TAP
Technology
Advisory Panel
IPCO

04 September 2019

Challenges in Data Tracking

1. In the context of understanding and controlling data flow in and across organisations:

    - Is it possible to add data tags to pieces of information shared with others so that you could later query to uncover if and where this information has been saved onto another system?

    - Could you query to see whether it has been deleted?

    - Can it work across organisations?

    - How are “blockchain” or “distributed ledger” technologies related to this?

2. Summary direct responses:

    - Tagging data and tracking can only work in cases where you own and operate all of the IT that can access the data and can enforce policies;

    - A distributed ledger can be a useful component of an internal tracking system;

    - Sharing across organisation IT boundaries relies on trust in other organisations to implement equivalent mechanisms;

    - However, even in this situation, preventing all copying by a person authorised to see the data remains unachievable, and if you cannot be sure there are no copies, you cannot be sure it has all been deleted;

    - Blockchain adds further complexity to a distributed ledger that adds no value in the scenario outlined.

    - In simple terms, the answer to the initial question is “no”. It is not possible by technical means to know whether copies of a document have been made or whether the document or any copies have been deleted. The only way to know that a recipient has not retained or copied a document is through old-fashioned trust.

State of the Art: Digital Rights Management and computer backups

3. The commonest example of the use of digital technology to restrict and track data use has been the various attempts to implement "Digital Rights Management" (DRM) to enforce licensing conditions for copyright content. This example highlights the challenges in building systems that try to control data use when data is shared – unless the computing environment into which the data is released is also closely controlled, then subversion of the enforcement mechanisms is always technically feasible. As such DRM-like mechanisms can be, and are, used within organisations to provide mechanisms to remind staff of the limitations of use of data, and record access, but they cannot be viewed as an enforcement mechanism - the enforcement is employment contract. However, they are cumbersome, and will, by their restrictions, inevitably impede research into new and creative ways of processing data. Sharing across organisations would simply require trust, as any technical means could be circumvented.

4. Such systems often use protection techniques such as digital watermarking, or encryption of all, or some, of the content, but in essence the protected file is still just a file – a bag of bits – and making multiple copies is straightforward. Note that the operation of computer system back up creates multiple copies of data in offline storage media which makes the confirmation that all copies have been deleted a human process of physical media management, not a digital challenge. While first observed in the 1970s when a "phone freak" database was found on a university computer and the GPO ordered the deletion of all copies of the files, this could only be complied with by destruction of all backups - the system was never designed to delete things. This matter continues to be a challenge, for example recent publications are concerned about the GDPR right to erasure which is in conflict with modern disaster recovery mechanisms implemented in the IT system of most companies. Without a regimen of constant erasure, at best we can put the data "beyond use" in these situations.

Merkle Trees and Blockchain

5. In 1979 Ralph Merkle patented the idea of "hash trees" – this is one key component of blockchain and has many useful properties all by itself – so much so that it is the most widely adopted mechanism for performing software revision control (e.g. git and the GitHub cloud service). A Merkle tree uses a digital signature to verify efficiently the integrity of data, and importantly can be used to perform consistency verification of a logging infrastructure (that is where we only ever append data), making the log tamper evident – someone could delete a record in the log, but it would result in the consistency verification failing. Much has been published on this in the academic literature[1].

6. Adding distribution to this sort of logging infrastructure by ensuring multiple copies exist is sensible for robustness and if it becomes clear that one copy has been tampered with the redundant copies allow recovery. This is a "distributed ledger".

7. Blockchain then combines a distributed ledger with a specific algorithm that aims to deal with contending updates to the log infrastructure and ensure consensus in the ordering of the contending updates; in the BitCoin case, these updates are the BitCoin transactions, and the algorithm is the energy-devouring "Proof of Work" mechanism - although the community continues to work on more energy efficient means to achieve the consensus goal. However, in the scenarios being considered here we simply do not have multiple entities trying to write to the

[1] "Efficient Data Structures for Tamper-Evident Logging", S Crosby and D Wallach, USENIX Security 2009

same log at the same time, and the complexities and power and performance cost of blockchain are unnecessary.

Summary conclusion

8. In summary:
    - Controlling the use of shared data can only operate within well-defined trust boundaries.
    - Prevention of copying by technical means is not possible.
    - Distributed Merkle trees can used to ensure data has not been tampered with; full Blockchain technology is only required when there are multiple, contending updates occurring at the same time.
    - While it is possible to demonstrate that a document has not been tampered with, it is not possible to confirm whether copies have been made or whether the file and all copies have been deleted.