Paris le 30 avril 2015

Éléments d'analyse technique du projet de loi relatif au renseignement

Dans le contexte de numérisation globale de l'ensemble des activités humaines et des informations qui leurs sont associées, la détection d'activités terroristes, dont l'extrême gravité est avérée, motive tout particulièrement ce projet de loi relatif au renseignement.

L'analyse qui suit a pour objectif de donner un point de vue technique sur les approches numériques proposées dans ce projet de loi, en se basant sur les compétences en sciences et technologies du numérique d'Inria[1]. Cette analyse est complémentaire de celles, de grande qualité, qui ont été conduites par la CNIL, le Conseil National du Numérique et la Commission de réflexion et de proposition sur le droit et les libertés à l'âge numérique de l'Assemblée nationale.

L'analyse que nous menons nous amène à attirer l'attention du législateur sur les principaux éléments suivants :

- **L'anonymisation est un problème de recherche largement ouvert et il n'existe pas aujourd'hui de technique d'anonymisation sûre.** Un texte de loi ne devrait donc pas se fonder sur la notion de donnée anonyme ou anonymisée mais plutôt utiliser les notions de données pseudo-anonymes ou encore de données personnelles ;
- **Les méthodes proposées d'observation des connexions numériques ou des contenus sont facilement contournables même sans connaissance technique informatique élaborée ;**
- **Le paradoxe statistique des « faux-positifs » doit être parfaitement compris lors du traitement automatique d'information.** Ce paradoxe conduit à devoir effectuer le plus souvent des traitements intrusifs de masse, formellement inopérants en pratique et pouvant conduire à des erreurs de classification avec des conséquences potentielles sérieuses ;
- **L'accessibilité des données et méta-données nécessaires au croisement d'informations issues de bases de données est fortement limitée, car elles sont le plus souvent localisées dans des centres de traitements très majoritairement situés hors du territoire français.**

[1] Inria, institut de recherche dédié au numérique, promeut « l'excellence scientifique au service du transfert technologique et de la société ». Inria emploie 2700 collaborateurs issus des meilleures universités mondiales, qui relèvent les défis des sciences informatiques et mathématiques. Son modèle ouvert et agile lui permet d'explorer des voies originales avec ses partenaires industriels et académiques. Inria répond ainsi efficacement aux enjeux pluridisciplinaires et applicatifs de la transition numérique. Inria est à l'origine de nombreuses innovations créatrices de valeur et d'emplois.

Domaine de Voluceau - Rocquencourt BP 105
78153 Le Chesnay Cedex France
Tél. : +33 (0)1 39 63 55 11
Fax : +33 (0)1 39 63 53 30

www.inria.fr

Nous développons ces points dans la suite de ce document.

L'anonymisation

L'anonymisation consiste à modifier un ensemble de données de manière à ce qu'on ne puisse pas identifier un ou plusieurs critères particuliers qui leur sont initialement attachés tels que l'identité de personnes, la localisation de faits, l'entité ayant recueilli les informations, etc.
L'anonymisation est aujourd'hui le sujet de nombreux travaux de recherche visant soit à augmenter son efficacité, soit au contraire à évaluer sa résistance à la dés-anonymisation. Les techniques d'anonymisation sont variées et pour certaines particulièrement sophistiquées, consistant par exemple à bruiter les données de manière appropriée. Mais aucune de ces techniques ne résiste actuellement de manière robuste au croisement des sources d'information. Par exemple, le croisement d'un fichier concernant, dans un hôpital, un ensemble de patients dont on a supprimé les informations nominatives (et donc a priori « anonyme ») avec les informations temporelles ou de localisation de personnes accédant à l'hôpital permet très largement de des-anonymiser ce fichier patient.
Par conséquent un texte de loi ne devrait pas se fonder sur la notion de donnée anonyme ou anonymisée, mais parler plutôt de données pseudo-anonymes ou encore de données personnelles.

L'effectivité

Les mesures proposées par le projet de loi consistent en particulier à rendre possible la collecte de données dans les entités sous juridiction française, en particulier localisées sur le territoire français. Typiquement les données de connexion d'un dispositif informatique (ordinateur portable ou pas, smartphone, tablette, dispositif de régulation ou industriel, etc.) situé sur le territoire français vont nécessairement passer par un opérateur ou un hébergeur de juridiction française.
Ces collectes de données peuvent être facilement contournées, masquées ou perturbées, voici trois manières simples et efficaces de le faire.

La collecte de données de connexion est contournable.

Cette collecte peut être facilement évitée en utilisant par exemple une connexion chiffrée vers le serveur d'un opérateur ou d'un hébergeur extérieur à la juridiction française. Techniquement cela se réalise en utilisant un tunnel de communication sécurisé de type VPN (Virtual Private Network) ou la mise en place d'une communication chiffrée avec un proxy hors juridiction française. Dans tous les cas, ces contournements techniques sont faciles à mettre en œuvre et la seule information exploitable sera l'établissement d'une connexion chiffrée entre une machine et un serveur étranger. En particulier, aucune information sur le destinataire final de l'information ou le contenu du message ne sera possible dans ce cas.

Le contenu des communications est chiffrable

Indépendamment des données de connexion, le contenu des messages ou des informations échangées peut être chiffré avec des programmes librement disponibles et très bien documentés. Leur utilisation permet, dans l'état actuel de nos connaissances, et en l'absence d'autre méthode d'acquisition de renseignement (eg par canal auxiliaire) de préserver l'information de manière forte. On estime par exemple actuellement que déchiffrer une information chiffrée en utilisant le protocole RSA avec des clefs de 2048 bits prendrait plusieurs milliers d'années sur les plus gros calculateurs actuels.

Des contremesures simples peuvent être développées

La collecte d'information basée sur l'accès à des sites considérés comme suspects peut être fortement perturbée ou brouillée par des contremesures consistant à faire accéder massivement à ces sites suspects des internautes sans intention malveillante. Des entités malveillantes créent ainsi indirectement des attaques de type DOS (Denial Of Service) consistant à noyer les sites

d'observations par des accès massifs créés intentionnellement. Techniquement on pourra pour cela soit utiliser des réseaux de botnets ou des techniques de spam, faisant cliquer sans intention malicieuse des internautes sur des adresses numériques de sites considérés comme suspects.

Le paradoxe des « faux-positifs »

Le traitement par des programmes informatiques des données collectées, en particulier évoqué dans l'article L. 851-4 du projet de loi, doit faire l'objet d'analyses formelles correctes. En particulier nous attirons l'attention du législateur sur ce que les statisticiens appellent le paradoxe des faux-positifs. Son principe est le suivant. Supposons que l'on recherche des terroristes dans une population. Tout algorithme de détection a une marge d'erreur c'est à dire va identifier des personnes sans intention terroriste (des « faux-positifs »). Si la marge d'erreur est de 1%, ce qui est considéré à ce jour comme très faible, l'algorithme identifiera quelques 600 000 personnes sur une population totale de 60 millions de personnes. Si le nombre de vrais terroristes est par exemple de 60, ces vrais terroristes ne représenteront que 0,01% de la population identifiée.
Ce phénomène scientifique bien connu et lié à l'identification statistique d'évènements rares a donc des conséquences que le texte du projet de loi actuel ne prend pas en compte. Raffiner les informations obtenues dans un tel contexte peut se faire en croisant les sources d'informations, avec les limitations actuelles que nous évoquons dans le paragraphe suivant.

Le croisement d'informations

Le croisement d'informations, basée sur le croisement de bases d'informations de nature variées, incluant de manière fondamentale les méta-données, est aujourd'hui très efficace. L'acquisition quasi systématique d'information faites par les applications ou les sites commerciaux, le traçage des activités des internautes permettent d'obtenir des profilages précis des utilisateurs. Ces traçages constituent un souci important quant au respect de la vie privée qui, comme le rappelle le premier article du projet de loi, est un droit fondamental qui concerne le secret des correspondances et l'inviolabilité du domicile, et auquel il conviendrait d'ajouter le secret des méta-données.
Si ce type de techniques peut certainement être efficace à des fins de renseignement il nous semble très difficilement utilisable dans le cadre de ce projet de loi. En effet ces techniques s'appuient sur des informations (données, méta-données,...) qui sont acquises et stockées quasi-exclusivement en dehors du territoire ou de la juridiction française par Google, Bing, Facebook, Twitter, Amazon....
Bien entendu, il sera possible de contraindre les entreprises ou les services relevant de la juridiction française, mais avec le risque fort de mettre à mal leur compétitivité qui ne se comprend qu'à l'échelle de la planète.

Conclusion

Inria est à la disposition du législateur pour détailler les éléments techniques ci-dessus et l'aider à évaluer les éléments scientifiques et techniques des textes juridiques que celui-ci jugera appropriés.

Une remarque complémentaire concerne la composition de la Commission nationale de contrôle des techniques de renseignement (CNCTR). Il nous semble en effet que compte tenu de la complexité scientifique et technique des sujets numériques abordés, la CNCTR bénéficierait d'une représentation équilibrée entre les compétences numériques et juridiques. Dans ce cadre, les statuts de la CNCTR pourraient prévoir la nomination de membres par l'ARCEP, la CNIL et Allistene, l'alliance des organismes, universités et écoles en sciences et technologies du numérique.