绪论

大数据正深刻影响着人们的生产方式、生活习惯、思维模式和研究方法。大数据不仅是学界和业界的前沿课题，而且已上升为国家基础性战略资源。大数据的开放、开发与利用已成为国家重大战略需求与重大工程需求。同时，大数据应用中面临的数据安全威胁与隐私泄露也极大地破坏了正常的社会经济秩序，甚至危及国家网络空间安全。

大数据是学界和业界的前沿课题，正深刻影响着人们的生产方式、生活习惯、思维模式及研究方法。随着互联网、物联网、云计算等技术的迅猛发展，包括自媒体数据、日志数据和富媒体数据在内的网络大数据飞速发展。然而，网络大数据在收集、存储和使用等全生命周期中面临着诸多安全挑战，大数据所带来的隐私泄露给用户带来了严重损失和潜在风险，极大破坏了社会经济秩序，影响了政务大数据、商务大数据、健康大数据等更多大数据的产业化应用。目前，大数据已上升为国家战略性资源，大数据安全与隐私保护不仅是国际学术前沿，也是国家重大战略需求。本项目致力于网络大数据环境中数据保护、隐私保护以及数字水印隐藏等方面的基础研究，具有重要理论意义和应用价值。

现有的数据隐私保护对象主要是用户的身份、社交关系以及用户属性等关键隐私信息。在大数据环境中，由于数据存在更广泛的关联性，使得原本孤立的信息成为隐私；同时，对经过传统匿名等技术处理后的数据，通过大数据关联挖掘和深度分析后，依然可能分析出用户的隐私，这些给大数据环境中的隐私甄别和隐私保护技术提出了新的挑战。

基于大数据环境下的数据安全和隐私保护理论，面向时空大数据进行数据与隐私保护示范应用，并对大数据环境下的数据与隐私保护性能与效率进行评价。针对位置大数据，突破北斗高精度定位时间同步技术，构建具有高精度时空信息的网络大数据环境，对大数据环境下融合时空信息的数据与隐私保护理论进行实验分析。课题的研究内容将本项目提出的数据与隐私保护理论机制落实到具体应用中，根据应用反馈修正数据与隐私保护理论，响应了指南的要求。

大数据的独特之处，除了规模巨大、类型多样、增长迅速等特性，最重要的是这些特性所导致的“全息”意义上的数据关联性，这种关联性将是实现未来商业模式、生产生活方式、管理流程等颠覆性变化的驱动力。譬如国家电网智能电表的数据可用于估计房屋空置率，淘宝销售数据可用来预测经济走势，移动通讯基站定位数据可用于优化城市交通设计，微博上的关注关系和内容信息可用于购物推荐和广告推送等。同时，数据关联性也是导致常规的数据保护与隐私保护方式失效的根本原因之一。例如，关联性挖掘分析使得仅通过匿名技术不能很好地保护用户隐私。但是，如果施加过强的数据保护策略，必将割裂这些数据的关联性，从而形成一个个数据孤岛并导致大数据服务的不可用。

高精度时空信息的网络大数据环境构建：以本项目提出的数据与隐私保护理论和方法为基础，针对位置大数据，通过北斗高精度定位时间同步技术，构建大数据环境中融合时空信息的数据与隐私保护理论验证环境。

1.2文章结构

本论文以时空大数据为背景，结合北斗导航系统和网络协议构建时空坐标系，为提供基于高精准时分信息的时空大数据研究提供研究基础，改进了Vivaldi算法并提出了稳定抑制Vivaldi算法，其目的是为了增加网络坐标系统的准确性。

本文分为6章，主要内容如下：

第1章引言阐述了时空大数据和网络坐标系统的重要性，以及网络坐标系统的课题背景和意义，介绍了国内外时空大数据研究的进展和对网络坐标系统的部分研究成果和几种构建算法。

第2章介绍了什么是时空大数据，时空大数据的相关特征，四种现有的典型网络坐标系统构建算法GNP[2]、PIC[10]、NPS[11]和Vivaldi[3]，北斗导航系统（GNSS），网络实验环境和协议OpenVSwitch和实验环境MiniNet的介绍。

第3章介绍了时空大数据平台构建的方案，通过构建开放的体系架构为实现范围更广泛的信息资源共享与多层次多节点的协同工作提供崭新的运行环境。

第4章利用Mininet和OpenVSwitch协议，构建系统原型，实现了对时空数据的标签化处理和打标签操作，通过数据的时空戳属性，融合了空间坐标系和网络坐标系，构建了时空坐标系。

第5章通过数据分析说明了随机延迟污染现象的普遍存在，介绍了一种现有的随机延迟污染抑制方法MP-Filter[6]，并提出TO—Filter的随机延迟污染抑制方法。并分析了恶化网络性能的其中两个因素：随机延迟污染现象以及TIV现象，同时介绍了对非中心式网络坐标系统安全性产生影响的几种攻击。三种现有的用于抑制TIV现象的网络坐标距离预测算法，并对其优劣点进行了分析。同时简单介绍了稳定抑制Vivaldi算法。提出了用于抑制随机延迟污染现象以及TIV现象的稳定抑制Vivaldi算法，并通过仿真实验分析其准确性和抑制抖动能力。

第6章是全文的总结。

1. 项目研究背景

大数据是学界和业界的前沿课题，正深刻影响着人们的生产方式、生活习惯、思维模式及研究方法。随着互联网、物联网、云计算等技术的迅猛发展，包括自媒体数据、日志数据和富媒体数据在内的网络大数据飞速发展。然而，网络大数据在收集、存储和使用等全生命周期中面临着诸多安全挑战，大数据所带来的隐私泄露给用户带来了严重损失和潜在风险，极大破坏了社会经济秩序，影响了政务大数据、商务大数据、健康大数据等更多大数据的产业化应用。目前，大数据已上升为国家战略性资源，大数据安全与隐私保护不仅是国际学术前沿，也是国家重大战略需求。本项目致力于网络大数据环境中数据保护、隐私保护以及数字水印隐藏等方面的基础研究，具有重要理论意义和应用价值。

“互联网+”、大数据、时空数据、时空大数据，是学界和业界讨论最多的问题。时空大数据的提出有其重要的意义，从哲学层面看，空间与时间一起构成运动着的物质存在的两周基本形式。空间指物质存在的广延性；时间指物质运动过程的持续性和顺序性。空间和实践具有客观性，同运动着的物质不可分割。没有脱离物质运动的空间和时间，也没有不在空间和时间中运动的物质。空间和时间也是相互联系的。现代物理学的发展，特别是相对论的提出，证明空间和时间同运动着的图纸的不可分割的联系。

从运动着的万事万物表达的层面看，事务都可以分为空间维度（S-XYZ）、属性维度（D）和时间维度（T）。时间维度（T）是指信息随时间的变化，具有时态性，需要有一个精确的时间基准；空间维读（S-XYZ）是指信息具有精确的空间位置或者空间分布特征，具有可量测性，需要一个精准的空间基准；属性维度（D）是指空间维度上可加载随时间变化的要素（现象）的各种相关信息（属性信息），具有多为特征，需要有一个科学的分类体系和标准编码体系。

从提升社会治理体系和治理能力现代化水平的层面看，随着全球化进程的加快，当今社会的一个重要特点是，世界（区域、国家、城市）管理和治理对事件和空间的依赖程度越来越高，时空大数据正日益成为全球（区域、国家、城市）治理体系和治理能力现代化的核心驱动力。进入21世纪，太空成为继核武器之后的新型战略威慑力量，与核威慑、网络威慑交织融合，共同构成新的战略稳定架构；太空实现全球作战力量一体化，不受传统的陆地、海上、空中飞越限制，具有天然的全球性和跨域性，全球任何地点的作战力量和手段都能通过“天地一体网络”连接起来，形成一体化作战力量体系；太空力量是战斗力的倍增器，可以在任何时间、任何地点、任何气象条件下打击地球上任何一个目标。而这必须有全球一体化的时空大数据平台保障。

本章小结

1. 项目关键技术分析

2.1 时空大数据

大数据（Big Data），指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

时空大数据，指基于统一的时空基准（空间参照系统、时间参照系统），活动（运动变化）于时间和空间中与位置直接（定位）或间接（空间分布）相关联的大数据。之所以这样界定时空大数据是基于以下三个事实：一是，世界是物质的，物质是运动的，包括人类活动在内的万事万物的运动变化都是在一定的时间和空间中进行的，而所有的大数据都是世界万事万物运动变化的产物；二是，随着智能感知技术、物联网、云计算技术的发展，各个领域开始了“量化”的进程，这种一切皆可“量化”（数字化）的趋势导致了大规模海量数据的产生，而空间参照与时间参照是大数据的两个基本特征；三是，从可视化角度讲，正是因为一切大数据都具有空间参照与时间参照特征，所以才能直观地为人们提供大数据的空间位置，空间分布和时间标识。

大数据具有以下特征：容量（Volume）：数据的大小决定所考虑的数据的价值和潜在的信息；种类（Variety）：数据类型的多样性；速度（Velocity）：指获得数据的速度； 可变性（Variability）：妨碍了处理和有效地管理数据的过程；真实性（Veracity）：数据的质量；复杂性（Complexity）：数据量巨大，来源多渠道；价值（value）：合理运用大数据，以低成本创造高价值。

时空大数据除了具备一般大数据特征外，还具备其他特征：位置（Location）：点、线、面、体的三维位置（X、Y、Z）拓扑、方向、度量很复杂；时间（Time）：位置、属性等随时间的变化而变化；属性（Attribute）：每个位置点、线、面、体上都以自己的数量特征和说明信息；尺度（Measure）：空间尺度变化；多维（Dimension）：所有大数据与空间数据集成构成空间数据立方体，即“多维”可视化。

时空大数据的产生、存储和管理需要一个时空大数据平台来完成。当前，全球正在经历一场持久而深远的数据化（一切皆可“量化”）革命，跨界、融合、开放、共享是大数据时代的核心特征。时空大数据平台，是把各种分散的（点数据）和分割的（条数据）大数据汇聚到一个特定的平台上—时空大数据平台。这种聚合效应就是通过数据多维融合和关联分析与数据挖掘，揭示事物的本质规律，对事物做出更加快捷、更加全面、更加精确和更加有效地研判和预测。从这个意义上讲，时空大数据平台是大数据的核心价值，十大数据发展的高级形态，是大数据时代的解决方案。

2.2 网络坐标

2.2.1 概述

网络在服务人们、提供共享信息的同时，网络性能已成为了人们关注的重点，这是因为网络服务质量的提示有赖于网络性能的提高，因而如何更快的获取网络上的信息已经成为了一个研究热点。网络坐标系统就是为了提高互联网距离测量效率而提出的。自2002年提出GNP[2]算法以来，现在已有基于中心式的如GNP[2]、PIC[10]和基于非中心式的如NPS[11]、Vivaldi[3]等时延预测机制，它们都以如何有效快速的获取网络节点间时延作为研究重点，同时都将网络节点放入N维的坐标系统，通过计算节点坐标距离来作为网络时延的预测值。

2.2.2 GNP简介

GNP[2]是最早提出的网络坐标系统，将网络中节点实测时延映射到数学的几何空间上，将获取网络节点间时延转化为获取节点坐标间距离。其思想是将网络构建在一个几何空间上，如三维欧式空间，并用该几何空间中的一个点作为网络中任何一台主机的位置。

在众多网络坐标系统中，GNP是具有代表性的基于锚节点的时延预测机制，其将网络中的主机分为两个部分。一个部分是事先选定好的在网络中均匀分布的节点，这些节点被称为锚节点(landmarks)，它们会首选在几何坐标中计算并确定自己的坐标，用于作为其他普通节点坐标定位的参考。锚节点保持自己的坐标，同时散布给所有任何想要参与进来的主机节点。另一部分是网络中的普通主机，它们通过获得锚节点的坐标，来计算确定自己的坐标，因此该方法具有很强的可扩展性。

在GNP网络坐标系统时，首先要确定网络中一小部分主机作为特殊节点，即为锚节点，其作用是在坐标系统中为其它普通主机提供一组必要的参考坐标。

假设有*N*个锚节点*P1,P2……PN*(*N* > *d*，*d*为几何空间维度)，使用Ping方式可以很容易的获得锚节点间的实测时延*Li,j*，从而构造一个*N*×*N*的时延矩阵，该矩阵沿对角线对称，则GNP需要找到一组坐标*X1*,*X2……XN*来表示N个锚节点坐标，从而使锚节点间坐标距离||*Xi*-*Xj*||与实测延迟*Li,j*的误差平方总值最小，即下面公式的值最小。

(2.1)

在建好了以锚节点组成的网络坐标系统后，在加入普通节点，同样使用Ping方式可以很容易的获得与锚节点间的实测时延*Lk*，并通过该普通节点与各个锚节点实测时延来确定普通节点的坐标*X*，并使得其与各个锚节点的实测时延与预测时延误差平方总值最小，即下面公式的值最小。

(2.2)

GNP系统是集中式的坐标定位系统，具有很强的可扩展性和较高的网络距离预测准确性。对于该系统，每有一个主机加入其中，这个主机就需要测量到所有锚节点的距离，因而可能导致节点负担过于沉重，开销太大，同时系统的距离预测准确度还受到锚节点的分布情况、数量、位置等影响。

2.2.3 PIC简介

由于GNP[2]过于依赖锚节点，普通节点需要测量到锚节点的距离，使得锚节点承担了相当的通讯开销，锚节点所能承受的通讯量则成为了网络系统性能的瓶颈，因而有研究者提出了PIC[10]。PIC是在GNP的基础之上，改进了锚节点的选择方案，由于PIC不需要固定的锚节点，其系统中任何已经有坐标的节点都可以被其他节点选作锚节点，从而使计算开销和通信开销能够均匀的分摊到各个节点上，同时也有效的避免了因为存在个别锚节点的坐标失效造成的误差。同时，PIC提出了一种锚节点选择机制，其有效性以及系统坐标准确性与GNP几乎一致。

与GNP系统一样，PIC也是采取了将网络节点映射到几何空间的策略，在节点P加入PIC系统中时，它从已存在坐标的*N*个节点中选择*M*(*M* > *d*，*d*为几何空间维度)个节点作为其锚节点，与上述GNP类似的，同样使用Ping方式可以很容易的获得与锚节点间的实测时延*Lk*，并通过该普通节点与各个锚节点实测时延来确定普通节点的坐标*X*，并使得其与各个锚节点的实测时延与预测时延相对误差平方总值最小，即下面公式的值最小。

(2.3)

而当系统中已存在的节点的数量*N* < *M*时，其节点的坐标计算方法不同：此时会将该*N*个节点都作为锚节点。与上述GNP类似的，对*N*个锚节点*P1*,*P2……PN*，使用Ping方式获得锚节点间的实测时延*Li,j*，从而构造一个*N*×*N*的时延矩阵，并找到一组坐标*X1*,*X2……XN*来表示N个锚节点坐标，从而使锚节点间坐标距离|| *Xi*-*Xj*||与实测延迟*Li,j*的相对误差平方总值最小，即下面公式的值最小。

(2.4)

关于节点*P*的锚节点选择有着以下三种不同的策略：

1)随机策略：*M*个锚节点是从系统中已有坐标的*N*个节点中随机选择出来的。

2)最近策略：选择网络系统中离节点*P*最近的*M*个节点来作为锚节点。

3)混合策略：*M*个锚节点中部分节点是通过随机策略选择出来的，另一部分则是通过最近策略选择出来的。

在最近策略中，PIC提出了两种方法来寻找离节点*P*最近的*M*个节点：

方法一：首先在已知坐标的节点中随机的选择一个节点暂作为目标节点，接着对该节点与其邻居节点的距离进行测量，如果存在一个邻居节点离自己的距离更近，则这个邻居节点作为目标节点，然后反复执行这个过程，直到无法找到离自己更近的节点。

方法一对节点间的距离的反复测量会对网络带来更多的开销，因此PIC又提出了方法二：使用坐标计算节点间距离来取代直接测量。该方法可以有效的减少网络开销，但该方法只有在新加入的节点已有坐标的前提下才能实现，因此PIC提出：首先使用随机策略来求出该节点的坐标，接着再使用该坐标来寻找离该节点最近的锚节点。

PIC通过数据实验表明，不同的锚节点选择策略对于节点坐标的准确度也会有着不同的影响：在使用随机策略选择锚节点时，对预测长距离的准确度更高；在使用最近策略选择锚节点时，对预测短距离的准确度更高；而使用混合策略时，对长短距离预测的准确度整体最高，因而在PIC中，都是选择使用混合策略。

2.2.4 NPS简介

NPS[11]是一种分层的非中点式的网络坐标系统。它主要解决的是：在网络坐标分布式计算时遇到的适应性、一致性以及稳定性问题。NPS能够让节点在同一的坐标系上，适应网络环境变化，同时准确的反映网络拓扑结构的变化，并能减少不必要的节点坐标更新。

NPS将网络节点分为三种：服务器节点，锚节点，和普通节点。其中，服务器节点储存着系统参数以及其他节点的信息；锚节点与GNP中的锚节点类似，参与几何空间坐标系的构建，供普通节点作为参考；其他的节点则为普通节点，这类节点也可以作为其他节点的参考节点。

NPS的网络坐标系统构建采用了分层层次结构，将网络节点分布到0到*L*层。锚节点作为NPS坐标系统的基础，被放在第0层；而普通节点的放置时，记所在层数为*Li*，则其选取的参考节点*j*所在层数*Lj*需要满足*Li*< *Lj*，该约束条件可以有效的避免节点间互相作为参考节点，使得坐标的计算保持了一致性。

NPS中，普通节点加入系统并完成计算坐标需有以下步骤：

1)在普通节点加入系统时，从服务器节点获得系统参数和参考节点的信息。

2)测量获得普通节点与参考节点间的实测距离。

3)计算普通节点坐标，使其满足普通节点与参考节点间预测距离与实测距离误差平分总值最小。

同时，每隔一段固定时间，节点将会重新计算自身坐标，使其适应网络拓扑环境的变化。

2.2.5 Vivaldi简介

在众多网络坐标系统中，Vivaldi[3]算法是具有代表性的基于模拟的时延预测机制，其使用物理模拟来预测时延。Vivaldi算法将网络节点通过将物理弹簧弹性定律，即“弹簧在拉伸和压缩时，其动能和势能会相互转化”，应用到网络节点坐标系统，将网络节点视为由弹簧相连的点，从而将节点间预测距离误差之和最小化问题模拟转化为求弹簧间系统势能最小值问题。

Vivaldi[3]算法是完全分布式的，节点仅需要获得任何具有坐标的邻居节点的实测时延并可完成自身坐标的更新，当实测时延与预测时延不一致时，则转化视为节点间弹簧发生了拉伸或者压缩，通过弹簧形变使得节点到达一个合适的位置，从而使整个弹簧系统势能最小，即网络坐标系统误差总值最小。Vivaldi算法的具体实现为：对于节点*i*，获得来着邻居节点*j*的测试时延*RTTi,j*:

1)计算误差权值*ω*：

(2.5)

2)计算两点距离相对误差*es*：

(2.6)

3)更新自身误差估计*ei*：

(2.7)

4)更新本地坐标*xi*：

(2.8)

(2.9)

其中*xi*为节点*i*的坐标，*u*(*xi*-*xj*)为从节点j坐标到节点*i*坐标的单位向量，*ei*为节点*i*的误差因子，*ce*和*cc*为调节因子，其取值越大，则节点每次更新的程度越大。

Vivaldi[3]算法是公认的具有优秀性能的网络坐标系统构建算法，其优点在于其实现是全分布式的，无需额外设立基站设施，对Internet拓扑变化有较好的适应性，并能及时更新网络坐标，将网络坐标系构建所需的时延探测任务交给应用级业务上，减少了网络开支。

2.3 北斗导航系统

2.4 软件定义网络（SDN）

2.5 OpenVSwitch

2.6 mininet

2.7 RYU控制器

2.8 本章小结

1. 时空大数据平台建设方案

3.1 基于Web Service的时空大数据平台体系

3.2 多层协同的架构体系

3.3 功能平台的服务体系

3.4 开放方向的数据体系

3.5 高效运行的管理体系

3.6 时空大数据标准体系

3.7 本章小结

1. 系统原型设计与实现

4.1 虚拟局域网

4.2 Mininet和RYU安装

4.3 OpenVSwitch 协议解读

4.4 网络风暴解除，ARP回复

4.5 胖树结构

4.6 本章小结

1. 时空网络坐标系性能分析与优化

5.1 网络坐标系统

5.2 随机延迟污染现象及抑制方法

5.3 网络坐标中三角不等式现象级抑制方法

5.4 基于Vivaldi算法的抑制方法

5.5 本章小结

1. 机遇和挑战

6.1 机遇

6.2 挑战

1. 总结

致谢

引文