Towards Automatic Ranking App Risks via Heterogenous Privacy Indicators

Deguang Kong* Lei Cen† Hongxia Jin‡

Abstract

To inform the users of dangerous levels of mobile apps, assessing privacy risks of mobile apps becomes an urgent task. This paper presents the *first* systematic study on privacy risk ranking of mobile apps via incorporating the heterogeneous privacy indicators (*i.e.*, permission access, user review, developers' description and ads library). We formalize the risk ranking problem as an optimization problem, which uses *"risk propagation"* technique to automatically rank the risks of mobile apps by considering the privacy indicators from different aspects, such that the ranking order can be automatically learned by considering data *manifold* information. Our method can automatically rank the risks of mobile apps given a few number of labeled mobile apps, The exploration on the impacts of different privacy indicators will give insight on which privacy indicators are more closely related to the privacy risks of mobile apps.

1 Introduction

Nowadays, people spend more time on using mobile apps on smart phones and tablets because of the convenience they bring to people's daily life. Personalized service (*such as* targeted advertising, personal recommendation) is possible on mobile devices when users' personal information *such as* contact and location is accessible by mobile apps. However, disclosing personal information to mobile apps could lead to serious privacy issues. Mobile app risk assessment is an effective way to display the risks of mobile apps by summarizing the information that related to the unauthorized access to users' personal information. It makes the risk transparent and warns the users of potential personal information leakage. A risk score strategy has been shown to have a "significant positive effects" [2] for users, which allows the users to better perceive the levels of security risks.

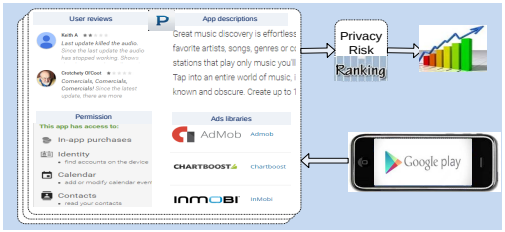

Figure 1: Motivation: Ranking the risks of mobile apps using app meta data such as *description*, *user review* and *permission access*, *ads library*. To automatically ranking *more* different mobile apps, a ranking model is proposed to capture the relations between the ranking score and privacy indicators from different aspects.

In android systems, permissions indicate the resources that the apps can access, and thus can be viewed as a privacy indicator [3]). From users' perspective, the meta data such as users' reviews and developers' descriptions reflect users' perceptions and developers' expectations for the apps, and thus are also correlated [2] with risks of apps.

Given the heterogeneous privacy indicators of mobile apps, a question that naturally follows is: *can we design an automated approach to analyze the risks of mobile apps by utilizing the heterogeneous indicators*? On one hand, there are millions of mobile apps on Google play and labeling the risk score for each mobile app is time consuming and tedious. The proposed method is required to label the risks of mobile apps efficiently and effectively when *only* a very small number of mobile app risk scores are available. On the other hand, the proposed

*Samsung Research America, San Jose, CA, US 95134, doogkong@gmail.com

†Purdue University, West Lafayette, IN 47907, lcen@purdue.edu

‡Samsung Research America, San Jose, CA, US 95134, hongxia@acm.org

approach should utilize the privacy indicators from different aspects, and make a comprehensive assessment. How to combine all the heterogeneous privacy indicators to accurately estimate the risks of apps is under-explored but highly desirable. Recent works, including permission usage pattern mining [1], app permission prediction from meta-data, mobile app recommendation, however, do not essentially solve this problem.

We propose a new approach to rank the privacy risks of mobile apps via *heterogeneous* privacy indicators. The proposed approach only requires a small number of risk score of apps labeled by experts and can automatically predict risk scores of other apps. The predicted scores can be used to improve the credibility of apps in app play store, and make the users be aware of the security risks of mobile apps.

2 Methodology

Assume we have n mobile apps, and each mobile app is abstracted as a data point $\mathbf{x}_i$ denoting the privacy indicators the app carries. We name the features extracted from v-th ($1 \leq v \leq V$) privacy indicator as the v-th aspect feature. In particular, let $\mathbf{x}_i^v \in \Re^{p_v}$ be the v-th view feature (*i.e.*, features extracted from permission, user review, *etc.*) of a mobile app i, p_v be the dimension of feature extracted from the v-th view. Consider all the mobile apps, $\mathbf{X}^v = [\mathbf{x}_1^v, \mathbf{x}_2^v, \cdots, \mathbf{x}_n^v]$, where each data column vector is $\mathbf{x}_i^v \in \Re^{p_v}$.

For the mobile app risk ranking problem, each mobile app is given a scalar value $y_i \in \Re^+$ as the risk score. Without loss of generality, we assume the risk scores for the first $\ell \ll n$ apps are already labeled by security experts, which are denoted as $\{\mathbf{x}_i, y_i\}_{i=1}^{\ell}$. The mobile app risk ranking task is to learn a function f: such that $y_i = f(\mathbf{x}_i)$, which can predict the risk scores y_i for *unlabeled*[1] mobile app $\mathbf{x}_i$ ($\ell + 1 \leq i \leq \ell + u$). The ranking/order of y_i reflects the severity of security levels for different apps. As a number of notations will be used in next sections, we summarize them in Table 1 for clarity.

Let $\mathbf{f} = [f_1, f_2, \cdots, f_n]$ be the desired risk scores[2] corresponding to apps $[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$, where $f_1 = y_1, f_2 = y_2, \cdots, f_\ell = y_\ell$ for the labeled apps. Taking all the above considerations, we propose to optimize the following objective function with respect to $\mathbf{f}$, *i.e.,*

$$\min_{\mathbf{f},\alpha} \quad \sum_{v=1}^{V} \alpha_v \mathbf{f}^T \tilde{\mathbf{L}}^v \mathbf{f} + \lambda \|\alpha\|_2^2 + \mathbf{f}^T \tilde{\mathbf{L}}^{\mathcal{W}} \mathbf{f} - \mathbf{f}^T \tilde{\mathbf{L}}^{\mathcal{S}} \mathbf{f}$$
$$s.t. \quad \alpha^T \mathbf{e} = 1; \;\; \alpha \geq 0; \;\; f_i = y_i \; (1 \leq i \leq \ell); \tag{1}$$

where V denotes the number of types of privacy indicators extracted from mobile apps. Eq.(1) consists of three parts:

(1) *risk propagation*: term $\sum_{v=1}^{V} \alpha_v \mathbf{f}^T \tilde{\mathbf{L}}^v \mathbf{f}$;

(2) *multi-view privacy indicator weight* α: term $\|\alpha\|_2^2, \alpha^T \mathbf{e} = 1, \alpha \geq 0$;

(3) *constraint* $\mathbf{f}$ *by incorporating prior knowledge*: term $f_i = y_i, \mathbf{f}^T \tilde{\mathbf{L}}^{\mathcal{W}} \mathbf{f} - \mathbf{f}^T \tilde{\mathbf{L}}^{\mathcal{S}} \mathbf{f}$, *etc.*

Table 1: Notations used in the paper

Notation	Description
$\mathbf{x}_i^v$	$\in \Re^{p_v}$, v-th view of feature
$y = [y_1, y_2, \cdots, y_i]$	$y_i \in \Re^+$, risk score for app i
$\ell; u$	# of labeled apps, # of unlabeled apps; $n = \ell + u$
α	$\in \Re^V$, contribution weight for each feature type
$\mathbf{f} = [f_1, f_2, \cdots, f_n]$	$\in \Re^n$, the desired app risk ranking score
W_{ij}^v	the similarity of app i, j in terms of v-th view indicator
$\mathbf{f}^T$	inverse of the vector $\mathbf{f}$

References

[1] M. Frank, B. Dong, A. P. Felt, and D. Song. Mining permission request patterns from android and facebook applications. pages 870–875, 12 2012.

[2] C. S. Gates, J. Chen, N. Li, and R. W. Proctor. Effective risk communication for android apps. *IEEE Trans. Dependable Sec. Comput.*, 11(3):252–265, 2014.

[3] C. S. Gates, N. Li, H. Peng, B. P. Sarma, Y. Qi, R. Potharaju, C. Nita-Rotaru, and I. Molloy. Generating summary risk scores for mobile applications. *IEEE Trans. Dependable Sec. Comput.*, 11(3):238–251, 2014.

[1] In the paper next, "*unlabeled*" refers to the apps whose risk scores are required to be labeled.

[2] For clarity purpose, we make a distinction between $\mathbf{y}$ and $\mathbf{f}$. Let $\mathbf{f}$ be the desired risk score for mobile app, but $\mathbf{y}$ only has the risk scores for the labeled apps, *i.e.,* $y_i = 0$ if $(\ell + 1) \leq i \leq n$.