Un schéma complet de traçage de documents multimédia reposant sur des versions améliorées des codes de Tardos et de la technique de tatouage " Broken Arrows "

Fuchun Xie, Caroline Fontaine, Teddy Furon

Un schéma complet de traçage de documents multimédia reposant sur des versions améliorées des codes de Tardos et de la technique de tatouage « Broken Arrows »

Fuchun XIE[1], Caroline FONTAINE[2], Teddy FURON[3]

[1]INRIA, Centre Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes, France

[2]CNRS/IRISA et INRIA Centre Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes, France

[3]Thomson Security Lab, Cesson Sevigne, France

Fuchun.Xie@inria.fr, Caroline.Fontaine@irisa.fr, Teddy.Furon@thomson.net

Résumé – Nous présentons dans cet article un schéma complet de traçage de documents multimédia, constitué d'un code du traçage à la Tardos couplé avec une technique robuste d'insertion zéro-bit informée « Broken Arrows ». Nous proposons ici une version renforcée de « Broken Arrows »qui lui permet de résister à l'attaque de Westfeld, la plus dangereuse à ce jour. Notre schéma est particulièrement adapté pour lutter contre les attaques de type fusion : la technique d'insertion est en effet suffisamment robuste pour permettre la détection simultanée de plusieurs symboles lorsqu'une telle attaque est réalisée. Nous proposons ici une adaptation des codes de Tardos à ce contexte « multi-symboles ». Ainsi, les attaques par fusion, habituellement considérées comme difficiles à contrer, perdent ici leur efficacité : les pirates sont alors obligés de se rabattre sur des attaques plus classiques, parfaitement gérées par le code de Tardos.

Abstract – We propose in this article a complete scheme for tracing multimedia content. Our solution is based on Tardos tracing codes and a robust informed embedding technique called "Broken Arrows". Our first contribution is to reinforce this embedding technique, in order to face the strongest attack that endangered it in the past. Our scheme is particularly efficient against the fusion class of attacks: the embedding is so robust that we are able to detect several symbols at the same time. Our second contribution is to propose a variant of Tardos codes which takes this "multi-symbol" context into account. Hence, fusions are no longer efficient and colluders have to come back to more classical attacks, well tackled by Tardos codes.

1 Introduction

Nous proposons dans cet article un schéma complet de traçage de documents multimédia. Le contexte de notre travail est celui de la diffusion à la carte de documents multimédia, e.g. vidéo à la demande. Chaque utilisateur reçoit une version personnalisée du document, contenant un identifiant personnel inséré grâce à une technique de tatouage robuste. Ainsi, si une copie est rediffusée telle quelle de manière illégale, on est en mesure de remonter à l'utilisateur malhonnête.

L'attaquant est alors contraint soit d'attaquer le schéma de tatouage sous-jacent, soit de s'allier avec d'autres utilisateurs du même document pour forger une copie contenant un identifiant qui ne correspond à personne. Ces deux stratégies sont contrées lors de la conception par l'utilisation d'une technique de tatouage robuste et d'un code d'identification adéquat, dit anti-collusion. On distingue trois types d'attaques : l'échange de blocs, la fusion, et le traitement en aveugle (compression, etc). Les codes anti-collusion sont destinés à contrer la première, l'algorithme de tatouage à contrer la troisième. La seconde est plus problématique, et c'est à elle que nous nous intéressons particulièrement ici.

Notre schéma repose sur l'association de la technique d'insertion « Broken Arrows » [1] et des codes anti-collusion de Tardos [2, 3]. Nous commençons par expliquer comment articuler ces deux couches, avant d'en proposer des améliorations. Nous avons tout d'abord travaillé à l'amélioration de l'efficacité du traçage par les codes de Tardos. L'insertion est en effet ici suffisamment robuste pour donner lieu à la détection simultanée de plusieurs symboles lors d'attaques de type fusion, comme le moyennage. Or, si les codes de Tardos classiques sont particulièrement attractifs par leur faible longueur et leur efficacité, ils ne gèrent pas la détection multi-symboles. Nous en proposons ici une variante qui permet de tirer partie de cette information supplémentaire et d'accuser plus de membres de la coalition, tout en augmentant la distance de Kullback-Leibler entre les distributions des innocents et des coupables. La technique de tatouage « Broken Arrows » a quant à elle été longuement évaluée lors du concours BOWS-2 [4], et sa robustesse est forte. La seule faille mise en avant par ce concours est sa faiblesse face à l'attaque par régression linéaire proposée par A. Westfeld [5]. Nous proposons ici une variante de « Broken Arrows » qui résiste à cette attaque. Pour plus de détails concernant ces résultats nous renvoyons le lecteur à [6, 7].

2 Un schéma de traçage complet

Les codes de Tardos sont connus pour leur simplicité et leur longueur optimale [2]. Nous en utilisons ici la version q-aire proposée par Skoric *et al* [3]. Voici comment ils fonctionnent. On note $\mathcal{X}$ l'alphabet q-aire. On utilise une distribution de Dirichlet de paramètre de forme k pour tirer au sort m vecteurs de probabilités indépendants $\{\mathbf{p}_i\}_{i=1}^m$. On tire ensuite pour chaque utilisateur j les m symboles indépendants X_{ji}, avec $\text{Prob}(X_{ji} = X) = p_i(X)$, pour $X \in \mathcal{X}$. L'identifiant de l'utilisateur j est précisément le vecteur $\mathbf{X_j}$. Le document à personnaliser est découpé en blocs (images, groupes d'images, ...), et le symbole X_{ji} est inséré indépendamment des autres dans le i-ème bloc à l'aide de « Broken Arrows ». Cette technique de tatouage est dite *zéro-bit*, car elle ne porte aucun symbole, mais seulement la présence ou absence d'une marque. Pour insérer les symboles d'un alphabet q-aire, nous définissons q clefs secrètes, et utilisons la clé $K(X_{ji})$ pour insérer le symbole X_{ji}.

On suppose que l'attaque s'applique bloc par bloc, et que la collusion comporte au plus c utilisateurs malhonnêtes. Lorsque le distributeur retrouve un contenu piraté, il en extrait la suite de symboles $\mathbf{Y}$ et cherche à identifier les utilisateurs qui l'ont créé. Il commence par calculer pour chaque utilisateur j le score

$$S_j = \sum_{i=1}^{m} U(Y_i, X_{ji}, \mathbf{p}_i). \quad (1)$$

Il peut alors accuser soit tous les utilisateurs dont le score dépasse un certain seuil, soit les utilisateurs présentant les scores les plus élevés. La fonction U utilisée par Skoric *et al.* est la suivante :

$$U(Y, X, \mathbf{p}) = \delta_Y(X) g_1(p(Y)) + (1 - \delta_Y(X)) g_0(p(Y)), \quad (2)$$

avec $g_1(p) = \sqrt{(1-p)/p}$ et $g_0 = -\sqrt{p/(1-p)}$.

3 Amélioration de l'accusation

Avec ce schéma, l'insertion est tellement robuste qu'on est très souvent capables de détecter simultanément pour un bloc donné, après la fusion de plusieurs copies, plusieurs des symboles associés à ces dernières. Or le processus d'accusation proposé par Skoric ne permet pas de prendre en considération ces multi-symboles. Nous l'avons donc modifié. Notons K_i le nombre de symboles détectés simultanément dans le i-ème bloc, et $\mathcal{Y}_i = \{Y_i(1), \ldots, Y_i(K_i)\}$ la liste des symboles détectés dans ce bloc.

3.1 Deux approches

Notre première proposition est d'utiliser comme score

$$S_j = \sum_{i=1}^{m} \sum_{k=1}^{K_i} U(Y_i(k), X_{ji}, \mathbf{p}_i), \quad (3)$$

avec la fonction U définie dans (2). C'est un peu comme si la longueur du code était augmentée de m à $m\bar{K} = \sum_{i=1}^{m} K_i$, ce qui rend l'accusation plus fiable.

Notre deuxième proposition est de conserver le calcul des scores de (1) mais de remplacer la fonction U par

$$U(\mathcal{Y}, X, \mathbf{p}) = \delta_{\mathcal{Y}}(X) g_1(p_{\mathcal{Y}}) + (1 - \delta_{\mathcal{Y}}(X)) g_0(p_{\mathcal{Y}}), \quad (4)$$

avec $\delta_{\mathcal{Y}}(X) = 1$ si $X \in \mathcal{Y}$, 0 sinon, et $p_{\mathcal{Y}} = \sum_{k=1}^{K} p(Y(k))$. On a alors l'avantage de diminuer la variance des scores des coupables : quel que soit leur symbole $X_{ji} \in \mathcal{Y}_i$, ils reçoivent la même pénalisation $g_1(p_{\mathcal{Y}_i})$.

3.2 Évaluation

Les paramètres utilisés lors des tests sont $m = 300$, $q = 4$ et $c = 20$; le paramètre de forme de la distribution de Dirichlet κ varie quant à lui de $0,1$ à $0,5$.

Nos statistiques sont établies à partir de 32000 scores d'innocents et 8000 scores de coupables. Nous comparons les résultats pour le code de Tardos symétrique [3] face à un échange de blocs, pour notre premiŕe variante face à une fusion, et pour notre deuxième variante face à une fusion.

Les résultats montrent que les espérances des scores des innocents μ_I sont nulles pour les trois méthodes. Les espérances des scores des coupables μ_C de nos deux variantes sont assez similaires, et beaucoup plus élevées que celle des codes de Tardos symétriques classiques. Pour les innocents comme pour les coupables, les variances des scores (σ_I^2 et σ_C^2) obtenus avec nos variantes sont plus petites qu'avec les codes de Tardos symétriques classiques.

Mais la mesure la plus importante est la Distance de Kullback-Leibler (DKL) entre les distributions des scores des innocents et des coupables, ici égale à :

$$D_{KL}(I; C) = \frac{1}{2} \left(\frac{(\mu_I - \mu_C)^2}{\sigma_C^2} + \frac{\sigma_I^2}{\sigma_C^2} - 1 + \log \frac{\sigma_C^2}{\sigma_I^2} \right). \quad (5)$$

Plus D_{KL} est élevée, plus les innocents et les coupables sont faciles à distinguer, et donc plus fiable sera le verdict. Nos résultats sont donnés dans la figure 1. On voit clairement que nos deux variantes permettent une meilleure distinction des innocents et des coupables. Ainsi, si les pirates procèdent à une attaque par fusion, ils sont encore plus efficacement identifiés que s'ils avaient échangé des blocs ! On les incite donc à se rabattre sur cette attaque simple pour laquelle les codes anti-collusion ont été initialement conçus.

4 Rendre « Broken Arrows » résistant à l'attaque de Westfeld

L'attaque proposée par A. Westfeld [5] peut être considérée comme un processus de débruitage. Elle repose sur l'estimation de l'amplitude de chaque coefficient de la transformée en ondelettes, estimation réalisée *via* une régression linéaire portant sur les coefficients situés dans son voisinage.

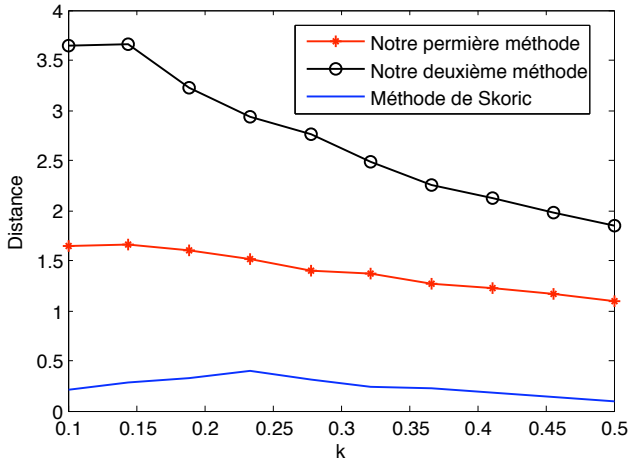

FIG. 1 – Distance de Kullback-Leibler entre les distributions des scores des innocents et des coupables, en fonction du paramètre de forme de la distribution de Dirichlet κ pour : 1) l'échange de blocs avec les codes de Skoric (bleu), 2) la fusion avec notre première méthode (rouge), 3) la fusion avec notre deuxième méthode (vert).

4.1 Le schéma « Broken Arrows » original

Voyons tout d'abord brièvement le principe de l'insertion d'un identifiant à l'aide de l'insertion proportionnelle de « Broken Arrows ». On extrait une longue séquence $\mathbf{s}_o$ de L coefficients d'ondelettes du document original. Cette séquence est divisée en m échantillons $\{\mathbf{s}_x^{(i)}\}_{i=1}^m$ de longueur l : $\mathbf{s}_x^{(i)} = (\mathbf{s}_o(il+1), \ldots, \mathbf{s}_o((i+1)l))$. Lors de l'insertion, le symbole X_{ji} est caché dans le bloc $\mathbf{s}_x^{(i)}$: $\mathbf{s}_y^{(i)} = \mathbf{s}_x^{(i)} + \mathbf{w}(X_{ji}, \mathbf{s}_x^{(i)})$, avec $\mathbf{w}(X_{ji}, \mathbf{s}_x^{(i)})$ le signal inseré. Collectant tous les blocs tatoués ensemble, on obtient la séquence tatouée $\mathbf{s}_{w,j}$, associée au j-ème utilisateur.

Le signal tatoué est en fait représenté comme $\mathbf{s}_y^{(i)} = \mathbf{s}_x^{(i)} + \mathbf{mask}^{(i)}.\mathbf{w}^{(i)}$, avec $\mathbf{w}^{(i)}$ le signal généré dans le domaine des ondelettes en fonction de la clé secrète $K(X_{ji})$, et $\mathbf{mask}$ le masque perceptuel qui module le signal $\mathbf{w}^{(i)}$. Dans le schéma original, on a $\mathbf{mask_{BA}} = |\mathbf{s}_x^{(i)}|$, où $|\mathbf{s}_x^{(i)}|$ désigne la valeur absolue des coefficients du signal hôte $\mathbf{s}_x^{(i)}$, qui sont les coefficients choisis parmi toutes les sous-bandes, sauf la sous-bande de basse fréquence LL.

Une telle insertion donne lieu à un PSNR acceptable pour les images tatouées, au-dessus de 40 dB dans nos expériences. Avec un tel PSNR, il semble que les amplitudes des échantillons de $\mathbf{w}^{(i)}$ sont presque toutes inférieures à 1. Par conséquent, « Broken Arrows » conserve les signes des coefficients. Ainsi, une attaque qui ne modifie pas les signes des coefficients conserve automatiquement cette « composante » du contenu original, et si les amplitudes des coefficients attaqués sont suffisamment modifiées, la marque ne peut plus être détectée. C'est ce qui se produit avec l'attaque de Westfeld. Nous allons maintenant présenter l'amélioration de « Broken Arrows », AWC, qui consiste en la définition d'un nouveau masque, $\mathbf{mask_{AWC}}$.

4.2 Insertion proportionnelle AWC (Averaging the Wavelet Coefficient with four neighbouring coefficients in the same subband)

Pour contrer l'attaque de Westfeld, on peut choisir de prendre en compte la dépendance entre les coefficients voisins lors de l'insertion. Nous proposons de remplacer chaque coefficient (sauf dans la sous-bande LL) par une moyenne de cinq coefficients : lui-même $\mathbf{s}_x^{(i)}(m,n)$ et ses quatre voisins $\mathbf{s}_x^{(i)}(m-1,n)$, $\mathbf{s}_x^{(i)}(m,n-1)$, $\mathbf{s}_x^{(i)}(m+1,n)$, et $\mathbf{s}_x^{(i)}(m,n+1)$. Nous obtenons alors le masque

$$\mathbf{mask}_{\mathbf{AWC}}^{(i)}(m,n) = \frac{1}{5}\left|\sum_{l=m-1}^{m+1}\sum_{s=n-1}^{n+1} \mathbf{s}_x{}^{(i)}(l,s)\right|. \quad (6)$$

En collectant tous les $\mathbf{mask}_{\mathbf{AWC}}^{(i)}(m,n)$ ensemble, nous obtenons le masque $\mathbf{mask}_{\mathbf{AWC}}^{(i)}$ de l'insertion proportionnelle AWC.

Intuitivement, cette insertion renforce la dépendance entre les coefficients voisins du signal tatoué. Pour une position donnée, le masque a une valeur plus grande que l'amplitude d'au moins un des cinq coefficients considérés. Selon la valeur du signal tatoué, l'insertion pourrait par conséquent modifier le signe du coefficient d'ondelette. Ainsi, la présence de la marque est non seulement cachée dans les amplitudes des coefficients, mais également dans certains de leurs signes. Dans nos expériences, environ 6% des coefficients d'ondelette voient leur signe modifié par cette nouvelle insertion.

4.2.1 Résistance aux attaques habituelles

Nous avons tout d'abord confronté AWC à la série de tests de robustesse utilisée lors de l'évaluation du schéma original [1]. Ces tests ont porté sur le même ensemble de 2000 images, et ont consisté en des attaques résultant de combinaisons de compressions JPEG ou JPEG 2000 aux facteurs de qualité variés, avec des changements d'échelle. La Figure 2 indique l'impact des 15 attaques les plus significatives, pour les deux techniques d'insertion. La probabilité de bon détection du marque est tracé par rapport à la moyenne du PSNR des images attaquées.

Puisque les attaques produisent presque le même PSNR moyen pour les deux insertions, les deux points correspondant à une attaque donnée sont presque verticalement alignés. L'insertion proportionnelle AWC est plus robuste face aux attaques 9-14, mais moins robuste face aux attaques 2, 5 et 6. Pour les attaques 1, 3, 4, 7, et 15, les deux techniques d'insertion offrent une robustesse comparable. Globalement, on peut considérer que notre nouvelle insertion n'est pas moins robuste que l'insertion originale face à ces attaques génériques.

4.2.2 Résistance à l'attaque de Westfeld

A. Westfeld a utilisé dans ses expérience un ensemble de 10000 images, incluant les 2000 images que nous avons utilisées dans toutes nos expériences [5]. Néanmoins, par souci

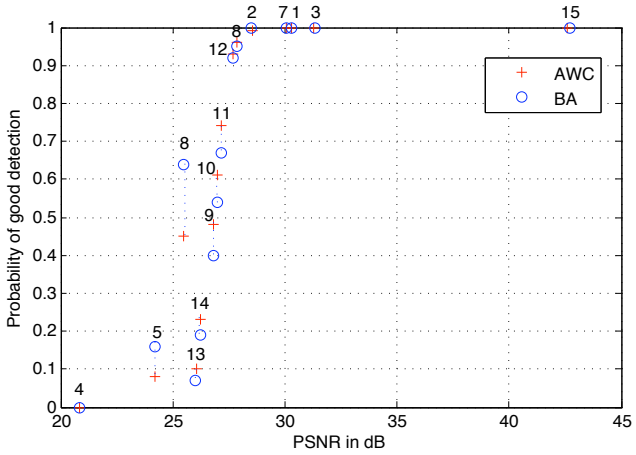

FIG. 2 – Probabilité de bonne détection contre PSNR moyen des images attaquées pour les deux techniques d'insertion de tatouage : l'insertion proportionnelle AWC '+', et l'insertion proportionnelle BA 'o'. Sélection des attaques : 1) Débruitage seuil 20 ;) Débruitage seuil 30 ; 3) JPEG Q = 20 ; 4) JPEG2000 r = 0.001 ; 5) JPEG2000 r = 0.003 ; 6) JPEG2000 r = 0.005 ; 7) l'échelle 1/2 ; 8) l'échelle 1/3 ; 9) l'échelle 1/3 + JPEG Q = 50 ; 10) l'échelle 1/3 + JPEG Q = 60 ; 11) l'échelle 1/3 + JPEG Q = 70 ; 12) l'échelle 1/3 + JPEG Q = 90 ; 13) l'échelle 1/4 + JPEG Q = 70 ; 14) l'échelle 1/4 + JPEG Q = 80 ; 15) aucune attaque.

d'homogénéité avec les tests décrits au paragraphe précédent, nous avons conservé cet ensemble de 2000 images pour les tests liés à l'attaque de Westfeld. Cette différence donne lieu à de légères variations entre ses résultats et les notres, mais sans conséquence sur les conclusions de ce travail.

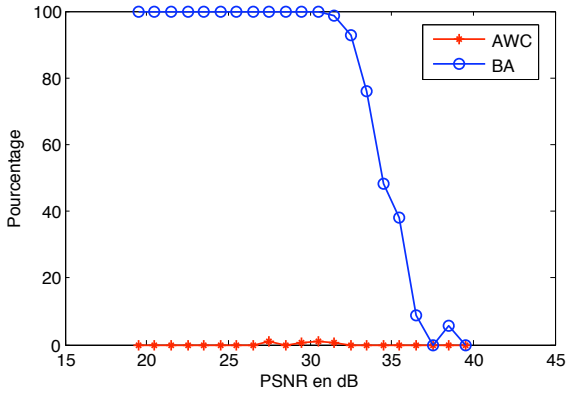

FIG. 3 – Résistance à l'attaque de Westfeld : pourcentage d'images attaquées avec succès.

Le PSNR des images attaquées s'étend de 19.9 à 46.2 dB (il varie de 19.7 à 45.0 dB dans [5]). La figure 3 montre la diminution du pourcentage des images attaquées avec succès quand le PSNR augmente. Pour l'insertion proportionnelle BA, l'attaque de Westfeld est vraiment dévastatrice, avec un succès de 100% pour les images lorsque le PSNR est inférieur à 30 dB, et même si son efficacité diminue lorsque le PSNR s'augmente, elle reste très efficace pour 40% des images quand le PSNR est autour de 35 dB. Cette figure montre aussi l'aptitude de notre variante AWC à contrer l'attaque de Westfeld. Le pourcentage des images attaquées avec succès est presque à 0 pour n'importe quel PSNR. Ainsi, AWC remplit parfaitement son rôle : offrir, comme « Broken Arrows », une très bonne robustesse générale, tout en résistant efficacement à l'attaque de Westfeld.

4.3 Conclusion

Nous avons proposé dans cet article un schéma complet de traçage de documents multimédia, constitué d'un code du traçage à la Tardos couplé avec une technique robuste d'insertion zéro-bit informée « Broken Arrows ». Nous avons proposé une version renforcée de « Broken Arrows »qui lui permet de résister à l'attaque de Westfeld, la plus dangereuse à ce jour. Notre schéma est particulièrement adapté pour lutter contre les attaques de type fusion : la technique d'insertion est en effet suffisamment robuste pour permettre la détection simultanée de plusieurs symboles lorsqu'une telle attaque est réalisée. Nous proposons ici une adaptation des codes de Tardos à ce contexte « multi-symboles ». Ainsi, les attaques par fusion, habituellement considérées comme difficiles à contrer, perdent ici leur efficacité : les pirates sont alors obligés de se rabattre sur des attaques plus classiques, parfaitement gérées par le code de Tardos.

Remerciement : Les auteurs remercient P. Bas pour ses suggestions et discussions concernant « Broken Arrows ».

Références

[1] T. Furon and P. Bas. Broken arrows. *EURASIP Journal on Information Security*, 2008.

[2] G. Tardos. Optimal probabilistic fingerprint codes. *Proc. of the 35th annual ACM symposium on theory of computing*, pages 116–125, 2003.

[3] B. Skoric, S. Katzenbeisser, and M. Celik. Symmetric Tardos fingerprinting codes for arbitrary alphabet sizes. *Designs, Codes and Cryptography*, 46(2) :137–166, February 2008.

[4] BOWS-2. http ://bows2.gipsa-lab.inpg.fr/. 2007.

[5] A. Westfeld. A regression-based restoration technique for automated watermark removal. *Proc. of 10th ACM Multimedia and Security Workshop, Oxford, UK*, September 2008.

[6] F. Xie, T. Furon, and C. Fontaine. On-off keying modulation and tardos fingerprinting. *Proc. of 10th ACM Multimedia and Security Workshop, Oxford, UK*, September 2008.

[7] A. Charpentier, F. Xie, T. Furon, and C. Fontaine. Expectation maximisation decoding of tardos probabilistic fingerprinting code. *Proc. of SPIE on Media Forensics and Security XI, San Jose, California, USA*, January 2009.