Tetrane : retour d'expérience sur la construction collaborative d'un cours Software Reverse Engineering pour l'Université du Maryland (USA)

RESSI, 19 Mai 2017

Auteurs:
Frédéric MARMOND, fmarmond@tetrane.com
Benoit BRODARD, bbrodard@tetrane.com

Contributeur interne :
Louis DUREUIL, ldureuil@tetrane.com

Résumé

Introduction - La genèse du projet

La mise en oeuvre du projet
- Choix pédagogiques
- Choix des outils
- Les apports de REVEN
- Préparation du cours
- Déroulement du cours

Bilan du projet
- La suite

Résumé

Le domaine du reverse engineering logiciel doit actuellement relever un certain nombre de défis pour alimenter les administrations et entreprises en ingénieurs/chercheurs directement opérationnels et évolutifs :

1. L'élévation des niveaux de cybermenace et leur prise en compte ont créé une pénurie d'ingénieurs compétents. Les institutions et entreprises de cyberdéfense peinent ainsi à pourvoir leurs équipes et ont le besoin de former de nouveaux ingénieurs/chercheurs immédiatement.
2. Or le reverse engineering reste un sujet complexe à aborder, nécessitant beaucoup de pratique, avec de nombreux outils à appréhender, parfois coûteux. En outre, l'état de l'art évolue en permanence et nécessite des formations prenant en compte les techniques les plus récentes et permettant aux étudiants de s'adapter aux inévitables futures évolutions du métier[1].
3. Enfin, la pénurie qui touche le domaine touche également les enseignants : les acteurs du domaine sont très sollicités sur des missions opérationnelles. Les nouvelles formations doivent donc être conçues avec un effort modéré et réplicables.

Dans ce contexte, Tetrane est une entreprise française éditrice d'une solution de Reverse-Engineering logiciel proposant des fonctions de type "timeless debugging". Posant les constats ci-dessus et suite à différentes rencontres et échanges, Tetrane a souhaité s'investir avec des partenaires académiques et privés aux Etats-Unis dans la conception d'un cours universitaire d'un semestre à l'université de Maryland (USA).

Le premier semestre de cours délivré et achevé, nous présentons ici nos premiers retours d'expérience, s'agissant d'une proposition nouvelle d'enseignement du Reverse Engineering autour d'un outil d'analyse binaire mais aussi d'une construction de cours multi-fuseaux horaires, multi-secteurs (académique, conseil, industrie) et d'une collaboration franco-américaine.

D'ores et déjà, l'université du Maryland a décidé de proposer de nouvelles sessions de ce cours. Pour notre part, il apparaît que la démarche est tout à fait duplicable dans d'autres contextes.

Introduction - La genèse du projet

Tetrane est une entreprise française, conceptrice de la technologie REVEN (REverse ENgine) et éditrice d'un logiciel en Reverse-Engineering (RE) logiciel. Sa solution REVEN-Axion est utilisée par des équipes prestigieuses de cyberdéfense, et notamment au sein de ministères de défense sur 4 continents. Cependant ces équipes ont du mal à recruter des ingénieurs ayant les compétences opérationnelles nécessaire pour accomplir

[1] Le lecteur désirant approfondir ces problématiques et l'état de l'art du domaine pourra consulter avec intérêt l'article *Reverse Engineering Cognition, Maura K. Tennor, Cyber Operations Research, Lead, J52B, Internal Research and Development Portfolio, July 2015, MITRE*."

leurs missions de plus en plus pointues en reverse-engineering. Ceci impacte également le potentiel de développement de Tetrane, limité par le manque d'utilisateurs compétents.

Le domaine du RE logiciel se distingue d'autres domaines par divers aspects :

- Il est très volatile car de nouvelles protections et contournements apparaissent quotidiennement.
- L'attaque ou la défense sont souvent entremêlés techniquement. Ce domaine comporte donc une dimension juridique, avec une portée internationale et des spécificités selon le pays dans lequel le RE est effectué et selon l'origine du code analysé.
- Il revient parfois à chercher une aiguille dans une botte de foins, chaque bit étant potentiellement important. La connaissance du fonctionnement interne de blocs logiciels ou matériels est souvent primordiale, et une masse de connaissances de multiples composants en profondeur est nécessaire pour résoudre certains cas. Les experts sont souvent spécialisés par thématiques (Windows, Linux, user-land, kernel-land, etc.).
- La valeur de certains ingénieurs en RE réside dans l'exclusivité de leur savoir, l'état de l'art est donc bien souvent seulement partiellement public.

Lors d'une discussion avec le directeur du cyber-center de l'Université du Maryland-UMD (USA), ce dernier nous a confié qu'il souhaitait pousser la formation de ses étudiants pour les rendre directement opérationnels auprès de l'industrie et des équipes étatiques.

Conscients des enjeux du domaine (cf ci-dessus) et forts des capacités de la technologie REVEN de Tetrane, avec l'aide d'une entreprise américaine partenaire, nous avons proposé la construction d'une formation en Reverse Engineering à l'UMD, pour :

- Former les étudiants selon les attentes des équipes qui voudraient les recruter.
- Enrichir l'enseignement dispensé et ainsi valoriser l'université dans un environnement académique très concurrentiel aux USA.
- Faire connaître notre technologie et nos solutions.

La discussion initiale a eu lieu en octobre 2015, l'idée de monter une formation a été émise le jour même, le partenaire américain a donné son accord de principe dans la semaine. Nous avons rapidement produit un squelette du cours et dès février 2016, le cours a été programmé par UMD pour Fall 2016.

Nous avons testé le cours sur des stagiaires chez le partenaire sous la forme d'une formation interne et affiné le fond et la forme avec leurs premiers retours.

La mise en oeuvre du projet

Choix pédagogiques

Afin de maximiser l'efficacité de la formation, nous avons élaboré un premier contenu pédagogique puis nous l'avons fait valider par certains de nos utilisateurs, futurs recruteurs potentiels d'étudiants formés.

Les thèmes proposés couvrent les concepts techniques (fonctionnement interne des compilateurs et linkers, analyse statique, analyse dynamique, analyse timeless, etc.), les outils (open-source et propriétaires) utilisés par les experts dans la vraie vie, les sujets connexes indispensables à connaître pour de futurs professionnels (législation, savoir-être, éthique, rédaction de rapports, notions sur l'environnement contextuel d'une entreprise, etc.).

Une très grosse priorité a été apportée à la pratique, avec beaucoup de Travaux Pratiques (TP) et Devoirs Maison (DM). Cette volonté de notre part d'avoir un cours très pratique correspond à la spécificité du Reverse Engineering de nécessiter de balayer énormément de cas réels pour monter en efficacité. Cette approche correspond également à l'état d'esprit de la "culture hacker", parfois rebutée par l'enseignement traditionnel.

Le cours est dispensé par un consultant expérimenté et en activité, qui transmet aux étudiants autant de savoir-faire que de savoir-être. Les rapports produits par les étudiants sont ainsi évalués sur des critères réels de la profession et les exemples ou mises en situation sont issus d'expériences vécues en mission.

Choix des outils

Le RE logiciel est un domaine vaste avec de nombreux outils disponibles. Nous avons sélectionné les outils que nous allions présenter durant la formation en cherchant à couvrir un éventail représentatif de ce qui est utilisé par nos utilisateurs :

- IDA-pro, qui est la référence en analyse statique. Nous avons utilisé la version gratuite.
- gdb comme représentant des debuggers.
- REVEN-Axion pour le domaine "timeless analysis" et sa capacité à montrer précisément les interactions logiciel-logiciel et logiciel-matériel.
- Wireshark pour l'analyse des trames réseau.
- Un ensemble de petits outils très courants, tels que objdump, readelf, strings, ldd, ltrace/strace,

Afin de faciliter leur déploiement, nous avons fourni aux étudiants des machines virtuelles (VM) avec ces produits et outils installés.

Les apports de REVEN

REVEN-Axion, qui était de notre point de vue un facilitateur de la construction de la formation, s'est confirmé être un support pédagogique de premier plan.

REVEN-Axion est en effet capable d'enregistrer une trace d'exécution x86 de manière non-intrusive et de rejouer cette trace à volonté sur son simulateur, en permettant de visualiser n'importe quel bit mémoire (aux niveaux logiques et physiques) à tout instant de la trace. Les étudiants peuvent ainsi voir et "toucher" le fonctionnement du CPU, les mécanismes d'IRQ, les échanges de données sur le bus PCI ou par DMA, l'interaction des différentes couches logicielles avec le matériel, etc.

Les plugins/connecteurs de REVEN vers IDA, gdb ou Wireshark ont permi aux étudiants de prendre conscience de la complémentarité des divers outils et approches, et de commencer à prendre des réflexes qui leurs seront utiles dans leur vie professionnelle.

En terme d'infrastructure, REVEN fonctionne en mode client / serveur. Le simulateur, qui nécessite une grosse puissance de calcul, tourne sur un serveur hébergé par Tetrane. Pour une vingtaine étudiants, nous avons implémenté un serveur de 16 coeurs (32 threads) à 3GHz, 128Go RAM et 1To de SSD, hébergé au Canada (géographiquement proche de l'université pour limiter la latence réseau). Les étudiants utilisent l'interface graphique depuis leur laptop, ce qui permet entre autre de lancer des calculs en asynchrone sur le serveur et de s'y connecter de temps en temps pour voir l'avancée.
La location de cette machine coûte environ 200€ par mois. Nous avons constaté qu'elle a été sous-utilisée et qu'une configuration moins "musclée" pourrait être tout à fait suffisante.

REVEN-Axion a également apporté au cours une dimension intéressante de collaboration enseignants / étudiants et étudiants / étudiants :

- Les fonctions de gestion de projet en mode client / serveur ont permis au(x) formateur(s) de suivre et aider les étudiants au fil de l'eau en se connectant sur leurs projets.
- Certains projets ont été pré-enregistrés afin de travailler sur une trace spécifique, le but étant de préparer des traces qui montrent facilement les points sur lesquels le projet porte spécifiquement, et d'orienter les questions aux étudiants sur des endroits précis de la trace pour les faire progresser pas à pas.
- Les fonctions de collaboration fournies par REVEN ont aussi été appréciées des étudiants, qui ont pu avancer à leur rythme, de manière asynchrone, et s'entraider en se connectant ponctuellement aux projets des autres groupes. Par effet de bord, le mode hébergé où les données sont centralisées sur le serveur impose moins de pression sur le poste client (sauvegarde et persistance, possibilité de laisser le serveur travailler en tâche de fond et d'éteindre le laptop pour s'y reconnecter plus tard, etc.).

Préparation du cours

Le contenu du cours a été produit conjointement par Tetrane et l'entreprise partenaire de conseil en sécurité IT basée à Washington D.C., selon les domaines propres d'expertise de chaque entreprise. Le formateur du partenaire a été formé par Tetrane sur les compétences spécifiques en reverse-engineering qui lui manquaient.

Tetrane a ensuite assuré des sessions par visio et sur place pour transmettre son retour d'expérience au formateur.

S'agissant d'un nouveau cours, optionnel, avec un nombre de place limité, une auto-évaluation a été demandée aux étudiants intéressés afin de faire la promotion du cours par de petits challenges techniques, fun et inhabituels et d'inscrire les candidats les plus aptes.

Déroulement du cours

Une séance hebdomadaire type dure 3 heures et comporte :

- Un debrief sur la séance et les devoirs de la semaine précédente.
- L'introduction d'un nouveau concept : le formateur transmet du savoir sous forme de discussion interactive et renvoie à des papiers et articles à lire en dehors de la séance.
    - Par concept présenté dans une séance, on entend par exemple : une méthode d'obfuscation, une méthode d'attaque, un outil (IDA-pro, gdb, REVEN, etc.), le linker linux, etc. La présentation théorique ne fait qu'effleurer le sujet, qui sera découvert plus en profondeur par les étudiants par la suite.
    - De solides connaissances préalables sont nécessaires pour avancer, l'étudiant doit tirer profit de ses acquis précédents pour comprendre rapidement l'essence du nouveau concept. La présélection est donc primordiale pour ce cours.
- Un TP de prise en main du nouveau concept.
- Une mise en commun au cours de laquelle le formateur revient sur des points à préciser, ou, au contraire, pousse le sujet plus loin si les bases sont acquises.
- Une présentation des devoirs, correspondant à de "mini-missions" similaires à ce que les étudiants pourraient avoir à accomplir dans un environnement professionnel, représentant un travail personnel ou en groupe d'une durée de 4 à 6h. Chaque devoir donne lieu à un rapport et est noté.

Le semestre comprend 14 séances et est clôturé par un projet final : challenge de reverse engineering complexe comprenant plusieurs techniques de protection vues durant le semestre, à faire sur 2 semaines.

Le plan global est le suivant :

1. Review and Static Analysis – Review pointers, assembly, memory management,

binaries
2. Memory Forensics – Analysis of memory usage to determine code behavior
3. Shell Code – Shell scripting and how it can be used as malware payloads, packing, etc.; review of buffer overflows
4. Process Injection and API Hooking – Placing and finding code in the virtual space of another process
5. Memory Handling – Tracking memory accesses to determine the cause of a crash
6. Network Traces – Working with network data; introduction of Wireshark
7. Client-Server Interactions – Reversing code that sends or receives data from the network
8. Middleware and Compilers – Static and dynamically linked libraries, compiler functionality, reversing C++ code
9. Flash – Reversing Flash malware
10. Self-Modifying Code – Reversing code that alters its own instructions while executing
11. Obfuscation – Working with code that is designed to be opaque
12. Obfuscated JavaScript – Working with intentionally obfuscated JavaScript
13. Non-Technical Considerations – Understanding non-technical factors that play into the exploit and reverse engineering ecosystem (the market for zero-day attacks, vulnerability risk analysis, etc.)
14. Final Project

Bilan du projet

Pour Tetrane, la formation n'est pas (encore) notre coeur de métier et ce cours constituait une première expérience. Il convient malgré tout de garder à l'esprit que le système de formation américain est basé sur la performance :

- Les cours sont notés par les étudiants (qui s'endettent pour étudier).
- Les universités sont fortement financées par les entreprises et des mécènes.
- Les classements des universités sont donc très importants.

Le cours dispensé a donc été évalué de manière très pragmatique par les différents acteurs. Il en ressort les points suivants :

- **Un niveau technique très exigeant** (trois désistements d'étudiants sur la première quinzaine), jugé positif par le corps enseignant de l'université. La matière est pointue et challenger les étudiants pour les forcer à se dépasser est vu comme une bonne chose. Il est normal que tout le monde ne soit pas apte (ou motivé) à exceller dans cette thématique.
- **Le format adopté, très "mise en pratique"**, et les retours d'expérience de cas professionnels réels constituent un très bon complément à d'autres cours théoriques plus généralistes et abstraits proposés par l'université.
- **La collaboration entre les étudiants**, qui se sont instinctivement regroupés par petites équipes pour les TP et les DM. Cela s'est avéré très efficace pour l'auto-stimulation, la recherche de complémentarités, et au final reflète une organisation professionnelle.

- **La nécessité pour l'enseignant formateur d'investir du temps :** la préparation du cours reste chronophage et le professionnel qui le dispense doit rester joignable par mail en cours de semaine ou de week-end pour répondre aux étudiants qui seraient bloqués sur leurs devoirs. Cette contrainte, bien qu'à très forte valeur ajoutée pour les étudiants, peut être difficile à assumer pour le professionnel formateur tenu par ses missions professionnelles et ses congés / repos. Grâce à la complémentarité des fuseaux horaires américains et française, avec la participation de Tetrane, un support quasiment 24/24 a été proposé aux étudiants, qui travaillaient sur leurs devoirs à tout moment du jour et de la semaine.
- **REVEN** a été particulièrement apprécié sur la forme (aspect pédagogique) et le fond (résolution de problème), certains étudiants nous ayant déjà questionné sur les modalités pour l'utiliser lors de leur stage du printemps prochain.
- Pour TETRANE, cela a constitué un test grandeur nature de l'utilisation de notre produit en "milieu hostile" (les étudiants sont parfois 'joueurs', et leur comportement diffère de celui d'experts plus expérimentés). REVEN semble donc particulièrement bien adapté pour ce genre de cours.

Il est encore trop tôt pour avoir les retours des équipes ayant recruté les étudiants qui ont suivi ce cours. Nous espérons avoir des éléments à partager d'ici RESSI 2017.

La suite

Plus globalement, l'université nous a déjà indiqué qu'elle tirait un bilan très positif de ce cours, et nous a demandé de réitérer pour septembre 2017. Depuis, des discussions avec l'Université GeorgiaTech montre que cette méthode (beaucoup de TP/DM et utilisation de l'outil REVEN comme support pédagogique) peut également être appliquée pour des cours sur l'architecture des ordinateurs et l'ingénierie.

D'ores et déjà, nous entrevoyons des évolutions à apporter au cours d'ici la prochaine session :

- Réévaluation du contenu pour l'actualiser si besoin avec de nouvelles techniques.
- Sur la forme, l'absence de réel support de cours impose un formateur déjà aguerri ou formé par Tetrane sur les sujets qu'il va développer devant les étudiants, et disponible.