Data-Purpose Algebra

Chris Hanson, Lalana Kagal, & Gerald Jay Sussman

Restrictions on the use of data

Data is often encumbered by restrictions on the ways it may be used. These encumberances may be determined by statute, by contract, by custom, or by common decency. Some of these restrictions are intended to control the diffusion of the data, while others are intended to delimit the consequences of actions predicated on that data.

The allowable uses of data may be further restricted by the sender: “I am telling you this information in confidence. You may not use it to compete with me, and you may not give it to any of my competitors.” Data may also be restricted by the receiver: “I don’t want to know anything about this that I may not tell my wife.”

Although the details may be quite involved, as data is passed from one individual or organization to another the restrictions on the uses to which it may be put are changed in ways that can often be formulated as algebraic expressions. These expressions describe how the restrictions on the use of a particular data item may be computed from the history of its transmission: the encumberances that are added or deleted at each step. A formalization of this process is a *Data-Purpose Algebra* description of the process.

One pervasive assumption behind our formalization is that a data item, annotated with its provenance, may be restricted, but this restriction is not on the content of the data item. For example, a law-enforcement official may not act on improperly obtained evidence, but if the same information was redundantly obtained through lawful channels the official may act.

Of course, there are other real-world circumstances where this assumption is invalid. For example, consider the fact that *Joe ate ice cream at 3:12 PM on 13 August 2006.* Now suppose that in the playground *Anne told me that Mary told her that Mary observed* *that Joe ate ice cream at 3:12 PM on 13 August 2006. Don't tell his mother!* We see that this item has restricted distribution. But if also *Jim told me that Joe ate ice cream at 3:12 PM on 13 August 2006* our formalization would allow me to tell Joe's mother that he ate ice cream before dinner. However, I would feel inhibited, as a matter of courtesy, by the fact that Anne told me not to pass this information along.

Data-Purpose Algebra

To formally describe the ways that the use of data may be restricted and the way in which the restrictions are transformed as the data is processed and passed from one agent to another we decorate each data item with extra information. Each data item i has, in addition to its content $q = Q_{\mathrm{D}}(i)$, its source $a = A_{\mathrm{D}}(i)$, a category $k = K_{\mathrm{D}}(i)$, and a set of purposes $p = P_{\mathrm{D}}(i)$ for which it can be used. An item is constructed from its content, agent, category, and purposes $i = I(q, a, k, p)$. The agent is the agent that produced this data. The category is a set of data items, from which the particular data item is chosen. This set may be named but is not likely to be enumerated. For example a typical legal category is "US person." The set of purposes is explicit; a typical purpose in the set of purposes is "criminal law enforcement."

A data item i may be processed by some agent a' to produce a new data item. (See figure 0.1.) The new data item has the same kinds of annotations as the original one, but the process generates new content and new annotations as functions of the original. The functions are specific to the kind of process performed by the agent.

For example, medical data about a definite person is quite restricted as to its allowable uses. But it may be anonymized for use in the education of medical doctors. In such a case, the allowed purposes of the anonymized data may be wider than the original data, and the category of the data will be different. On the other hand, when a person enters a medical establishment for treatment a record is made of the patient's name, address, and date of birth. This data is usually unrestricted, but the fact of it appearing in a medical record adds restrictions required by the HIPAA law.

An agent may combine data from multiple sources to produce new data. (See figure 0.2.) In this case the functions may be considerably more complex.

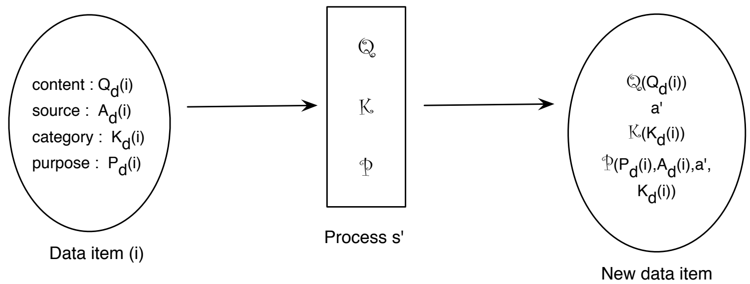

Figure 0.1 A data item i may be processed by some agent a' to produce a new data item. The new content is some function $\mathcal{Q}(Q_{\mathrm{D}}(i))$ of the given content. The source of the new data item is a', the new category is a function $\mathcal{K}(K_{\mathrm{D}}(i))$ of the given category, and the allowed purposes of the new data item is a more complex function $\mathcal{P}(P_{\mathrm{D}}(i), A_{\mathrm{D}}(i), a', K_{\mathrm{D}}(i))$ that may depend on the original purposes, the sources, and the category of the original data.

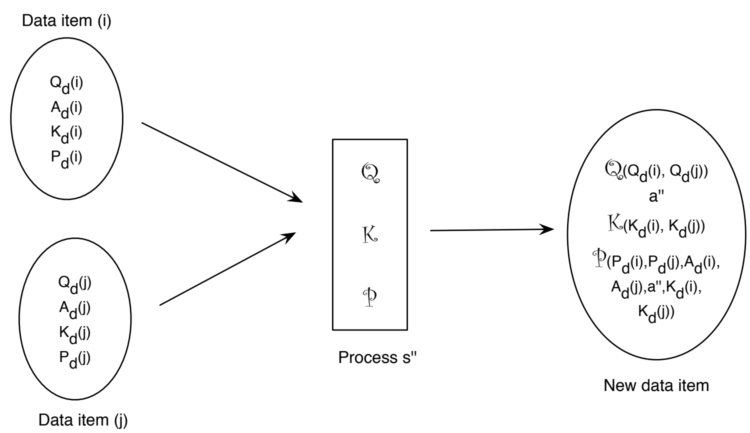

Figure 0.2 The process of combining data from multiple sources may be more complex.

For example, a person at the medical office may use a public source, such as a telephone directory, to verify the recorded telephone number of a patient. This process combines highly restricted information from a medical record with unrestricted public information, but the result remains restricted. It has been shown that sets of "anonymized" data can often be combined to discover the identities of the parties.[1]

An Example Formalization

In the illustration that follows we consider a simplified formulation of the rules for data passed among government agencies and officials, specified by the Systems of Records Notices associated with Systems of Records, as defined by the Privacy Act.

Let r be a data source, for example, a System Of Records (a SOR). A system of records r may be owned (controlled by) an organization $o = O(r)$, authorized to use information for purposes $P_{\mathrm{O}}(o)$. This restricts the use of data in r to be at most:

$$R_{\mathrm{ORG}}(r) = P_{\mathrm{O}}(O(r))$$

Also associated with the system of records may be a System Of Records Notice (a SORN) $n = N(r)$, which gives information about the permissible uses of the SOR. If there is a SORN, it specifies input conditions: the allowed sources $S_{\mathrm{S}}(n)$ from which data may be collected, the data categories $K_{\mathrm{S}}(n)$ that may be collected, and the purposes $P_{\mathrm{S}}(n)$ for which data that is collected may be used. It also specifies a set of routine-use notices $E(n)$ for data extracted from that SOR.

Each entry $e \in E(n)$ specifies a set of possible recipient organizations $O(e)$, categories of data $K_{\mathrm{R}}(e)$ that may be transferred to those organizations, and the set of authorized purposes $P_{\mathrm{R}}(e)$ for which the specified recipient organizations may use data from

[1] Latanya Sweeney did substantial work here. See `http://lab.privacy.cs.cmu.edu/people/sweeney/` for details. Is the combined data restricted? It depends on the laws. If the law says that medical records are restricted, then it is independent of how they are derived. On the other hand, it is possible to combine two restricted pieces of information to produce a less restricted deduction. For example, if the *same* information is available from two different sources, then the restriction on the combination may be relaxed to be the uses allowed by each source separately, or it may not, depending on the details.

the source. Any particular recipient r_1 may be a sub-organization of a possible recipient organization r_2 specified in a SORN. This relation is notated $r_1 \prec r_2$.

The purposes allowed for data i that has been transferred from a SOR s to a SOR r depend on the purposes that came with the data and the input conditions on the SORN. So, if s is not one of the allowed sources or the category of the data is not one of the allowed categories the data may not be used for any purpose:

$$\begin{aligned} R_{\mathrm{IN}}(i,s,r) &= P_{\mathrm{S}}(r) \text{ if } s \in S_{\mathrm{S}}(N(r)) \wedge K_{\mathrm{D}}(i) \in K_{\mathrm{S}}(N(r)) \\ &= \{\} \text{ otherwise} \end{aligned}$$

The set of applicable routine-use notices $A(i,s,r)$ for transfer of data item i from a SOR s to a recipient r is just the set of those entries for which the recipient is a sub-organization of an organization specified as a recipient organization of an entry in the SORN and for which the category of the data $K_{\mathrm{D}}(i)$ is in the allowed categories $K_{\mathrm{R}}(e)$ for that routine use e:

$$\begin{aligned} &A(i,s,r) \\ &\quad = \{e \in E(N(s)) \mid (\exists o (o \in O(e)) \wedge (r \prec o)) \wedge K_{\mathrm{D}}(i) \in K_{\mathrm{R}}(e)\} \end{aligned}$$

The restriction on authorized purposes of a transfer from a source to a recipient is that the purposes must be authorized by any of the entries that contain the recipient organization.

$$R_{\mathrm{OUT}}(i,s,r) = \bigcup_{e \in A(i,s,r)} P_{\mathrm{R}}(e)$$

The authorized purposes $Z(i,s,r)$ to which a recipient r may put a data item i extracted from a source s is then restricted to be those purposes particular to that data item that are also allowed by one of the purposes in the authorized routine purposes obtained from the SORN:

$$Z(i,s,r) = R_{\mathrm{IN}}(i,s,r) \cap R_{\mathrm{ORG}}(r) \cap R_{\mathrm{OUT}}(i,s,r) \cap P_{\mathrm{D}}(i)$$

So $Z(i,s,r)$ is the set of purposes of the new item held by the recipient r with the content of the old item i held by the source s.

The result of a transfer process A_{XFER} of an item i from a source s to a recipient r is a new item:

$$I(Q_{\text{D}}(i), A_{\text{XFER}}, K_{\text{D}}(i), Z(i, s, r))$$

To be done

The example shown above shows how to cover many kinds of formalizable requirements, such as those of the Privacy Act. But there are harder problems. We have not begun to consider the informal and implicit restrictions on the use of data required by cultural considerations, such as courtesy. However, we must confront the problem of being able to formally describe such currently informal notions to ensure that we can make a system that is sufficiently general to cover real-world situations.

Another problem is revealed by the situation where an entity is allowed to discuss the consequences of a secret it knows with any other entity that already knows that secret. Similarly, it is possible that an entity may hold a secret that it is only allowed to divulge if the reality is that the information is generally available through other channels.

Summary

The algebraic approach is well suited to modeling the allowable uses of information when the restrictions on that use are determined by the path by which the information is obtained, but it is not so good at dealing with restrictions that are time dependent or inherent in the content of the information, independent of the path. We will have to design means of modeling information with time-dependent and content-dependent restrictions.

When formalized algebraically, computations are directly representable as purely functional computer programs. There is no complex translation required. This makes it easy to verify that a program that implements the data-purpose algebraic computations is correct. In addition, because there are no side effects required there is no problem for synchronizing concurrent processes, making it easy to get good performance from massively parallel and possibly distributed processes.