美国国安局如何利用“元数据”找出恐怖分子

最近一段时间,不管是美国国家安全局的批评者还是维护者都在热烈讨论“元数据”——即美国国家安全局通过“棱镜”计划等秘密监控手段为美国政府搜集来的公众的电话、网络等个人信息。

美国国家安全局收集了电信巨头威瑞森以及其他所有美国电话运营商旗下所有用户的一切通话记录:包括通话的日期、时长以及通话的参与者等等。

一些人认为,搜集这些数据严重侵犯了美国公民的隐私权。另一些人认为,只要有助于美国免受恐怖主义威胁,隐私问题可以忽略不计。

美国《外交政策》杂志发表文章,详细解释了“棱镜”项目是如何帮助美国政府防范恐怖袭击的。

现代快报记者 潘文军 编译

美国国家安全局内一片繁忙景象

起点

“基地”筹款人和79个联系号码

这里的确涉及一些个人隐私问题,但可能没有媒体报道的那么夸张。这里采用的样本数据来自一个真实的社交网络。样本数据并非来自电话记录,但在分析这些数据时已经能够让人产生个人隐私被曝光的感觉。

虽然举这个例子是为了说明美国国家安全局里究竟发生了些什么,但是真相不是用文字就能说得清的。这已经是一个非常接近现实情况的例子,但大量的假设和保密程序使得事实与这个例子还是有所不同。

让我们从一个经典的场景开始。美国情报官员截获“基地”组织的一个关键暗语,并且获得了也门“基地”组织筹款人的电话号码。

假设你是国家安全局的分析师,你被授权通过“元数据”找出这个“基地”筹款人的社会关系网络,你的起点就是这个电话号码。

第一步很简单:你将这个筹款人的电话号码输入“元数据”分析软件,并单击“确定”。

在我们所举例子的数据中,你会得到79个在过去30天与“基地”筹款人有过联系的电话号码。由于“基地”筹款人的活动相当隐秘,他所使用的这个手机号码也是专门为秘密活动准备的,所以几乎任何一个拨打这个号码的人都是有很高情报价值的目标。

通过“元数据”,我们可以称量出每个与这个号码联系的电话号码的分量——通话的时间长短、联系人所在位置以及通话的时间段都可以帮助判断联系人与这个“基地”筹款人的关系。美国国家安全局的培训手册中详细地列出了不同数据代表的这个联系人的不同的威胁级别。你将要画出数据曲线图,每个点代表一个电话号码,点越大、代表威胁“分数”越高。

对于情报搜集者来说,这几乎等同于天降横财,你获得这一切所需不过5分钟。现在,你可以回到“元数据”中,去查出这79名与“基地”筹款人通过电话的人中有没有互相通过电话的。

于是,通过一种常用的数学计算方法,每个电话号码都可以被算出在这个网络中的重要程度,高得分号码的主人显然在这个社会网络中是更为重要的角色,虽然我们不能将其等同于这个网络中更重要的恐怖分子。

通过搜索你会发现,这个网络中的许多人会互相通话,他们并不仅仅和“基地”筹款人通话。这表明他们可能在协调自己的行动,也许他们的通话中就涉及到与“基地”有关的事务。

扩展

调查第二层关系网

做完这一步如果所获不多,下面你就该检查所有这79个号码的通话记录。这就是所谓的通往“大数据”之路。

我们将“大数据”与普通的旧数据区分开来的原因是:你无法只通过阅读和理解信息本身发现其背后的联系,在这个例子中,信息只是一个很长的通话清单,在你搜集了大量数据后,你将要做数学分析让重要的关系显现出来,这不是普通的人工检查可以完成的目标。

你又加入了新的号码,现在,以“基地”筹款人为中心,你就建立了两层关系网,第一层是30天内与“基地”筹款人有过通话记录的79个电话号码,第二层则是与这79个电话号码有过通话记录的电话号码,数据显示,第二层关系网一共有47923个电话号码,有超过67535种关系,有不计其数的通话、每个都有不同的时长、不同的时间段和不同的地点。

你已经受雇于美国国家安全局,所以你肯定很擅长数学,但是你所用的软件并不需要你理解其工作原理,你所需要做的只是输入数据,然后单击“确定”,软件会利用数学公式算出结果。


美国国家安全局长亚历山大

分析

对海量数据先筛选再分组

此时,它给你提供了一份每个号码重要性排名的数据。根据数据,你可以将这份名单从将近50000个电话号码削减至1200–22500个电话号码。

在数学意义来说,削减至22500个号码是最有利的,你可以削减掉超过一半的号码,但只有不超过7.5%的重要号码被漏掉,这是最具“性价比”的一个削减数额。

当然,你削减的号码越多,你漏掉的重要号码也就越多,如果你将号码削减至1200个,这表示你削减掉了97.5%的号码,那么,93%的重要号码会被遗漏,从数学意义上你依然有的赚,但从情报意义上,这已经没有什么价值。

不过,无论怎么削减数据,我们要调查的号码依然很多。这就需要你从菜单中选择一些额外的分析选项,然后单击“确定”,软件会迅速根据数据将这些号码分成许多不同的组。

如果你个别检查每个组的内容,就有很大几率获得有价值的情报。你会发现某些“专题组”,这个组里号码的拥有者可能居住在同一个地区,或者支持同一个组织。但这一切,只有在你知道了这些人的姓名、住址,以及他们通话的内容之后才能知道。

通过寻找这种关系,你可以从大量的电话号码中发现较小的“团伙”,一般情况下,一个“团伙”涉及的电话号码少于400个,他们之间有着最集中、最活跃的联系,这是一个非常小的群体,他们很可能有着某种直接的联系,你可以找出哪一个“团伙”与那个“基地”筹款人的关系最为密切。

你有权分析半打这样的“团伙”,每个“团伙”都有不同的潜在目标,有的很大,有的很小。

你也可以继续将网扩大,比如将那个“基地”筹款人的联系网络再扩大一个层次,调查那47923个电话号码的通信记录,但这样你就会要分析不计其数的电话号码,你将面临比以往更复杂、更具挑战性的任务。所以还是喊停吧。

判断

根据8大原则确定深入调查对象

是时候做出决定了。

我们应该弄清哪些电话号码拥有者的姓名、住址和其他数据?我们应该对哪些电话号码进行更深入的调查(这些调查包括将他们的信息通报给FBI或CIA,对他们启动窃听电话等监视手段)?

你可以按下列原则推荐号码,要求对其进行更深入的调查:

1.与“基地”筹款人电话号码有通话记录的79个号码;

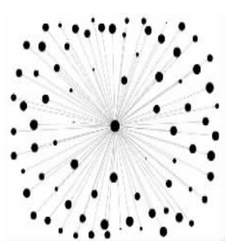
2. 那79个号码中最重要的24个;

3.“基地”筹款人电话号码的第二层外围号码,即那47923个电话号码;

4.从47923个号码中挑选出来的1200个“重要性得分”最高的号码,不过这里包含的有价值号码只占所有有价值号码的7%;

5.4500个“重要性得分”最高的号码,这里包含的有价值号码占所有有价值号码的21.5%;

6. 得分最高的22500个号码,这可以囊括所有有价值号码的92.5%;


美国国家安全局分析师的分析图

7. 你通过分析将这些号码分成一些“团伙”,选择其中一个“团伙”进行调查,其包含的电话号码可能从数百到数千不等;

8. 选择其中最活跃的几百个号码进行调查,但这有一定风险,因为真正的“基地”组织成员为了确保安全,可能刻意保持与那个筹款人通话的低密度。

你会怎么做?

风险

“最小化”程序尽量避免侵犯美国公民隐私

你已经“触摸”了成千上万的电话用户的记录,这些用户包括许多美国公民。绝大多数通话记录只是为了帮你完成甄别程序,把调查范围缩得更小一些。到目前为止,你尚未找到一个实际的目标号码。

但只要你仔细看一下就会发现,电话号码是非常结构化的数据。清单上的47923个电话号码起码有好几千个是美国的号码。你该感谢电话号码的区号制度和交换技术,你只需轻松地按一下按钮,就能得到所有号码所在地区的分布图。

美国国家安全局说,它有一个“最小化”程序,以防止对美国公民隐私的不必要的侵犯。据估计这个程序就是阻止分析师夸大或缩小美国电话号码的情报价值。但是如果列表中有十多个联系号码都在明尼阿波利斯,这是否意味着需要进行深入调查呢?另外在何时保证安全的重要性超过侵犯隐私的负面效应?是以那79个电话号码为界限,还是以那22500个电话号码为界限?

这就导致了另一个关键的问题:我们该在多大程度上相信数学?你可以做很多不同类型的分析,每个分析都有长处也有短处。那么哪些分析是适合这个案例的?还是任何一个都适合这个案例?

网络分析已证明了自己在发现重要节点方面的可靠性,但那只是网络结构方面的重要节点,和发现重要的、危险的恐怖分子并不完全是一回事。而如果你把网撒得越大,你就越可能发现自己在分析一张社会关系网,而不是一个恐怖网络。

你的分析之中的一个重要部分就是利用电话的持续通话时间和通话时间在一天中的时段来确定哪些电话更有可能和恐怖活动有关。但是,这些标准反映的是历史的趋势,有没有不断更新?更重要的是,这种标准有没有做过精确性测试?

有没有一种办法可以测试可信度而不用顶着“侵犯隐私”的恶名?有没有一种办法可以搜集到样本网络中所有的通话内容,然后让你比较一下你的预言和事实的出入?如果没有进行过精确性测试,你还能相信你的预测吗?

这些问题的大部分都没有明确的答案,但是一些因素影响了政府选择如何去回答它们。

一方面,美国的政策仍然是“所有的恐怖袭击都可以事先预防”。一个被经常表达的结论是:当你在面对一个定时炸弹时,去谈公民自由是否遭到了破坏是毫无意义的。

但公众并不知道,通过这些监控技术,究竟有多少“迫在眉睫”的恐怖袭击被阻止了。谁来判断你的分析是不是及时?

抛开隐私政策不谈,保持你的数据的集中性,避免使其膨胀也是非常重要的。当你从那个“基地”筹款人开始着手调查时,如果从和他有通话记录的79个号码中就能查出名堂,不需再扩大范围到那47923个电话号码,那么不管是在保证公民自由方面还是在反恐方面都会产生较好的效果。

但是如果你的起始调查点不止“基地”筹款人这一个“种子”,而是有多个“种子”,比如说包括“基地”筹款人、他的庄家以及为他们工作的4个快递员,那么你就可以进行更强大的数学分析,当然你需要分析的电话号码数量也会成倍增加。

分析师对于数据都是“贪婪”的,即使他们不一定用到它。而很多政府机构充满了相信“大数据”魔力的人。不可避免的结果就是:当总统、国会议员和法官被人用一些专业术语灌输了数据越多代表恐怖袭击越少的思想时,他们就倾向于开一张“空白支票”。

问题

十大问题亟待解决

问题很复杂,但并不令人费解。一旦明显越过了“国外情报监视法”所允许的界限,如果还想继续使用这些技术,就必须解决一些问题:

1.在调查成为棘手的“侵犯隐私”之前,一个情报分析师可以在多大程度上接触美国公民的个人数据?如果分析师并没有单独查看,而只是将电话记录输入分析程序中,违法吗?如果分析师只是看电话号码,但并没有查看号码的拥有者是谁,违法吗?如果查看了号码的拥有者,但并没对其进行额外的调查步骤,违法吗?

2.数据越完整,元数据分析就越精确。那么还应该用“最小化”程序将美国公民的数据过滤掉吗?过滤意味着最终的目标可能不准确,具有讽刺意味的是,难道调查无辜的人比侵犯个人隐私更经得起推敲?

3.属于美国国外运营商的“目标号码”在所有目标号码中占多大比例?这些数据的缺乏是否使得整个调查有点过于针对美国公民了?

4.在最基本的层面上,人们愿意信任数学公式和行为模式,并由其决定谁该接受监视调查吗?

5.“元数据”分析中很少涉及确定性,它几乎总是产生概率。那么,多大的概率就能促使美国政府打开权力滥用的潘多拉魔盒,对公众进行监视,或者对私人住所进行搜查,或者使用无人机进行攻击?90%?60%?当概率处在政府采取进一步行动的边缘时,我们应该允许侵犯更多的公民隐私、搜集更多的数据,以使这个概率更加明确吗?

6.我们有没有测试一下自己的分析数学,看看实际通话内容在多大程度上和自己的预测相关联?如果测试了,那么这个测试是如何做的呢?如果没测试,我们该相信在其他领域取得成功的模式吗?还是需要专门对它们进行一个反恐方面的测试?

7.如果我们认为这些模式需要进行精确性测试,我们愿意忍受在做这些测试时几乎无法避免的侵犯个人隐私的行为吗?更精确的模式会减少许多对无辜者的调查,从长远来看,那不是更好地保护了个人隐私吗?

8. 美国民众愿意信任政府,让政府持有这些数据吗?虽然政府说这些数据主要涉及国外的反恐行动,但美国民众相信总统不会在美国本土内生的恐怖袭击有可能发生时,命令国家安全局查看美国公民的信息吗?

9.如果像波士顿马拉松爆炸事件那样无法立刻确定恐怖袭击的起源,那么会发生些什么呢?国家安全局应该立刻展开更广泛的“元数据”分析直到找出该对恐怖袭击负责的人吗?

10. 如果允许在美国国内的反恐调查中使用这些技术手段,那么在政治危机发生时,怎样才能避免政客使用这些技术去调查自己的反对派?例如,这种分析很可能在无政府主义运动——“占领运动”中分析出一长串的领导人名单来。而针对民兵组织的分析则可能找出一大批枪支卖家。当你深入一个巨大的数据库时,你有时会得到自己并不想要的结果。

这些问题中没有一个是简单的。没有一个可以通过投票解决,也没有一个可以由专家提出见解解决。想通过合理和准确的方式对这些问题进行讨论都不是一件容易的事。

“大数据”是一个交叉点,反恐主义和包括“宪法保护个人不受无理搜查”在内的美国法律体系在这里汇集、交锋。美国的开国元勋无法想象,有一天一台机器采用先进的数学分析程序就能提供批准搜查令的依据。

现在的技术能力已经远远超过了美国宪法第四修正案提出者最疯狂的梦想,不管是法院还是法律本身都很难跟上技术的脚步。

如果美国不能激发全民的智慧解决这些问题,那么美国可能会失控,不是失去自由,就是失去安全,或者两者都失去了。

而且,没人能够解释这一切为什么会变成这样。