中国科协学会学术部 编

新观点新学说学术沙龙文集 92

大数据时代隐私保护的挑战与思考

中国科学技术出版社
CHINA SCIENCE AND TECHNOLOGY PRESS

新观点新学说学术沙龙文集 ⑫

大数据时代隐私保护的挑战与思考

中国科协学会学术部　编

中国科学技术出版社
·北　京·

图书在版编目（CIP）数据

大数据时代隐私保护的挑战与思考 / 中国科协学会学术部编 .—北京：中国科学技术出版社，2015.5

ISBN 978-7-5046-6838-7

Ⅰ . ①大… Ⅱ . ①中… Ⅲ . ①隐私权—法律保护—研究—中国 Ⅳ . ① D923.04

中国版本图书馆 CIP 数据核字（2015）第 139626 号

选题策划 赵 晖
责任编辑 赵 晖 夏凤金
责任校对 杨京华
责任印制 张建农

出 版 中国科学技术出版社
发 行 科学普及出版社发行部
地 址 北京市海淀区中关村南大街 16 号
邮 编 100081
发行电话 010-62173130
传 真 010-62179148
投稿电话 010-62103182
网 址 http：//www.cspbooks.com.cn

开 本 787mm × 1092mm 1/16
字 数 200 千字
印 张 10.25
印 数 1-2000 册
版 次 2015 年 12 月第 1 版
印 次 2015 年 12 月第 1 次印刷
印 刷 北京长宁印刷有限公司

书 号 ISBN 978-7-5046-6838-7 / D · 94
定 价 18.00 元

序 言

由中国科协主办，中国密码学会承办的“大数据时代的隐私保护的挑战与思考”学术沙龙经过两天的热烈讨论，取得了良好的效果，收获颇丰。目前即将出版的这个册子，就是根据大家的发言和讨论内容整理而成，供更多人参考。

随着移动互联网、物联网、云计算等技术的快速发展，全球数据量出现爆炸式增长，人类已经进入了“大数据时代”。大数据正以前所未有的速度，丰富人类探索世界的方法、驱动产业间的融合与分立，在政治、经济、军事、国家安全等方面的用途和潜在利益也引起了各界的广泛关注和重视。

大数据在提高经济效益和社会效益的同时，对信息安全和隐私保护提出了新的挑战。众多案例表明，数据大量收集后会暴露用户隐私，互联网便捷的传播方式更是加剧了隐私泄露的广度和深度，而且遭到泄露的隐私很难得到充分和即时的控制。2013 年发生的“棱镜门”事件加剧了人们对信息安全和隐私的担忧，为反思与应对大数据时代的隐私与安全提供了范本。

在 2014 年 10 月 19~20 日，中国密码学会在贵阳市组织召开了第 92 期新观点新学说学术沙龙，就大数据时代的隐私保护挑战与思考进行了深入而热烈的讨论。

此次学术沙龙得到了国内外专家学者的响应和支持。来自中国科学院软件研究所、北京信息科学技术研究院、杭州师范大学、清华大学、中国信息安全研究院、国家信息技术安全研究中心、西安电子科技大学、中国科学院信息工程研究所、复旦大学、解放军信息工程大学、北京电子科技学院、重庆大学、华南农业大学、福建师范大学、广州大学、贵州大学、上海大学、大连海洋大学、电子科技大学、华为技术有限公司、阿里巴巴集团、卫士通信息产业股份有限公司、沃通电子认证服务有限公司、郑州信大捷安信息技术股份有限公司、可信计算与信息保障实验室、信息物理社会可信服务计算教育部重点实验室等科研机构、高等院校、企事业单位的 34 位科技工作者作为讨论专家参加了学术沙龙，还有来自

国内有关科研院所、大专院校、企业应用单位的专家学者、科研技术人员及在校博士等对主题感兴趣的80余名代表旁听了沙龙讨论。

围绕“大数据时代隐私保护的挑战与思考”这一主题，此次沙龙活动安排了28个主题发言，内容涉及大数据对网络安全带来的挑战、适用于大数据安全的密码学技术、大数据的隐私保护关键技术、大数据应用中的安全和隐私问题等。与会代表按照沙龙管理办法，严格控制发言时间，言简意赅，切中主题，深入讨论。通过报告和讨论，系统总结和梳理了大数据时代的信息安全与隐私保护问题，深入分析和研讨了大数据时代隐私保护发展前景及存在的问题。与会专家一致认为，大数据时代的到来对隐私保护问题提出了前所未有的挑战，需要大力加强大数据安全技术的研究和应用，为敏感信息的传输、存储、发布和使用保驾护航。

科学的本质是批判，交流的本质是质疑。本次沙龙秉承了这一理念。因为是新观点新学说学术沙龙，而大数据时代的隐私保护又是一个新的信息安全问题，探索性很强，所以这次沙龙上出现了不同观点和意见的交锋，实现了大数据时代隐私保护的研究特色、思维方法以及交叉特色的深入探讨和交流，对具有代表性与前瞻性的研究领域进行深刻分析与辩论，针对如何平衡大数据时代信息共享与隐私保护这一矛盾问题，激发了新问题、新思想，推动了新的学术观点的诞生。

通过此次学术沙龙，我们认识到，大数据时代的隐私保护技术涉及匿名身份认证、全同态加密等新型密码、密文检索与安全数据查询关键技术、适用于大数据的属性密码和灵活授权技术、安全数据去重技术、PKI技术的全面应用等密码技术，也包括了增媒体的挖掘与反挖掘、隐私保护的数据挖掘、面向大数据的访问控制、大数据的完整性检测、基于安全事件的大数据实时处理等技术，涉及互联网应用、社交网络、远程医疗卫生、信息物理社会融合环境、政务移动办公、私有云安全存储等应用环境，值得深入研究和发展。大数据也对网络安全带来了新的冲击，互联网大数据在量子计算继续取得进展的情况下存在着潜在的威胁，如何寻求大数据技术发展与隐私保护的平衡，并通过网上隐私保护立法来保护用户的隐私，是需要进一步思考和解决的问题。

希望通过此次学术沙龙，适用于大数据的隐私保护技术能够得到有效发展和应用，在大力发展大数据的同时较好地保障安全和隐私。

张振峰

目　录

时间：

2014 年 10 月 19 日上午

地点：

贵州省贵阳市

主持人：

冯登国

贵州大数据产业发展及信息安全需求

◎ 陈　坚

我来自贵州省科技厅，下面介绍三个方面的情况。

第一个就是贵州选择大数据。大家都知道，今年贵州把大数据作为一个新兴产业来进行发展，这有几个方面的考虑。一个是贵州是一个欠发达地区，我们的人均 GDP 现在是全国倒数第一，工业不发达，错过了工业发展的机遇，另外城镇化水平低于全国平均水平。按照总书记的要求，要实现到 2020 年与全国同步全面建成小康社会的目标，这点对我们特别重要。另外，我们经济财税主要来源于煤电，煤和电还有其他化工对我们的经济影响比较大，相对来说工业这块重工业比较多。总书记对贵州提的要求就是要我们守住发展生态这条主线，牢牢守住生态和发展的衔接，既要发展得快，还要保持生态发展。要守住这个的话，再搞一些工业发展、城镇化发展的老路，对于我们来讲是没有前途的。所以说，这个是对我们发展要求的抉择，贵州这时候选择了大数据这个产业。另外，贵州也可以做好大数据。大数据是基于现代信息化发展的阶段，尤其是互联网发展到现在，数据越来越多。有专家说，今年是大数据的元年，这个数据来自于很多方面，既有政府所掌握的一些数据，商业创造的一些数据，另外还有一些社交场合的大数据，现在包括我们的各种传感器也可以带来很多数据。大数据元年，对谁来说都是在同一个起步线上。

第二个是贵州的条件比较好，我们贵州的电能比较充足。贵州这几年各方面条件都比较好，交通条件也比较好，明年我们可以达到 5000 千米的高速公路里程，可以实现县县通高速。另外，我们已经有 9 个支线机场，今后要发展 13 个支线机场，所以各方面条件越来越好，而且信息化这方面网络也是比较多的。另外像贵阳这地方叫“爽爽贵阳”，气候条件比较好，比较适合于在这里发展。我们夏天都不用吹空调的，一年四季都盖被子，空气质量也很好，这些方面应该讲

是一个有利条件。

第三个就是我们出台了一些科技方面的政策。在大数据方面应该说是有一些政策的。

第四个就是在党政的引导下，全社会全方位地对大数据予以推进。在招商引资上把它作为主要内容，另外在对外合作交流中，我们也把它作为主要内容。基础设施投资方面现在也在各方面给予保障，三大运营商都在这里建立了他们自己的数据中心。还有一个就是学校院所在加强这方面的研发和培训，这也为我们的大数据发展提供一些条件。政府在引导这方面的投资，企业也在选择这方面的发展方向，包括我们的科技部门现在也在这些方面做一些科技小学，还有我们领军型企业的科技人才计划。我觉得在这些政策下，包括各种实践，我们贵州做大数据这个产业是可以做得出来的。

第三点，我想举几个例子。第一个我想谈谈贵州省一个叫创客的公司。这个公司在社交的场合，利用机器人的方式可以为愿意搞设计的人提供一些指导条件，把他的软件在网上公布，别人用他的软件进行设计之后，就可以转化为有关协助人员公司给他加工零件，通过零件组装，可以使飞行器做各种各样的，满足一些创造发明开发者的创造条件。创客组织在中央电视台也播过，贵州有这么一个组织，而且这个团队的人员都来自世界各地，现在他们有十多个人了。整个创客公司是在全球范围内，通过众筹模式得到一些全国各地的资金，有点像原来腾讯搞 QQ 这样一个社交的模式，它的创客组织也可以通过这个来结交国内的一些爱好者。这些爱好者可以通过创造还有实践最后做出来样品。现在这个公司在准备筹办世界创造大赛，我们想这是一个比较成功的案例。第二个就是我介绍一下我们的朗玛公司，朗玛公司现在也是上市公司了，原来也是搞社交，语言的即时通信社交，另外就是游戏。贵州省把大数据作为新的领域以后，朗玛公司已经在搞智慧医疗方面，2014 今年收购了中国最大的健康网站——三九健康网。另外现在与成都电子科大合作，要成立一个研究院，专门围绕着医疗大数据方面做工作，它还要收购三甲医院，将来在全国推大数据医疗。第三个我想介绍的就是我们的食品安全云，特别是 2014 年我们贵州省 7+N 政府云启动的部分。现在我们的食品安全云除了把农村农产品的食品安全监测这项体系建立起来以外，最近我们还与京东商城合作。我们与京东商城签订协议，京东商城的第三方食品安全认

证就由我们贵州省食品安全云来为它承担。实际上我们是给京东商城的食品安全贴一个食品安全的标签，主要是把食品的各种监测数据搬到互联网上。我们除了自己搞食品安全以外，现在还与电商之间在整合。作为科技主管部门，今年我们也召开了农村信息化示范试点，也在大力推动智慧医疗、智慧物业等。我们在想，在省委省政府的正确领导下，我们大数据这个产业将会越来越红火。

信息空间、网络空间、“大数据”对网络安全的冲击

◎ 王育民

大家早上好，很高兴有机会和大家在一起交流，今天我想给大家介绍的有这么几个问题。第一个是信息化社会里面什么叫信息，信息指的什么东西，信息怎么来定量的研究，还有就是 Shannon 信息论，作为传输理论的信息理论；第二个是关于互联网的一些看法；第三个就是关于信息空间、cyber 空间和互联网，怎么来看这些东西；第四个就是开放网络中的信息安全；第五个是关于网络安全；最后一个就是大数据对网络信息的冲击。我想可能没有时间很细致的来论证这些问题，只是讲一些观点。

第一个是关于信息量和信息论，“信息”概念是非常泛泛的，到处都用“信息”。但是信息怎么来定义，是一个非常困难的问题。像控制论的创始人 Norbert Wiener（诺伯特·维纳），他就讲了信息的问题。其中讲了一句很重要的话，他说信息就是信息，它不是物质，也不是能量。好像没有给信息真正下出定义，但是这对于我们认识世界具有重大划时代意义，就是说，他把信息作为一个非常重要的问题，和物质、能量并列了，作为一切系统的三大要素之一。任何一个简单的系统，一个开关系统它得有物质，得有开关，得有电力控制能量，还得有开和关信息来维持整个电力照明系统。大到宇宙，到生物系统，它都是由这三个要素组成的，所以这句话是非常重要的一句话。物质和能量是客观存在，是有形的，信息是抽象的，无形的，物质和能量是系统的具体，信息是系统的灵魂。所以信息要借助物质能量才能产生传输、存储、处理和感知，物质和能量借助于信息来表述和控制。我们在 20 世纪 80 年代的时候开过信息论年会，张兆哲老师曾经列出 100 多种信息量的定义。但是实际上现在我们能够对付得比较好的信息大概只

有一种，就是不确定性信息。Wiener 也讲了一句非常重要的话，什么叫信息，信息不是熵，而是熵差。熵差是什么呢，熵是描述物质运动的空间里面的不确定性，熵差是解除的不确定性，这是你得到的信息，这句话也是非常重要的。

Shannon 就是按照不确定信息给它研究透了，创造了一系列的理论问题，然后创立了信息论。这个信息论就带来了从 1948 年一直到现在整个信息传输处理的数字化革命，它是数字化革命的一个指导思想。数字化革命的物质基础就是 VLSI 技术（Very Large Scale Integration）和计算机。

物理空间与网络空间、信息空间与 cyber 空间怎么来看。数字化革命中人类创造了一个网络空间，这个网络空间是人造的，这是控制论里面的一个观点。网络空间里面要传输信息，要存储、处理信息，这就构成了人类生存的第二空间，叫 cyber 空间，它是整个宇宙信息空间的一个子空间。cyber 空间属于宇宙中早已存在的信息空间的一个子空间。自从有宇宙之后，宇宙中就充满着信息，动物与动物之间有信息，植物之间也有信息传递，人和动物和植物之间也有信息交流，宇宙又从很远的地方在给我们地球传来很多与宇宙有关的信息。但是我们想，这里面要特别强调的是，不要把网络空间和 cyber 空间等同起来。网络空间是物质的，属于物理空间的范畴。cyber 空间是宇宙信息空间的一个子集，它是抽象的，无形的。将 cyber 空间译为网络空间，我觉得是值得大家商榷的一个问题。所以网络空间的安全与 cyber 空间，也就是与信息空间的安全、信息的安全也不能够等同起来，这里面还是有区别的，我们后面还要讲。

第二个就是关于物联网，怎么来看物联网。Shannon 在 1948 年给出的信息传输的传输系统模型，里面有信源和信素。信源是产生信息的及它要送给大家的，它要送给信宿的，信宿是接受信息的，这么一个简单的模型。在这个模型里面，他讲了一句很重要的话，就是“信源和信宿，既可以是人，也可以是物”。这里面就预示着人与人之间的通信交流，这是我们大家非常熟悉的，我们从有人类以来就在交流。但是人与物能不能交流信息、物与物能不能交流信息？长期以来我们主要研究和发展人与人之间的通信，在控制中会涉及人与物间的通信，但很少提到物与物之间的通信。这主要是因为物不具有对周围环境的感知能力，也不具有与外界通信交流的能力。什么时候有了呢？人类创造了传感器以后。我们可以把传感器绑定给每一个物，这个物就具有了对周围环境的感知能力，它可以

采集信息，然后它也可以跟其他的物、也可以跟人进行通信，这时候就有了物与物之间的通信和物与人之间交流的问题，这才出现了物联网的问题。所以物联网的连接我认为应该有两种不同的：一个是通过传感器所构建的传感器网络，这些网络大多是为完成特殊任务的专用网络，可将这类网称为物联网（Networks of Things）；再有一个是物与物之间通过互联网的 TCP（Transmission Control Protocol，传输控制协议）/IP（Internet Protoco，互联网之间互连的协议）协议所构成的物连接的网，这叫物联网（Internet of Things）。所以物联网本身应该是整个我们互联网的一个子集，它必须得用 TCP/IP 协议，然后才加入到互联里面去。所以不要把物联网吹得太过，好像是一个新的发展，是高于互联网的发展，实际上它是互联网的发展和延续，是扩充。

物联网的关键技术，一个是传感器技术，另一个是能源的技术。有些物的连接不一定就是上互联网，比如说一个桥的监测系统。你把这个传感器固定在桥墩里面，固定在钢筋水泥里面，它要对整个桥的应力情况进行感知，然后告诉大家这里面有没有问题。把传感器埋在里面就再也取不出来了，所以它的电源供应很重要，也要埋在桥墩里面来支持它继续工作。其实人和人之间经过网络通信，也需要借助传感器。我们人可以进行直接的交流，但是要通过电的手段来传送信息，你没有送话器行吗？你没有受话器行吗？送话器、受话器本身是人发明的传感器，所以从这个意义上来讲，人与人之间的交流和物与物之间、人与物之间的交流没有本质上的区别。

第三个问题，就是信息空间、cyber 空间和互联网。信息空间我们已经讲了，它自有宇宙以来就一直都存在着，包含宇宙中的所有信息构成的空间，其中有宇宙生成和发展的历史、地球出现生命、人类社会发展、现有动植物等的信息。cyber 空间我们也已经讲了，就是由人所创造的，利用计算机和通信设备构成的一切通信网络中的信息所形成的空间，所以 cyber 空间应该是属于信息空间的子集（图 1）。互联网大家都非常清楚，就是用 TCP/IP 协议所连接出来的虚拟网络。互联网本身是虚拟的东西，它有承载网络，各种各样通信网络就是它的承载网络，承载网络是物理上的东西。互联网里面还分出内域网、外域网。这些名字在 20 世纪 90 年代出现，我们讨论过，其实是很好的名字。比如内域网，就是各个公司、各个单位的一个内部网络，例如一个国家的外交系统，都可以构造它自

己的内域网，都可以用这个 TCP/IP 协议。互联网是没有地域限制的，例如我们外交部的网络系统，整个地球凡是有外交关系的国家和地区都要建立它的终端。

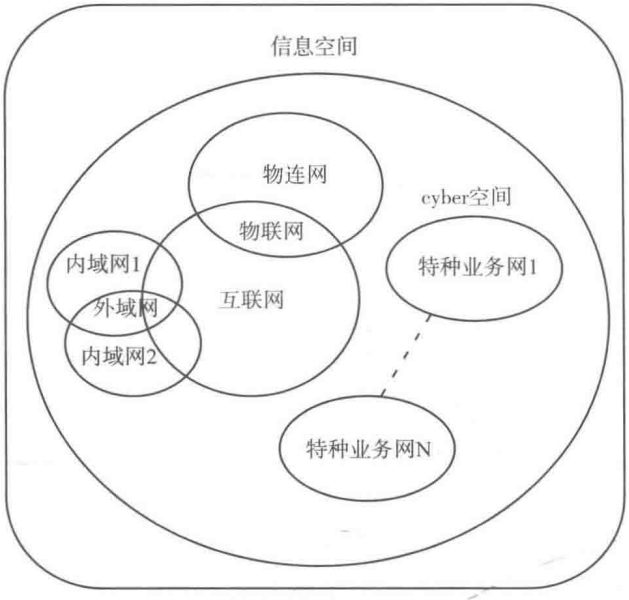

图 1　信息空间与 cyber 空间

第四个就是开放网络中的信息安全问题。我们知道，Shannon 指出，通信的基本目的是“在此时彼地或彼时彼地精确或近似地重现信源的输出”。这里面的彼时是我们考虑到有存储器以后可以延迟通信的情况，所以做了一点点修正。在开放网络传输、存储、处理信息的环境中，影响将信息传递给意定接收者，（我们这个通信一定不是说泛泛的，是要想把这消息传递给他想要给人家知道的那个接收者，叫意定接收者）影响这个任务的因素有两个：一个是物理噪声，这个在信息论里面处理得非常好，把物理噪声用随机模型来描述怎么样对传输产生干扰，造成了含糊度，造成了你不能够接收恢复的原始信息。第二个影响这个任务完成的因素就是有敌对的敌手存在，敌手是我们通信信息安全的大敌，也是我们通信能够进行正常可靠的大敌。敌手可以有两种攻击通信系统的办法，一个是被动攻击，它不影响你正常传输，但是它窃听你，而且还可以分析利用系统的漏洞。被动攻击只是接受信号、恢复数据，对它进行分析。比如说你是加密的，我

要进行密码攻击，能够破译你的密码来提取信息，这就是被动攻击。一种是主动攻击，就是侦测、分析和利用系统的漏洞，向系统注入有害的数据，实现假冒（身份）、窃取（信息、权限）、嵌入（陷门、木马、病毒、蠕虫等）、破坏数据的完整性，直至部分地控制系统，达到破坏系统的正常工作的目的，这就是主动攻击。过去我们在研究保密通信，研究密码学的时候很少考虑主动攻击的问题。到了信息时代，这种破坏你通信的手段有了很多新的发展，这时候主动攻击成为更主要的对象。被动攻击和主动攻击两者是相辅相成的，常常是配合使用的。对关注的信息进行侦测、定位、查找、提取、认知，实现对信息的全面觉察，这是美国在“9·11”前后提来的问题。信息全面觉察，比如人肉搜索，对特定的人，将其隐藏于网络中的信息进行搜索、定位、认证、解读，并最终把他暴露于光天化日之下。比如对拉登的追踪、定位和处置就是信息觉察的一个案例。上述论述表明，噪声和敌手的存在威胁着系统要将信息传送给意定接收者的任务。更具体地说，噪声和敌手的存在会威胁或影响信息的下述几个方面，也就是我们在研究信息安全的时候常常提到的，比如信息的可靠传输的问题，信息的保密性、隐私性的问题。

最后我想就大数据的问题再提几句，我觉得大数据问题是在于人类发明了一种新的无形的显微镜和望远镜来在信息空间中进行观察，进行搜索，进行解读，就像我们18、19世纪发明物理上的显微镜和望远镜一样，这是一个观点。再有一个观点就是“大数据”带来科学研究方法的更新，这是一个新的阶段了。第三点就是大数据使机器进入了智能化阶段，我们过去发明的机器人都是代替人类体力的比较多，而现在是代替人类脑力的比较多。未来的诺贝尔奖金要授予谁？曾经一个法国杂志说，未来很可能授予像Google这样的搜索工具。它要得诺贝尔奖，这个也是有道理的，就是说它现在利用大数据做了很多新的科学发现，很多新的东西是用人力、用科学家的脑袋，个人处理能力都已经不能解决的问题，是要用人发明的智慧工具来解决的问题。

向　宏：

我问王教授一个问题，我们现在做网络空间研究的时候，您刚才说了，这个概念确实有很多，有叫信息空间的，有叫网络空间的，还有叫cyber空间的，而

且我们所给的这些定义还跟国外不太一样。比如说 cyberspace，美国国防部的定义，就有点类似于您刚才说的整个网络空间，包括物理的设备、人等东西，而您刚才说的 cyber 空间，就是 cyberspace，按照我的理解，应该是一个纯信息的子空间。所以我在想，咱们是不是有必要在研究的时候，国内相对来讲统一一下，我们说的信息空间也好或者网络空间也好，或者 cyber 空间也好到底具体是什么，定义一下。这个争论很长时间了，现在也应该差不多给出一个相关研究的工作者统一的一个定义，不然的话跟国外的研究老是错位。

王育民：

对，我也同意这个意见，但是这个可能是要靠很多学会共同运作的，比如说计算机学会，密码学会，还有通信技术统一运作，大家来碰碰头，才能够统一这个思想。因为现在像河南郑州信息工程大学，他们已经成立了网络学院，但是网络的安全要比信息安全的面广得多，大得多，怎么样防止网络被物理破坏都是网络安全的问题。但是信息安全我认为从理论来说，用 Shannon 在 1948、1949 年和后来搞信息论人发展了的一些东西，像认证码的问题，这个问题我认为还是比较系统地来解决的。我刚才没时间讲，来学嘉到我们那儿做报告就讲了这个问题，他说主动攻击里面最主要的是单向函数的问题，就是研究密码的单向性问题这事很重要，这个是对的，现在公钥体制里面都是有一个单向函数。但是这个问题，Massey 在他们学校有一个内部的讲义，我看过，Shannon 很明智地避免了对单向函数给出定义，因为这个你没法用数学严格来描述，你就是可证明安全也只能说这个等价于那个，像这个。就跟信息一样，信息很多是你不能定量研究的，爱的问题你不能定量地来研究，“你说你对这个爱有多深，有几分，只能说你去想一想、你去看一看，月亮代表我的心”，就类似这样的东西，他说不出来这一份爱跟那一份爱哪份爱更多一点，没法比较的。美学也是这样的，好多信息的属性是不能用定量来研究的，包括我们隐私性也是不能用定量来研究的。所以安全性，安全的定量没法弄，这些都是没有解决的问题。

黄继武：

王老师，我请教个问题，大数据基本上是一个在不同程度下的统计数据，比