Forráskód szerzőjének felismerése programozói stílus alapján

Naghi Mirtill–Boglárka, Koncsárd Tünde, Antal Margit

Sapientia Erdélyi Magyar Tudományegyetem, Marosvásárhelyi Kar, Matematika-Informatika tanszék

nagy.mirtill@gmail.com, tundekoncsard3566@gmail.com, manyi@ms.sapientia.ro

Egy forráskód szerzőjének beazonosíthatósága fontos információbiztonsági kérdés. Felhasználható programozási házi feladatok esetében a csalások kiszűrésére, de akár hackertámadás után hátramaradt kódrészletek alapján a betörő is beazonosítható.

A szerző felismerésére leggyakrabban gépi tanulási módszereket alkalmaznak [1]. A forráskódokból lexikális (pl. függvénynek olvashatósága), kinézeti (pl. kódsorok hossza), illetve szintaktikai (absztrakt szintaxisfa mélysége) jellemzőket nyernek ki, majd ezekkel betanítanak egy osztályozót. Amíg a lexikális és kinézeti jellemzőket a fejlesztési környezetek automatikus formázása jelentősen befolyásolhatja, addig a szintaktikai jellemzők invariánsok kód obfuszkációra nézve is.

Munkánk során többféle lexikális, kinézeti és szintaktikai jellemzőket nyertünk ki a programkódokból, amelyeket felhasználva különböző osztályozókat tanítottunk be a szerzők azonosítására. Méréseinket két adathalmazon végeztük: a 2015-ös Google Code Jam programozási verseny forráskódjai (1617 szerző, 9 forráskód szerzőnként), illetve a Sapientia Egyetem C++ programozás tantárgy házi feladatait felhasználva (14 szerző, 27 forráskód szerzőnként).

Kulcsszavak: biztonság, programozói stílus, gépi tanulás

Hivatkozások

[1] Aylin Caliskan-Islam, Richard Harang, Andrew Liu, Arvind Narayanan, Clare Voss, Fabian Yamaguchi, Rachel Greenstadt, De-anonymizing Programmers via Code Stylometry, 24th USENIX Security Symposium, 2015, pp. 255–270.