

# Protecting Privacy While Revealing Data

Bernardo A. Huberman and Tad Hogg  
HP Laboratories  
Palo Alto, CA 94304

**There is an alternative to the familiar debate of privacy versus the public interest in access to medical data for epidemiological studies. It consists of cryptographic mechanisms that allow investigators to have access to an individual's data and to contact them person with further questions while at the same time preserving their full privacy.**

The ability to collect and disseminate fine-grained data in the medical field has led to expressions of concern about privacy issues and to public reactions that in some cases have translated into laws [1,2]. In the case of some European Community nations, strong restrictions have been placed on the ability of those who collected personal data to release it without explicit individual consent.

The coming use of genetic information to personalize medical treatments has the negative flip side of allowing finer-grained distinctions by insurance companies of the individuals concerned. Genetics introduces a further complication in that information about one person is statistically relevant for that person's relatives as well, due to their common genetic characteristics. Thus, even if one person is not concerned about revealing genetic information, it may nevertheless be a concern for some relatives.

While these concerns are important, it should be pointed that the release of medical data can also help the community at large, particularly through epidemiological studies to identify new diseases. In this case, there is a need to balance the social benefit of these studies with the loss of privacy that they seem to entail [3]. The current policy proposals often fail to provide this balance and in many cases put restrictions on data sharing that can be detrimental to the public interest, as in the case of epidemiological studies. In fact, new interpretations of these privacy protection laws seem to preclude even the access to data collected by doctors, and in the UK even the names of the doctors who already have relevant data for studies cannot be revealed [2].

It appears that countries face two alternatives: , full disclosure or full privacy. In reality neither option is appealing. On the one hand, full disclosure will likely make individuals more reluctant to use medical services for rather routine problems. On the other hand, full privacy, achieved through anonymous services, limits the range of epidemiological studies by preventing researchers from following the health of particular groups identified through initial contact with the medical community. For instance, it may only be apparent after a study is underway that additional questions about the individuals or their relatives would be appropriate.

An alternative, and simplistic, approach would be to resort to a trusted party or entity that would act as an intermediary between the subjects and the researchers while protecting their privacy. The difficulty with this alternative is that it is hard to find someone or an institution that everyone likes. Worse, it provides a single point of failure, for if this entity were compromised all data files could suddenly become public. Even with legal protections, citizens might anticipate that laws could change with time, as in the case of adoption rights, where today it is possible to obtain the identity of parents who gave children away for adoption at a time when the legal standard offered them anonymity for life.

As we point out below, there is a technical solution to the problem that allows for investigators to have access to individual data and to contact them with further questions while at the same time preserving their full privacy. This solution relies on zero-knowledge cryptographic techniques developed in the context of secure distributed computation. Our scheme allows a researcher to issue a survey to a number of individuals who can answer in what effectively amounts to an anonymous fashion, while they can still be tracked over time and queried on additional items without the researcher learning the identity of the subjects. Moreover, the solution we propose does not even require a trusted third party, which for the reasons stated above is not a suitable solution.

The “magic trick” behind this solution can be explained in simple terms by using a simple analogy. We first describe this analogy and then explain how to implement it computationally.

Consider a bulletin board where survey questions are posted for all members of a community to see. For the sake of simplicity in the exposition, we'll assume that the answers to these questions are of the form "yes" and "no", although the mechanism is much more general. Each subject answers the question by effectively *anonymously* "placing" on the bulletin board two unlocked boxes, labeled "yes" and "no" with locks designed in such a way that the subject only has the key to the one corresponding to his answer. This is shown in Figure 1

## Bulletin Board

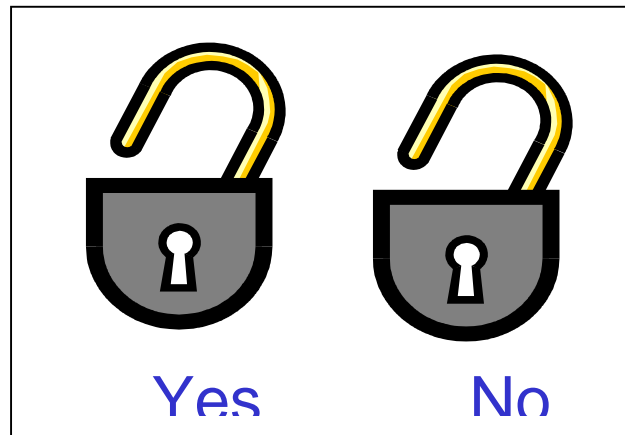


Figure 1



Since no one else knows which of the two keys the subject has, others, including the researcher, cannot tell how a given subject responded. And yet, he can contact each of the respondents that answered the question in a given way by creating a box that can be unlocked only by members of the selected group, as shown in Figure 2.

## Bulletin Board

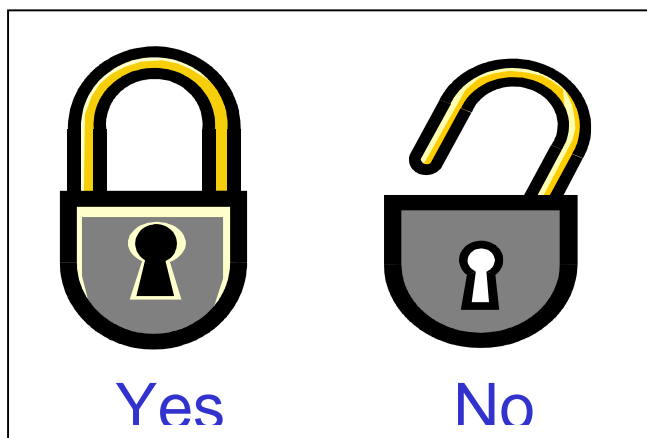


Figure 2



Placing messages in this box and then locking it, allows communication with members of this group, defined by their answer to the question. Thus the researcher can ask group members further questions. This mechanism need not be restricted to the researcher: it can also allow members of the group to communicate with each other (e.g., as a chat forum) without them learning the identities of others in the group. All of this occurs in full view of the whole community, but with decrypting abilities possessed only by those who answered in a given fashion.

This method or technique provides a potential solution to the dilemma of protecting privacy or making it public. Notice that it does not require a trusted third party,

although the underlying implementation, which we discuss below, does require the users to follow standard and tested cryptographic protocols. This trust is no different from the trust we put in a locksmith when asking for a copy of our household key, or on the manufacturer of a garage door opener.

A simple application of this technique counts individuals with a given property. All that is required is to post a message with a key requesting an acknowledgment from all members using that key. The number of answers compared to the whole population yields a useful frequency. Another form of panel research would follow a group over time, effectively conducting prospective surveys by simply adding more questions to the bulletin board and watching what happens to the frequencies. This would also allow looking for correlations among members of different groups. That is accomplished by repeating the original procedure in a more refined fashion.

This physical metaphor can actually be implemented and automated in a transparent fashion by using public key cryptographic systems[4]. These systems rely on a pair of related keys, one secret and one public, associated with each individual participating in a communication. The secret key is needed to decrypt (or sign), while only the public key is needed to encrypt a message (or verify a signature). A public key is generated by those wishing to receive encrypted messages, and broadcasted so that it can be used by the sender of the message to encode it. The recipient of this message then uses his own private key in combination with his public key to decrypt the message. Popular public key systems are based on the properties of modular arithmetic. In our particular application we use the additional property that by constraining the product of two or more public keys to be equal to a specific large number, it is only possible to generate a set of such keys in which only one of the keys has a corresponding private key. This provides the computational basis for the analogy of the two locks described above: each person answers the question by posting two public keys, constrained so that their product matches a value given as part of the question. The person can only have a private key for one of the posted public keys, and selects the private key corresponding to the answer.

The security of the full system requires addressing additional issues. For instance, to what extent do laws protect people from having to reveal their secret keys? Another issue is the size and diversity of the group, enabling people to effectively hide among

other members. In some cases, incentives for participation and correct answers can be important and some possible answers have been proposed, like markets for secrets[5].

This proposal provides a third alternative to the dilemma of having to choose between privacy and the public interest. While these two have been part of the public discourse for many years, the new developments in genetic research and information systems raise these issues to a higher level concern. While the social benefits of novel privacy mechanisms are not usually considered in policy discussions of the use of cryptography, they illustrate an important opportunity for allowing widespread use of these technologies.

#### References

1. Haim Watzman, Israel split on rights to genetic privacy *Nature* **394**, 214 (16 July 1998)
2. David Adam, Data protection law threatens to derail UK epidemiology studies *Nature* **411**, 509 (31 May 2001).
3. Patricia A. Roche and George J. Annas, Protecting Genetic Privacy, *Nature Genetics* **2**, 392 (May 2001).
4. Bernardo A. Huberman, Matt Franklin and Tad Hogg, Enhancing Privacy and Trust in Electronic Communities", in Proceedings of the ACM Conference on Electronic Commerce (EC99), 78-86 ACM Press" (1999).
5. Eytan Adar and Bernardo A. Huberman, A Market for Secrets. FirstMonday, August 2001. [http://www.firstmonday.org/issues/issue6\\_8/adar/index.html](http://www.firstmonday.org/issues/issue6_8/adar/index.html)