Proteção de dados na UE: a certeza da incerteza

*Por **Cory Doctorow**, escritor, ativista, jornalista e blogueiro, coeditor do portal Boing Boing, ex-diretor da Electronic Frontier Foundation e cofundador do Open Rights Group da Inglaterra*

Data da publicação:

Novembro de 2013

Quando uma regulamentação afirma que algum dado é “anônimo”, ela está desconectada das melhores teorias da ciência computacional. No momento em que escrevo, o Parlamento Europeu está envolvido numa acirradíssima disputa mundial sobre a nova Regulamentação Geral para a Proteção de Dados[1]. Estão em jogo as futuras regras para privacidade online, mineração de dados, big data[2], publicidade dirigida, ciências sociais guiada por dados (data-driven social sciences), espionagem governamental (via proxy) e milhares de outras atividades que se encontram no cerne de muitas das maiores empresas da internet, e das ambições mais obscuras e descontroladas de nossos políticos.

Os lobistas estão a todo vapor. Os ativistas que conheço e sei que vão a Bruxelas dizem que nunca viram algo assim: é o verdadeiro frenesi do lobby. Há na mesa centenas de emendas e propostas, algumas boas, outras ruins, e só para tomar pé de todas elas já exige trabalho em tempo integral.

Por mais complicadas que sejam as propostas, existe uma regrinha básica que devemos ter sempre em mente quando há na mesa alguma proposta para proteção de dados: sempre que alguém fala em relaxar as regras sobre compartilhamento de dados que tenham sido “anonimizados” (dos quais foram retiradas as informações de identificação) ou “pseudonimizados” (cujos identificadores foram substituídos por pseudônimos), devemos assumir, enquanto não houver provas, que esse alguém está dizendo besteira.

Trata-se de uma “lei férrea da privacidade”, que pode ser usada para descartar rapidamente as ideias sem sentido. O que sobra podem ser boas ou más ideias, porém, pelo menos não estarão baseadas em uma quase-impossibilidade.

Anonimizar dados é um negócio bastante difícil. Nesse quesito, há três falhas notórias que são muito citadas: o lançamento que a AOL fez em 2006 de pesquisa anônima na internet (anonymous search data); o lançamento feito pela Comissão de Seguros em Grupo do Estado de Massachusetts de cadastros de saúde anonimizados (anonymised health records); e o lançamento do acervo de 100 mil vídeos para aluguel que a Netflix fez em 2006.

Em cada um desses casos, os pesquisadores mostraram como se pode usar algumas técnicas relativamente simples para re-identificar os dados nesses conjuntos, normalmente escolhendo em cada registro os elementos que os tornam únicos. Há vários fumantes nos registros de saúde, mas quando se restringe a busca a um anônimo fumante negro que nasceu em 1965 e se apresentou na ala de emergência com dor nas articulações, na verdade é bastante simples fazer uma fusão do registro "anônimo" com outro banco de dados "anonimizado", de onde surgirá a identidade quase certeira do paciente.

DESANONIMIZAÇÃO

Desde meados da década de 2000, a desanonimização se tornou algo como um esporte de contato para os cientistas da informática, que vivem tirando da cartola esquemas de anonimização com espertos truques de reidentificação. Um artigo publicado recentemente na Nature Scientific Reports[3] mostrou como os dados "anonimizados" da empresa telefônica europeia (provavelmente uma na Bélgica) poderiam ser re-identificados com 95% de precisão, a partir de apenas quatro itens sobre cada pessoa (com apenas duas informações, mais da metade dos usuários do conjunto de dados poderiam ser re-identificados).

Há quem diga que isso não importa. Para essas pessoas, a privacidade morreu, ou é irrelevante, algo sem importância. Se você concorda com elas, lembre-se: a razão pela qual a anonimização e pseudonimização estão sendo contempladas na Regulamentação Geral para Proteção de Dados é que os próprios autores dizem que a privacidade é importante, e que vale a pena preservá-la. Falam sobre anonimização de conjuntos de dados porque acreditam que ela seja capaz de proteger a privacidade – e isso significa que estão dizendo, implicitamente, que vale a pena preservar a privacidade. Se for essa a meta das políticas públicas, então essas políticas devem procurá-la de maneira conforme à realidade que compreendemos.

De fato, toda a premissa básica dos big data está em risco com a ideia de que os dados podem ser anonimizados. Afinal, os big data prometem que, com os grandes conjuntos de dados, relacionamentos sutis podem ser destrinchados. No mundo da reidentificação, fala-se de abordagens de "dados esparsos" (sparse data) para a desanonimização. Embora a maior parte dos seus traços pessoais seja compartilhada com muitos outros, há certas coisas a seu respeito que se encontram menos comumente representados no conjunto – talvez a confluência dos seus hábitos de leitura com o seu endereço; talvez a sua cidade natal em combinação com as suas opções de carros. Essas raridades praticamente saltam dos dados e apontam diretamente para você, assim como se prontificam a fazer as demais conclusões dos big data. Se os big data conseguem encontrar a combinação de fatores ambientais sutis em comum para todas as vítimas de uma doença rara, eles também devem ser capazes de encontrar a combinação de identificadores sutis compartilhados com todos os diferentes conjuntos de dados nos quais você está presente, de fundi-los e de trazer a público a sua identidade.

FRENESI LOBISTA

A UE está sofrendo um lobby como nunca se viu igual. A Comissária Viviane Reding, da UE, diz: "Nunca vi na minha vida uma operação de lobby tão pesada!"[4] E está funcionando!

Uma quantidade imensa de textos escritos por lobistas está chegando até as emendas dos MPEs[5]. Os lobistas se tornaram legisladores de fato, só que recebem mais e não precisam comparecer a todas aquelas enfadonhas reuniões.

A cláusula quatro da Regulamentação Geral para Proteção de Dados contém definições usadas no documento, e trata-se de um dos principais campos de batalha. Ela estabelece a ideia de que existem dados "anônimos" e os isenta de regulamentação, e cria uma segunda categoria de informações "pseudônima" que pode ser tratada com menos restrições do que impostas às "informações de identificação pessoal".

Fui a dois dos meus cientistas da computação favoritos e lhes perguntei o que achavam da plausibilidade da anonimização ou pseudonimização de conjuntos de dados. Seth David Schoen (tecnólogo da Electronic Frontier Foundation) disse-me: "Os pesquisadores já mostraram que a anonimização é muito mais difícil do que parece. Só porque uma coisa parece anônima num primeiro olhar, não quer dizer que de fato seja; tanto por causa da matemática da distinção individual quanto por causa da quantidade imensa de bancos de dados que estão se tornando disponíveis. isso significa que devemos ser extremamente cautelosos quanto ao anonimato das coisas; não devemos nos fiar somente na nossa intuição.".

Ed Felten, que saiu da Comissão Federal dos EUA para o Comércio e agora está em Princeton, disse: “Uma década inteira de pesquisas da ciência da computação mostra que muitos conjuntos de dados podem ser re-identificados. Não basta remover os identificadores óbvious para evitar a reidentificação. Pode ser que não baste remover todos os dados sobre indivíduos. Até os conjuntos de dados totalmente compostos de informações agregadas podem ser usados, em alguns casos realísticos, para inferir informações sobre indivíduos específicos.

“Mas dizer que não existe a menor esperança para a reidentificação é ir um pouco longe demais. Existe uma ciência emergente da análise de dados que preservam a privacidade, que pode ser aplicada em alguns ambientes. Via de regra, dados oriundos das características dos indivíduos, inclusive os comportamentais, provavelmente irão passar informações sobre esses indivíduos, na ausência de uma rígida base técnica para se crer que não.

“A tendência é no sentido de tratar o assunto como criptografia, onde não vale o argumento de que ‘misturei os dados um bocado’ nem o de que ‘não consigo nem pensar num ataque’ – é preciso um argumento tecnicamente rigoroso de que um ataque é uma coisa impossível.”

Como se pode ver, ambos tomaram o cuidado de não eliminar a possibilidade de que alguém possa um dia apresentar um esquema de anonimização, mas tampouco se lançaram a criar uma categoria regulatória de dados “anônimos” que possa ser tratada como se não apresentasse riscos para as pessoas das quais eles foram coletados.

Pedi que ambos me indicassem uma leitura mais profunda sobre o assunto. Felten sugeriu “Privacy and Security Myths and Fallacies of ‘Personally Identifiable Information”, de Arvind Narayanan e Vitaly Shmatikov[6], excelente iniciação sobre as questões técnicas tiradas das Communications of the Association for Computing Machinery de junho de 2010. Shoen recomendou “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization”[7], de Paul Ohm, uma abrangente resenha jurídica sobre a ideia de anonimização na regulamentação publicada numa edição da UCLA Law Review de 2010.

Da minha parte, recomendo “On the Feasibility of User De-Anonymization from Shared Mobile Sensor Data”[8], um olhar fantástico (ainda que um tanto técnico) sobre as inferências de reidentificação que podem ser tiradas dos aparentemente inócuos dados de sensoriamento (sensor-data) que saem dos nossos telefones móveis, da Ata da Terceira Oficina Internacional sobre Aplicações de Sensoriamento nos Telefones Móveis, de 2012.

PRIVACIDADE DIFERENCIAL

A Microsoft tem feito pressão no sentido de uma abordagem que chamam de “privacidade diferencial”, e parece que podem cumprir com o prometido. Conforme Schoen descreve: “Os pesquisadores fazem as perguntas de suas pesquisas ao controlador original de dados, que devolve respostas intencionalmente distorcidas/ corrompidas, e pode-se dizer que dá para quantificar matematicamente o mal causado à privacidade no processo, discutindo depois se valeu a pena diante dos benefícios da pesquisa.”

Mas tudo isso são conjeturas: embora a quantidade de “distorção” dos dados seja uma questão quantitativa, o grau de proteção propiciada à sua privacidade pela distorção é, em última análise, uma questão pessoal, voltando-se à maneira como você se sente diante da divulgação e das suas consequências. Como costuma ser o caso, essa solução técnica incorpora um monte de premissas sobre questões que são sociais, em última instância, e calorosamente contestadas. Não se pode calar o argumento de que sua privacidade está sendo ou deixando de ser violada só com matemática.

É fascinante pensar nisso tudo, mas a maior dimensão é a seguinte: quando uma regulamentação vem faceiramente determinar que alguns dados são “anônimos” ou mesmo “pseudônimos”, essa regulamentação está gritantemente desconectada das melhores teorias de que dispõe a ciência da computação. Quando se encontra algo assim numa regulamentação, sabe-se logo que o autor não tratava a proteção da privacidade com seriedade ou não era qualificado para redigir uma regulamentação. De qualquer forma, é causa para alarme.

---

1. Ver http://www.guardian.co.uk/technology/data-protection

2. Ver http://www.guardian.co.uk/technology/big-data

3. Ver http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

4. Ver http://www.telegraph.co.uk/technology/news/9070019/EU-Privacy-regulation...

5. Ver http://www.motherjones.com/politics/2013/03/google-facebook-sopa-privacy

6. Ver http://www.cs.utexas.edu/users/shmat/shmat_cacm10.pdf

7. Ver https://papers.ssrn.com/sol3/papers.cfm?abstract_id1450006

8. Ver http://niclane.org/pubs/lane_phonesense.pdf

Categoria:

- poliTICs 16