonymity and data utility. We will continue our research to create a good technique for balancing them to contribute to effective use of information.

(Written by Madoka Tainaka)