Technologies for forgetting

Ian Brown

Department of Computer Science,
University College London
Gower St, London WC1E 7AU.

Computers' most basic functions are storing and processing information. As the costs of permanent storage have plummeted, hard disk sizes are rarely now a constraint in the amount of information that can be retained. It is often more expensive to seek out and delete such data than to save it indefinitely.

From a privacy perspective, this is disastrous. We have seen time and again that companies and governments will find all sorts of reasons and mechanisms to profile and use collections of personal information once they are built up (for marketing, surveillance and a range of other activities). What technological mechanisms can we use to allow data to age and be forgotten as it is in human memories and communities?

This article is a short non-technical overview of possible mechanisms, from simple minimisation of the collection of personal data through to more complex mechanisms to restrict access to data based upon algorithms from the field of information security.

Data minimisation

One of the key insights of data protection law as it has evolved since the 1970s is that the best way to "forget" personal data is that it should not be "remembered" in the first place. It is often easier for information systems to be designed to collect and store all available data, because filtering and selecting often takes more effort than grabbing and saving everything to hand.

With plummeting storage costs system designers frequently operate on the precautionary principle that data stored today can be processed in many new ways tomorrow; but data ignored today cannot be re-accessed for a new purpose tomorrow. (This of course is the entire point from a privacy perspective).

With a little care, most information systems can be designed in a way that achieves system objectives whilst minimising the collection of personal data.

Data encryption

Data in transit (across a communications medium such as the Internet) is technically easy to access by those controlling any of that communications infrastructure. It can also be accessed using a range of surveillance technologies operated by governments and even large corporations. Data at rest can be accessed by anyone with physical access to the storage medium.

Data encryption can be used to scramble data so that only those with permission can unscramble and access that data. Even intelligence agencies with multi-billion-dollar budgets are unable to use their computing capabilities to break well-designed encryption systems (although they are very often able to guess passwords that people have used or find vulnerabilities in software that can be used to bypass encryption).

Encryption can be used to efficiently "forget" large amounts of personal information, which may be stored in several different locations and be difficult to securely delete. If the data is encrypted, the deletion of the carefully controlled decryption "key" (usually at

most 256 bits of data or around 34 characters) effectively prevents further access to that information. Automatic deletion of keys on a pre-programmed expiry date can thus "age" and permanently wipe any quantity of data.

Trusted computing

Trusted computing is a technology for securely managing decryption keys that is about to achieve mass deployment on home and office PCs. It generally takes the form of a chip or "Trusted Platform Module" (TPM) included on computer motherboards. TPMs securely hold keys in a way that is much, much harder to subvert than those protected by software-only systems.

There are many potential applications of TPMs: most controversially, they are designed to restrict copying of digital media files. But one application would be to hold decryption keys that automatically expired a certain period of time after they were created, rendering unreadable any data encrypted using a related key.

Mixes and remailers

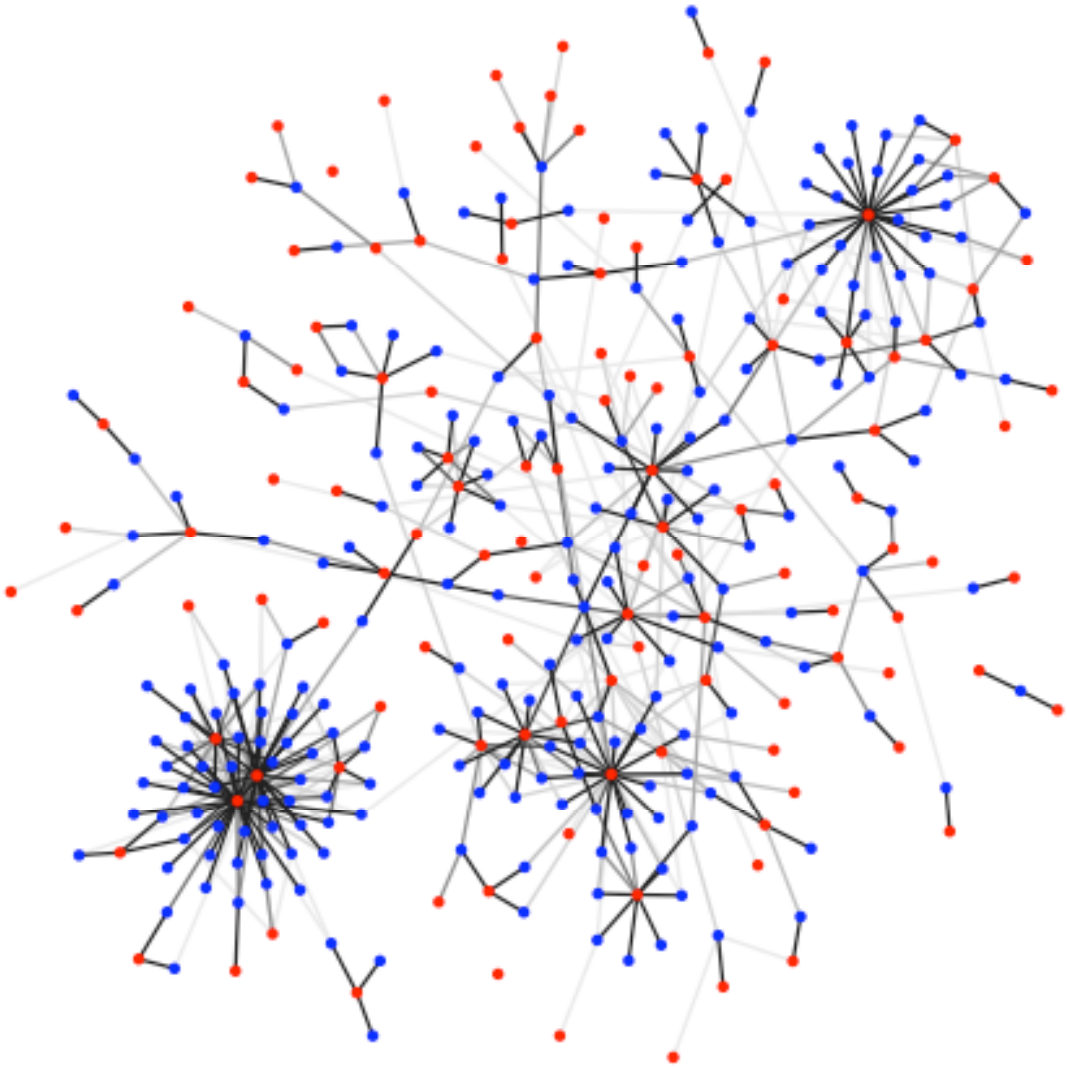

Figure 1: Friendship tree (source: George Danezis and Bettina Wittneben)

Even when communications (such as VoIP calls or e-mails) are encrypted, the fact that two individuals have communicated is visible to a number of parties (including their Internet Service Providers and many government agencies that have mandatory access to this data). Sophisticated algorithms can be applied to visualise "friendship trees" of groups and their communication patterns (see figure 1).

Encryption "mixes" mask these patterns of who is talking to whom, by relaying encrypted communications through a series of Internet servers that slightly reorder data packets before sending them on to the next server in the chain. They therefore stop the "remembering" of personal relationships and accesses to websites.

Private Information Retrieval

Whenever an individual retrieves information from a website or other data source, that source is able to link any identifying information given by the user with the data they have retrieved. So, for example: Apple knows which songs and album artwork users have downloaded; the New York Times knows which stories its web subscribers have accessed; and Google knows the terms in every search made through its site.

Private Information Retrieval uses cryptographic techniques to allow users to retrieve information from a source that cannot find out the data that has been accessed. Oblivious transfer is a technique that adds a further restriction that means a user cannot find out any more information from a server than they requested.

Conclusion

The most reliable way that information about individuals can be "forgotten" is to be very selective about information that is initially "remembered." Data minimisation techniques use simple system design to screen out personal data that is not directly required for the working of an information system. Mixes and remailers and Private Information Retrieval prevents the operators of networks and servers, and those in a position to observe their operation, from learning about links between people, organisations and the information they access.

If personal data must be collected, it can be protected from illegitimate access using encryption systems. Secure deletion of keys, particularly those held in trusted computing hardware, will definitively prevent future access to any quantity of personal data protected using those keys.