January 2020

Machine-Learning Based Evaluation of Access Control Lists to Identify Anomalies

Alejo Grigera Sutro

Machine-Learning Based Evaluation of Access Control Lists to Identify Anomalies

ABSTRACT

Access control lists (ACLs) are a commonly used mechanism to limit the distribution of data in an organization and to protect data from improper access. As ACL size grows one task to be solved is to keep the lists up-to-date and validate that each person in the ACL has appropriate levels of access to data. Currently, managing ACLs and detecting the presence of anomalous individuals in the lists is a manual task. This disclosure describes the application of a trained machine learning model that utilizes as input various factors such as employee roles and organization structure to detect and flag ACLs with dispersion patterns that are likely indicative of potential improper access.

KEYWORDS

- Access Control List (ACL)
- ACL dispersion
- ACL audit
- ACL management
- Data security
- File permission

BACKGROUND

Many organizations store data, e.g., codebases, customer data obtained with user permission, etc. which can be sensitive and/or business critical. Given the importance of maintaining security of such data, various security measures are typically implemented to protect the data and guard against unauthorized access. A commonly used mechanism for such protection is the use of Access Control Lists (ACLs). An ACL specifies the persons or user accounts that are authorized to access the data; anyone else that attempts to access the data is automatically restricted from access.

ACLs related to useful data tend to grow quickly, especially in large organizations with many organizational groups, each with several members. Manually ensuring that ACLs are kept up-to-date becomes more difficult as their size increases.

Dispersion of an ACL indicates the extent to which members in the list differ from each other. Persons (user accounts) that have the need and authorization to use certain data are likely to be similar to each other. Therefore, well-maintained ACLs can be expected to typically exhibit low levels of dispersion; dispersion on ACL is thus likely to be correlated with the risk of inappropriate access, e.g., indicative of possible presence of a user account in the ACL that should not be authorized to access the data.

Managing ACLs and detecting the presence of anomalous individuals in the lists is a task performed by humans who make the decisions manually. Given the sensitive nature of the task, those managing ACLs are often highly-specialized employees, thus making the process expensive for the organization.

Whitelists and blacklists are an additional approach employed to manage ACLs. However, blacklists typically list only a minority of user accounts. While whitelists are suitable for highly sensitive data, there is a potential for individuals to gain access via trusted internal connections that may be permissive and trusting toward others, especially when providing the access is perceived as likely to result in higher productivity and effectiveness.

DESCRIPTION

This disclosure describes the application of machine learning techniques to evaluate and manage ACLs. To that end, a machine learning model is trained to evaluate ACLs based on a variety of input factors and is used to evaluate a given ACL in terms of dispersion of ACL members.

To determine ACL dispersion, a trained machine learning model can take as input a variety of information pertaining to user accounts. Permission for use of such information can be obtained via consent from organization employees, contractors, and others, e.g., as part of contractual terms, or for use of the specific system. With appropriate permissions, such information can include factors such as:

- Job title and function
- Length of employment
- Employment location
- Organizational group, e.g., product group
- Management chain
- Internal professional collaborators
- Proposed use for the data
- Sensitivity of the data
- Business importance of the data
- Number of persons included in the ACL

Relevant factors can be taken into account when determining one or more of the pieces of information above. For instance, the list of an employee’s internal professional collaborators can be constructed from shared documents and files, joint project memberships, history of message exchanges via email or other collaboration tools, etc. The relative spread of the values associated with each of the pieces of information within an ACL group are provided as input to the trained machine learning model. The output of the machine learning model can indicate a variety of dispersion measures for an ACL such as:

1. **Absolute dispersion:** The net difference across all factors among all members.
2. **Relative dispersion:** The net difference across all factors among all members relative to the number of people in the ACL.
3. **Sensitivity-weighted dispersion:** The net difference across all factors among all members weighted by the sensitivity of the data controlled by the ACL.

The trained machine learning model can employ the various dispersion measures to flag ACLs that may be problematic and warrant further examination to detect and fix potential issues. Resolving ACLs that are too permissive can be performed by breaking apart data sources that have too many accessors, removing one or more individuals from the ACL, etc.

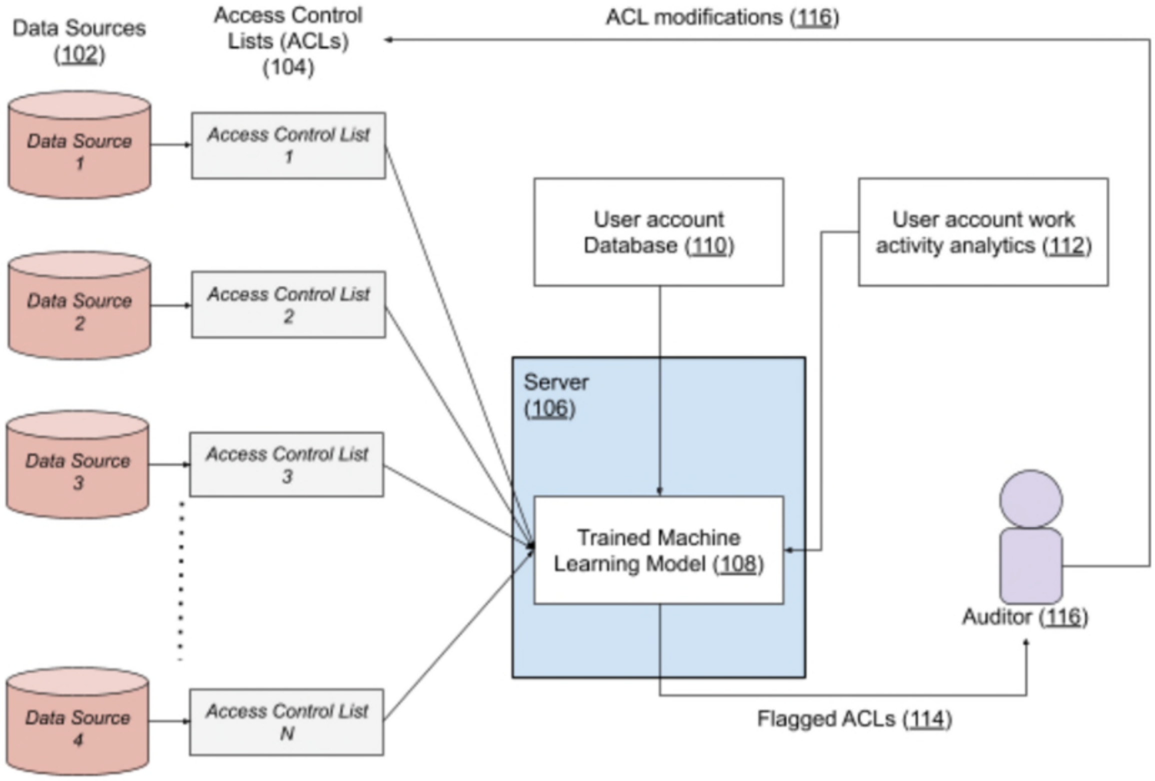

Fig. 1: Use of machine learning to flag potential risks of access control lists

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. Various data sources (102) with associated ACLs (104) are evaluated by a trained machine learning model (108) that executes on a server (106). The trained machine learning model is provided input about user accounts in the various ACLs via databases, such as a user account database (110) (e.g., that may include information about employee role, organizational structure, etc.) and analytics pertaining to employee work activities (112), obtained with user permission. The machine learning model produces as output various measures of ACL dispersion as described above, and flags a subset of the ACLs as potentially anomalous (114). The list is passed on to a human auditor (116) for further examination, e.g., to determine if any modifications to the corresponding ACL (116) are necessary. The output of the machine learning model can include specific factors such as user accounts that are anomalous for an ACL, input factors that indicate the source of anomaly, etc.

Employing a machine learning based approach allows scalable evaluation of ACLs. For example, automatic evaluation of ACLs can be performed on demand or at regular intervals (e.g., nightly). Moreover, the application of machine learning techniques can overcome the limitations of the current human-based approaches, increasing the efficiency, consistency, and fairness of the ACL evaluation and management processes, and lowering the cost. The factors used as input can be updated and the model can be retrained as the data sources and/or user accounts are updated.

The techniques can be embedded within the workflow of internal human auditors such that the audits can be performed at scale with human input reserved to handle the cases flagged by the techniques described above. The ACL evaluation functionality can be implemented as a service that can be invoked via an Application Programming Interface (API) or can be provided as a cloud-based service

CONCLUSION

This disclosure describes the application of a trained machine learning model that utilizes as input various factors such as employee roles and organization structure to detect and flag ACLs with dispersion patterns that are likely indicators that ACL edits are needed. The techniques can be utilized to flag ACLs for review by human auditors such that the audits can be performed at scale. The use of machine learning techniques can overcome limitations of the current human based approaches, and provide improvements in the efficiency, consistency, fairness, and costs of ACL evaluation and management processes and can lower organization risks related to data security. The techniques can be implemented in internal software within an organization and can also be provided as a cloud-based service for use by any organization.