Les développeurs démasqués par leur empreinte dans le code

Un morceau de code anonyme peut être utilisé pour identifier son créateur, d'après **7 chercheurs** du laboratoire de recherche de l'armée américaine (ARL – Army Research Laboratory) et d'universités (Maryland, Drexel, Princeton aux États-Unis, Göttingen en Allemagne). Ils expliquent, dans une étude repérée par *Itworld*, avoir développé une méthode utilisant le traitement du langage naturel et l'apprentissage automatique. Et ce pour **identifier le développeur du code source par son style de programmation** (espace ou tabulation, bloc ou paragraphe, etc.).

Le code source, signature des développeurs

Dans leur étude intitulée « **désanonymiser les programmeurs grâce à la stylométrie du code** » (« *de-anonymizing programmers via code stylometry* »), les chercheurs déclarent analyser les caractéristiques de style classiques (mise en page, attributs lexicaux…) et les arbres de syntaxe abstraite. Ces « arbres » captent un **jeu de fonctions syntaxiques** permettant de distinguer des particularités du code qui sont indépendantes du style d'écriture. Ainsi, même si certaines variables étaient modifiées (espaces, commentaires…) pour tromper le curieux, l'auteur du code pourra toujours être identifié car le jeu de fonctions syntaxiques ne change pas.

L'approche pourrait être utilisée pour régler des différends (**droit d'auteur, code malveillant**…). « *L'attribution du code source pourrait servir de preuve au tribunal, automatiser le processus de recherche d'un cyber criminel à partir du code source abandonné dans un système infecté, ou encore aider à résoudre des problèmes de copyright, copyleft ou plagiat dans la programmation* », selon les chercheurs.

Un taux de précision d'au moins 95%

Pour tester leur approche à grande échelle, les chercheurs ont utilisé des données disponibles dans le cadre du concours de programmation Google Code Jam. Ils se sont intéressés au code source écrit en **langage C++** par plus de 100 000 participants sur la période 2008-2014. La « stylométrie du code » aurait atteint une **précision de 95%** lors d'un test effectué auprès d'un échantillon de **250 codeurs**, avec une moyenne de **630 lignes de code par développeur**. Le taux a atteint 97% lors d'une autre expérimentation menée auprès de 30 programmeurs seulement, mais avec plus de lignes de code par codeur (1 900 en moyenne). Les chercheurs ont également constaté que le taux de précision est plus élevé lorsque les tâches de programmation sont plus complexes à réaliser. Même chose avec des développeurs plus avancés, dont le style de programmation est plus « *unique* ».

Lire aussi :
Métiers du numérique : les développeurs plébiscités
50% des freelances sur Hopwork sont des développeurs

crédit photo © Julien Eichinger – Fotolia.com