清 华 大 学

综 合 论 文 训 练

题目：基于 k 匿名算法的物联网元信息交互的隐私保护机制研究

系　　别：计算机科学与技术系

专　　业：计算机科学与技术

姓　　名：阮尚祥明

指导教师：蒋屹新　副教授

2011 年 6 月 4 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名：_____导师签名：_____日 期：_____

中文摘要

随着互联网迅速发展的时代，物联网的提出对人们的日常生活具有重大意义与广阔的应用场景。基于集成小型化智能传感设备的物联网将来也会引起很多安全和用户隐私的新挑战。安全与隐私保护问题将降低人们对物联网的信任和应用，是影响物联网成功发展的重要因素之一。

论文对物联网感知层中各传感设备的信息交互提供一种新的安全隐私策略：匿名空间算法。这算法是借助网络各行业近十年广泛采用的 k 匿名模型的思想来实现。文章中将建立算法的系统模型、详解内容以及对它进行仿真测试、分析比较与总体评估。最后对算法在物联网背景的将来应用提出扩展研究与建议，加强物联网所在需求提高的安全隐私保护机制。

关键词：物联网；感知层；k 匿名算法；安全；隐私保护

ABSTRACT

With the rapid development of the Internet, the proposal of the Internet of Things has a great significance and wide application scenarios on people's daily lives. IoTs which based on smart integrated sensor devices will lead to a lot of new security and user-privacy challenges in the future. The weakness of security and privacy protection will reduce the people's trust and restrict applications, is one of the influence factors in successful development of IoTs.

This thesis provides a new security privacy policy of information exchange on the sensors layer of IoTs: the spatial anonymous algorithm. It is based on the k-anonymity model which widely used in industries and network for nearly ten years. The paper establishes system model, describes detail content and simulates, analyses, compares, estimates the overall of the algorithm. Recommendations and extension researches for applications of the algorithm in IoTs background are also suggested to improve and strengthen the security privacy protection mechanisms required by IoTs.

Keywords： Internet of Things; Sensor Layer; k-Anonymity Model; Security; Privacy-Preserving

目录

第 1 章 引言 1
1.1 研究背景 1
1.2 k 匿名模型简介 3
1.3 研究现状 4
第 2 章 系统模型 6
第 3 章 匿名空间 SKA 算法 9
3.1 基本架构 9
3.2 详细流程 10
3.3 匿名空间 14
3.4 其他模块 14
第 4 章 性能分析与评估 15
4.1 固定匿名空间 16
4.1.1 数学模式 16
4.1.2 安全级别测试 17
4.1.3 节点密度测试 18
4.1.4 平均速度测试 19
4.1.5 半径测试 20
4.2 动态匿名空间 21
4.2.1 数学模式 21
4.2.2 安全级别测试 23
4.2.3 节点密度测试 24
4.2.4 平均速度测试 25
4.2.5 期望概率值测试 26
4.3 算法性能对比 27
第 5 章 结论 30
插图索引 31

表格索引 32
参考文献 33
声　　明 35
附录 A　外文资料的调研阅读报告 36

第 1 章 引言

1.1 研究背景

因特网从60年代第一次被 ARPANET 提出已经经历过不少的变化。初始四个节点的网络以快速节奏进入了高水平相互连接和自组织的网络时代，专门为日常工作如贸易、调研、经济等提供服务。而使用已逐渐被世界化的网络的人数已经爆发达到15亿，接近于全世界人口的 20% ，那还没算网络内在的各种服务器和路由器。这么让人惊讶的数目已经改变了我们的日常生活和习惯。随着设备的改进和小型化、计算能力的提高、以及能量开销的减少，新一代潮流已经出现：the Internet of Things，也是我们现在所熟悉的“物联网”。未来万物都可以连接到全球因特网以及物体可以互相通信，新的安全和隐私问题也开始提升，比如被物体所识别和交换数据的保密性、可靠性与完整性。人和物的隐私必须得到保证防治不可靠识别和跟踪，而且更多自动和智能物体，更多识别、隐私、责任等问题产生。

物联网自身的特点是由数量庞大的机器构成，设备集群；而这些机器/感知节点大多数部署在无人监控的场景，功能简单(如自动温度计)、携带能量少(使用电池)，完成一些机械的工作，那么攻击者就可以轻易地接触到这些设备，从而对他们造成破坏，甚至通过本地操作更换机器的软硬件。因此，互联万物的小型化系统和感知设备中最大挑战的主题就是隐私和安全问题：如果对私人信息的安全性和保护机制的充分度没有一个承诺确保，用户将不愿意采用这种隐形集成在他们环境和生活当中的新技术。除了提供隐私保护的解决技术之外，更多的手段需要及时建立和应用。

物联网结构可以分为八个部分：

- 通信：提供设备之间的信息交互；
- 传感：捕获和再现并转变实际物理信息成为数码信息；
- 促动：把实际物理的动态在数码信息上演示；
- 存储：存储从传感器、识别和跟踪系统所收集到的数据；
- 设备：在实际物理上与人进行互动；
- 处理：提供数据的使用和服务；
- 定位与跟踪：确定和跟踪实际物理界的位置；

- 识别：在数码信息中提供物体的唯一物理身份证。

每一个部分都有涉及到外部其他可以相同的技术，如通信、传感、识别部分都跟射频自动识别技术 RFID 拥有密切的关系。这些组成部分在完整性、可靠性、安全隐私性、可行性等属性又有不同的要求以及不同的研究现状。

接下来的内容将专门针对于物联网感知层的元信息交互进行分析探讨，它包括两个主要部分：传感与通信，其中“元”代表网络中参加信息互动的设备。首先，感知层是由无数个设备的集成传感器同时构成感知网络组成的。传感信息的完整性和可靠性现在是已相当处理好的研究目标，如水印的提出。而传感信息的安全性是一个弱点，因为一个攻击者可以简单把他自己的感知设备在物理距离不较近就可以感知到相同的信息。因此传感器自身的安全性还是比较低，需要依赖于通信部分的安全策略。传感器的隐私问题又着重针对于实际物理信息，各机制如录像数据的脸部模糊化是需要采用为了保护实际人和物的隐私。传感器的有效性又取决于交互通信的基础设备。另外感知层采用射频自动识别技术 RFID，节点之间是无线传播。而感知网络多种多样，从温度测量到水文监控、从道路导航到自动控制等，它们的数据传输和消息也没有特定的标准，没法提供统一的安全保护体系。因此攻击者很容易在节点之间传播信号中获取敏感信息，从而伪造信号。

可见物联网感知网络的安全隐私性在一定程度上需要借助于通信过程的策略。至今通信协议的研究已经相当的满足完整性、可靠性和安全性，如 TLS 或 IPsec 协议。与此相比，隐私需求方面虽然曾经采用过各种路由策略来解决，如洋葱路由（Onion Routing）或自由网（Freenet），但是都没有满足要求而能得到广泛使用。这也是本论文将提出在物联网感知层中一种新的信息交互机制，匿名空间算法（SKA），的一个重要启发因数，目的在于满足提高各方通信节点的安全隐私级别。

物联网的特殊安全问题很大一部分是由于物联网在现有移动网络基础上集成了感知网络和应用平台带来的。因此，移动网络中的大部分机制仍然可以适用于物联网并能够提供一定的安全性，如认证机制、加密机制等。但还是需要根据物联网的特征对安全机制进行调整和补充。而传统的加密机制则是端到端的，即信息只在发送端和接收端才是明文，在传输的过程和转发节点上都是密文；对于端到端的加密方式来说，它可以根据业务类型选择不同的安全策略，从而为高安全要求的业务提供高安全等级的保护。不过端到端的加密不能对消息的目的地址进行保护，因为每一个消息所经过的节点都要以此目的地址来确定如何传输消息，导致端到端加密方式不能掩盖被传输消息的源点与终点，并容易受到对通信业务

进行分析而发起的恶意攻击。这正好是论文中匿名空间算法 SKA 所要解决的一个主要隐私问题。

1.2 k 匿名模型简介

k 匿名隐私保护机制的模型最早在 1998 年由 P.Samarati 与 L.Sweeney 在论文[2] 提出，目的是为了解决数据发布中通过不同数据表进行的连接攻击（link-attack）所构成的隐私泄露问题。模型中先定义原数据表中所要隐私的信息为敏感属性，其他信息为准标识符；那么通过模型已设计的泛化（generalization）和隐匿（suppression）过程，把原数据表匿名化为新的数据表，使得：所发布的数据表中的每一条记录都有至少 $k-1$ 条记录与它在准标识符上的投影值相等，即每一条记录都有其他 $k-1$ 条记录跟它不可区分。这样也就是 k 匿名条件。以下是一个 $k=2$ 匿名模型的医疗统计例子：匿名化之后每一条记录都有至少其他一条记录与它相同，这样对病症统计工作还是有效，又保护了各人身份的隐私问题：

表1：原始数据表

Birthday	Gender	Zipcode	Disease
02-05-1955	Male	64312	Broken Arm
11-01-1958	Female	64354	Flu
19-10-1958	Female	64303	AIDS
27-07-1959	Female	64312	Hepatitis
12-12-1957	Male	64356	Bronchitis
…	…	…	…

表2：匿名化之后的数据表

Birthday	Gender	Zipcode	Disease
--195*	Male	643**	Broken Arm
--195*	Female	643**	Flu
--195*	Female	643**	AIDS
--195*	Female	643**	Hepatitis
--195*	Male	643**	Bronchitis
…	…	…	…

至今 k 匿名模型在很多领域上得到广泛的应用，如医疗、人口统计、治安系统或者房地产等行业。在运用上，k 匿名模型也不断地针对不同方面而发挥自己的思想，从敏感数据隐私改进在位置匿名隐私采用，后来还在身份隐私方面提起效果，从而各种匿名算法也随着陆续被提出来，如 L.Sweeney 的 Datafly 算法和 MinGen 最小泛化算法、K.LeFerve 的 Incognito 算法、 Mondrian 的多维 k 匿名划分算法等。其中对网络安全隐私领域中 k 匿名模型也有积极的应用，特别是对无线网络的信息传递路由方面，如 ad-hoc 和 cellular 网络中专门针对服务行业的多路匿名路由。

因此，为了满足前边讨论过的安全隐私目标，以及充分利用 k 匿名模型的隐私优点，把网络 k 匿名的概念概括到物联网。在这个环境中每一个设备，感知节点，可以借助于感知层的网络结构进行互相通信。使用 SKA 匿名空间算法让每个感知

节点通过分配不同数据包给其他 k-1 个相邻节点，包括它自己，进行通信而得到满足网络 k 匿名条件。这样真正的原感知节点在 k 个节点中是不可分辨，泄露自己的身份的概率小于 1/k 。很明显，算法提高了交互通信中的隐私水平，但也会需要牺牲开销提高，选择隐私和开销成为一个权衡取舍的问题。论文中给出了 SKA 算法的详解时将给出不同的解决方案，而每个方案都有不同的优缺点也会在文章后部分具体测试、分析和评估，提供给不同场合适当运用算法的目标。

1.3 研究现状

首先介绍 k 匿名模型的相关应用和不同构建算法。文献 [2] 是 k 匿名算法的第一次出现，它主要针对于包括准标识符和敏感属性的发布数据表进行泛化匿名，防止攻击者利用不同数据源实现连接攻击。在发展时期中，保护隐私信息的任务，尽量降低丢失现象、时空复杂度和系统的开销是 k 匿名模型的一个重要研究趋向。

至今 k 匿名模型的实现算法可以分为两大类：静态数据集和动态数据集。模型第一次提出来之后很长一段时间各研究学者在不同情况下创造纯基于静态数据集的各种不同 k 匿名算法：L.Sweeney 在文献 [15] 的 Datafly 算法和 MinGen 最小泛化算法；K.LeFerve 的 Incognito 算法；Mondrian 在文献 [16] 基于 K-D 树的多为 k 匿名划分算法；杨晓春、刘向宇的 Classfly 和 Classfly+ 算；刘向宇与其他人在文献 [7] 的基于特征类的高精度隐私保护数据发布方法等。从结构来看这些算法有一个共同的缺点及时所研究的基本数据集，即 dataset (DS) ，都是静态的，而实际生活当中，已发布的原始数据集却不断得改动：新数据的加入、旧数据的删除、属性的增减等，况且数据内涵的各信息之间都是相互关联的，如果数据集不能及时更新保持一致性，K-匿名算法实际上也是失去了有效性和实用性。解决方法之一是从原数据集从新开始K-匿名化，得到新的数据集满足K-匿名约束，再去发布。但在数据更新量大的情况下，这方法一是使得系统开销加倍得增大，二是产出多重版本的数据，引起隐私信息被泄露的可能上升。这样会使得K-匿名模型失去了原本安全目的的有效性和实用性。因此，在这种更为复杂的条件下，采取何种措施来维持动态数据的隐私保护，是一个值得花下功夫来探讨研究的新主流问题，而至今开始有一些针对数据插入、数据删除和数据更新的解决方法，如：宋金玲、赵威等人在文献 [8] 的 k 匿名数据集的增量更新算法；邓京璟与叶晓俊在文献 [9] 的基于 R 树多维 k 匿名算法；或李金才在文献 [10] 的基于多维桶的 k 匿名表增量更新算法等。例如文献 [13] 提出一个典型的位置路由算法是贪心周长无状态路由

（GPSR: the Greedy Perimeter Stateless Routing）。在 GPSR 中节点的位置信息是可知的，源节点和其他节点知道自己和相邻节点的位置。目的节点的位置携带在路由过程中，中间节点可以根据选择离目标节点最近的节点来确定下一跳传递。在基于位置的路由算法，所携带在数据包中的路由信息只是位置，而节点的 ID 是不需求。因此，匿名目的可以收到保护如果 ID 和位置之间的链接是隐私的。但是为了路由目的，位置信息需要共享给中间节点，使得当一个节点的位置是已知时，它的 ID 可能被泄露，例如一个内在跟踪者可以请求所有节点的位置以便于得到这个连接线索。这实际上是正文介绍的匿名空间算法所要针对解决的一种“连接攻击”。

从另一方面来看随着网络的发展趋向，k 匿名算法的对象也开始改变多样不同。开始时算法针对于数据隐私进行保护，后来它的思想不断地扩大到别的领域，对位置隐私和身份隐私起了不少作用。在无线传感器网络 Wireless Sensor Network (WSN) 中已经有很多基于 k 匿名模型的安全隐私策略研究，例如 K.Mehta 在文献 [17] 的周期采集 PC 与源模拟 SS；M.Shao 等人在文献 [18] 的流量实行强源匿名策略 FitProbRate；Y.Yang 等人在文献 [19] 的基于代理的过滤协议 PFS 和基于树的过滤协议 TFS 等。这些算法的共同优点是提供友好的原位置隐私保护机制，缺点就是换来的开销比较高，如流量大、能耗多、通信延迟长、实时性微弱等；在 adhoc 移动网络 W.Xiaoxin 与 B.Elisa 在文献 [20] 提出了基于 k 匿名的路由协议；Ardagna C.A 等人在文献 [21] 中对 adhoc 和细胞的混合网络服务业务介绍了基于 k 匿名的多路服务策略。

可见 k 匿名模型以良好的隐私保护机制在现代各领域中都得到广泛的应用和适当的研究。这也是本文的一个启发点，把 k 匿名模型的思想运用在物联网感知层的元信息交互问题上，提供给物联网安全隐私方面添加一个有效的算法。

论文接下来的部分将组织如下：第二部分是问题陈述和系统的结构模型。基于 k 匿名模型的匿名空间算法 SKA 将在第三部分进行描述和详解。第四部分针对提出的算法进行功能测试、分析、比较与评估。最后，第五部分总结与今后研究和发展趋。

第 2 章　系统模型

如果算物联网感知层中每一个感知节点当做一个“元”p，那么我们所关心的是各元之间的信息交互模式和它的隐私安全问题。而信息交互方式中当然是由多元参加通信，即存在着双方（两元）信息交互和多方（两元以上）信息交互。因为多方信息交互模式也是基于双元信息交互的基础上来实现。所以这篇文章专门针对于双元之间的信息交互模式来进行研究和分析。

假设（premilinary）：物联网感知层中通信网络能够感知现存的网络环境，通过对所处环境的理解，实时调整通信网络的配置，智能地适应专业环境的变化。在由各设备的传感器一起构成的感知网络中存在着无数感知节点p，它们通过所在网络的特性（如无线移动adhoc网络-MANET，cellular networks，WLAN，WRAN，WIFI，GPRS，3G）进行之间信息交互。

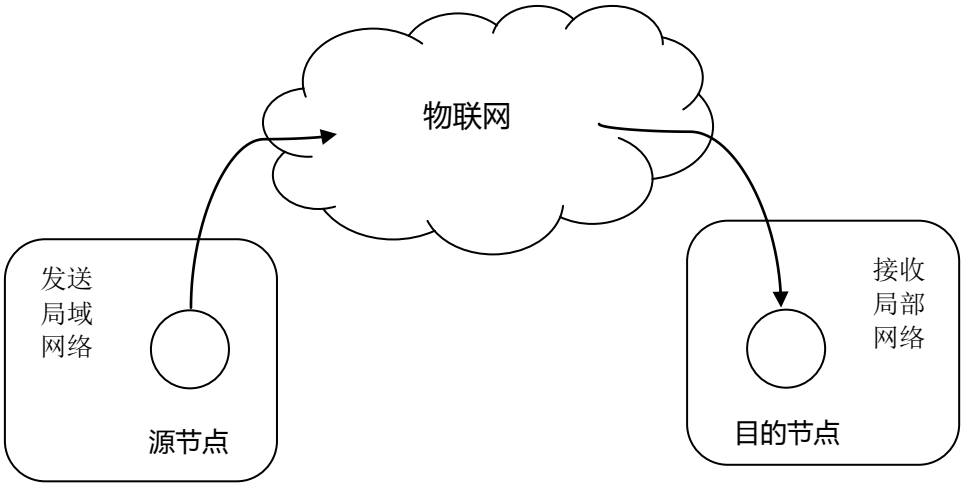

图1：物联网信息交互模型

问题描述：两个感知节点/元进行信息交互。在一个通信过程中，原信息M由一个节点，发送方p_s，发给另外一个节点，接收方p_r。这两个节点通过感知层网络中的各种中间节点来进行传递已被加密的信息M’。因为物联网中有多种多样网络结构，不同场景网络环境也有特殊不同，而每个结构模型都有自己的前提范围，所以在进入匿名空间算法详述介绍之前，先提出以下假设：

表3：符号解释

符号	解释
SKA	k 匿名空间算法
M	原始信息
M’	已加密的信息
P	感知网络节点集
p	感知节点
p_s	发送感知节点
p_r	接收感知节点
K	匿名系数，也是匿名安全级别系数
k_s	发送匿名系数
k_r	接收匿名系数
AZ	匿名空间
S-AZ	发送匿名空间
R-AZ	接收匿名空间
R_{AZ}	匿名空间的半径
U	数据包的发送集
$E_s(M)$	以对称密钥 s 对信息 M 进行加密

网络前提假设：

- 感知网具有相当高的节点密度，即在任何网络空间节点数都不会非常少。另外网络中传输性能的鲁棒性高，在传递空间内任何传递信息都可以正确到达目的地。对于物联网中的那些移动节点，它们以任何方向和任何速度移动。
- 每个节点都可以通过位置服务系统，如文献 [12] 和 [13] 所涉及到的，确定清楚自己的位置。基于位置信息，各位置路由算法如 GPRS 都可以使用。除此，传递路由是对称的，即当一个节点通过多跳连接传递信息时，接收节点也可以回信通过逆行发送路由。

安全隐私前提假设：

- 交互中的两个节点共有彼此之间设好的对称密钥，对称密码算法可以用不同已有的高效算法，如AES、3DES、IDEA等，便于利用它们不同安全方面的优点。

- 交互过程可能遇到内在攻击来自跟踪者或观察者。而且信息包中不包含真实节点的 ID 或者包含但是已经密码化。
- 位置服务系统自身的安全隐私能力是足够强大的，即攻击者没办法知道其他节点访问某个节点的位置的过程和结果。

设计目标：在物联网元信息交互中构造一个防止泄露元位置的交互机制。建立符合不同情况的安全级别。考虑移动设备负载能力，尽量利用不同策略，降低对元存储空间的要求。同时不断地改进算法为了减少传输过程中系统所产生的开销。

论文中所使用的符号意义如表3。

第 3 章　匿名空间 SKA 算法

3.1　基本架构

首先需要了解 k 匿名模型的思想在物联网的感知层中会怎么体现出来，以下介绍两个重要的基本定义：

- 网络 k 匿名：P 是物联网交互节点集，M 是原始从移动或固定的节点 $p \in P$ 发出去的信息。节点 p 是满足网络 k 匿名条件如果发送信息 M 被泄露的时候，p 被发现出来的可能性小于等于 $1/k$，其中 k 是节点p的隐私优先权。换另一种说法，就是当外部攻击在网络中不可分辨通信的真正节点 p 和其他 $k-1$ 个节点，那么节点 p 的通信方式满足了网络 k 匿名条件。
- 匿名空间 AZ (anonymity zone)：是由真实节点（参加交互的节点）和与它相邻的节点（参加匿名化的节点）组合构成的空间。在物联网中因为要符合动静态节点结合的状态所以匿名空间可以是任何结构和形状，文章接下来的算法将采用匿名空间的模型是一个半径为 R_{AZ} 的球形空间。

感知层的元信息交互中匿名空间算法 SKA 的基本过程是：假设节点（设备）p_s 要将信息 M 加密后通过物联网网络（互联网，局域网，无线网等）传送给节点（设备） p_r，一个完整通信过程分为以下三个步骤：

- 阶段 1：准备工作（源节点和目的节点与它们环境中的相邻节点负责）
- 阶段 2：传递过程（网络中的中间节点和定位管理系统负责）
- 阶段 3：接收工作（目的节点与它环境中的相邻节点负责）

其中第 1 阶段和第 3 阶段实现工作时采用 k 匿名模型来匿名化空间，保证通信中源节点和目的节点的安全隐私。

这里值得注意的是，实际上如果感知网络中采用端到端传输模式，是目的节点（接收方）的地址最容易被跟踪并泄露。但在信息交互过程中，信息是来回传递，每一个节点都需要担任发送和接收两个功能，所以每一个节点既是发送方又是接收方，对两方都进行 k 匿名算法是可以理解，而且后面会详解解决算法的另一个目的。

3.2 详细流程

第一个阶段：准备工作

这个阶段由发送方（源节点）和接收方（目的节点）与它们的相邻的节点构成匿名空间来完成：

首先源节点 p_s 根据自己的安全隐私要求选择一个适当的匿名值 k_s 。匿名值越大交互算法所能提供的安全程度越高。接下来，源节点在它的环境空间中选取 k_s-1 个可信相邻节点，收集它们的信息并一起构成 k_s 发送匿名空间 (Send Anonymity Zone, S-AZ) 。因为是匿名化的一部分，这些相邻节点需要有一定的可信度，可以通过统计、历史记录或过去关系等方式来选择，而 k_s 值得选择也有一方面需要先考虑可信相邻节点的数目。同时，目标节点 p_r ，即接收方，也做同样的工作，选出接收匿名值 k_r ，构建 k_r 接收匿名空间 (Receive Anonymity Zone, R-AZ) 。

源节点 p_s 根据它与目标节点共有的对称密钥s把原信息 M 加密得到密文：

$$M' = E_s(M)$$

对称密码机制（symmetric cryptography）可以基于原有的高效加密解密方法，如AES、3DES、IDEA 等来实现。之后源节点以对相邻节点收集过的信息中的负载和传递能力，把密文 M' 分成 k_s 条大小不同的子信息 m_i ，在匿名空间中发给 k_s 个信任相邻节点，包括他自己。

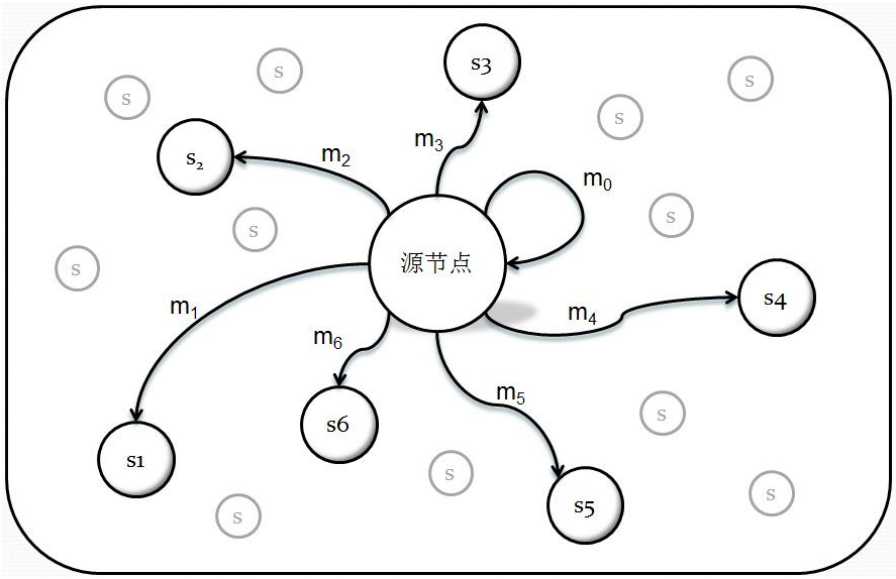

图2：阶段1 - 发送匿名空间

最后，发送匿名空间 S-AZ 中的 k_s 个节点按自己的身份发送它所拥有的信息 m_i

给匿名空间之外的中间节点，传递目标为接收匿名空间 R-AZ 的任意一个节点，开始物联网中感知网络传递数据的阶段。

第二个阶段：物联网传递过程

每一个中间节点收到 k_s 条 m_i 并根据传递路由在感知网络中把信息传给到下一个中间节点，直到 R-AZ。

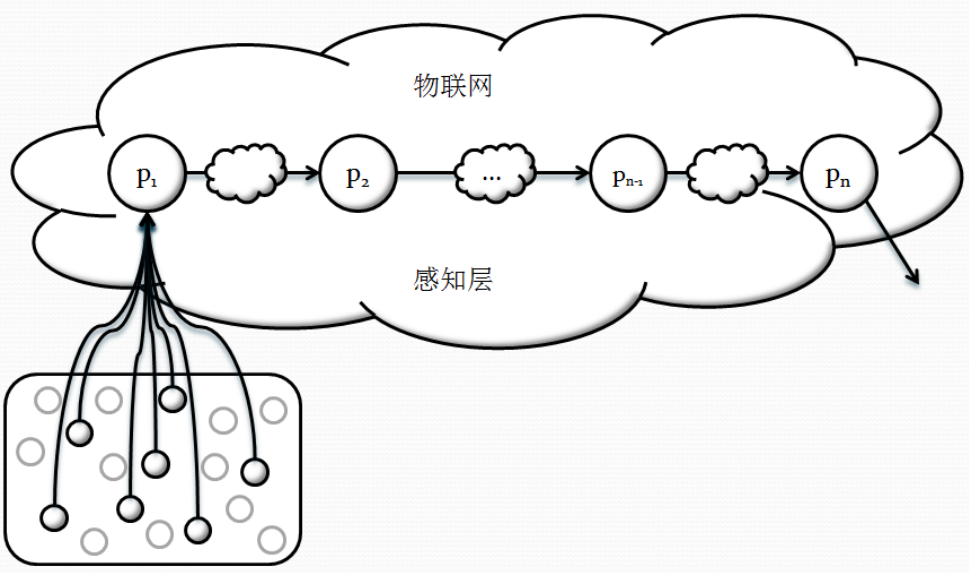

图3：阶段2 - 物联网传递过程

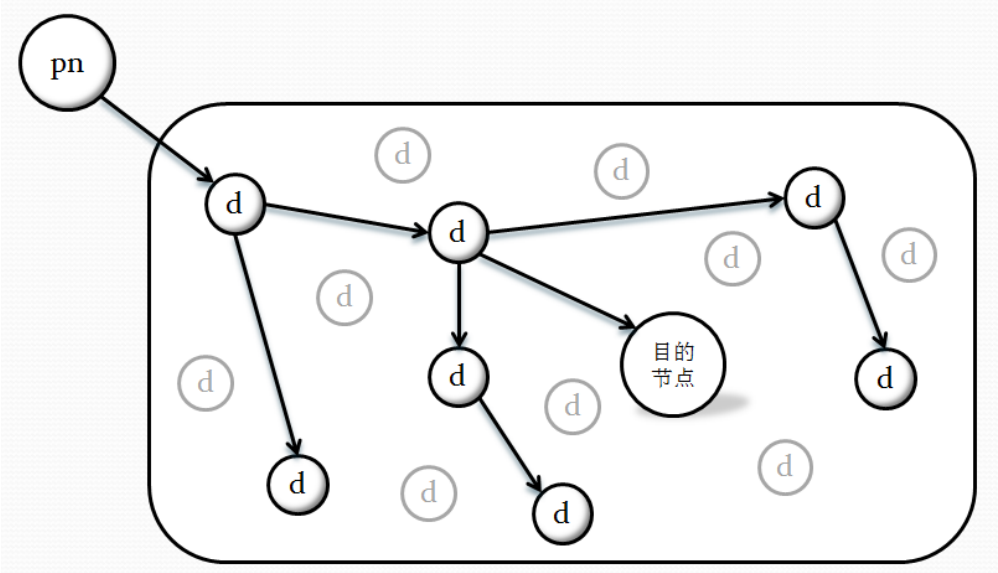

图4：阶段3 - 接收集名空间

值得注意的是，对于第一个中间节点 p_1 来说每一个来自发送匿名空间 S-AZ 的节点有相同的角色，即 k_s 个发送节点都是“源”而哪一个是信息的真正源节点是不确定的，也就是说满足了 k 匿名算法的思想和保护信息交互的隐私目标。

第三个阶段：接收工作

当接收匿名空间 R-AZ 中某个节点接收到从感知网中中间节点传过来的信息 m_i，R-AZ 里边的各节点马上以广播的方式互相传递信息，其中真实的目标节点 p_r 也因此而收到 m_i。

因为只有目标节点 p_r 具有与源节点 p_s 共享的对称密钥 s 所以只有它才可以对密文 m_i 并和之后进行解密，得到原文 M：

$$D_s[\sum_{i=0}^{k} m_{E_s i}] = D_s[M'] = D_s[E_s(M)] = M$$

也在这里特别地说，对于中间节点 p_n 接收匿名空间 R-AZ 的每一个节点d都有相同的角色，即 k_r 个接收节点都是“目的”而哪一个是真正目的节点是不确定的。

协议：**匿名交互协议**

初始：发送方p_s

结束：接收方p_r

参加：物联网中的感知网络（Perception Network – PN）

变量：原始信息：M

发送方p_s与接收方p_r的对称密钥：s

1. 准备工作（s&r）：
   - 1.1：选择适当安全级别 k_s 和 k_r
   - 1.2：构建k_s-发送匿名空间和k_r-接收匿名空间：S-AZ 与 D-AZ
   - 1.3：以对称密钥 s 对原信息 M 进行加密：$E_s(M)=M'$
   - 1.4：生成 M' 的 k_s 个子信息：

     $U_{ks}(M') = \{m_{1Es},m_{2Es},...,m_{ksEs}\}$
   - 1.5：确定包括 k_s 节点的目的匿名空间 D-AZ $\{q_{r1}, ..., q_{rk}\}$（真实接收方r在内），接上$U_{ks}(M')$：

     $U = \{U_{ks}(M')!D\text{-}AZ\} = \{[m_{Es1}!D\text{-}AZ] ,...,[m_{Esks}!D\text{-}AZ] \}$
   - 1.6：发送每一个节点 $q_{si} \in$ S-AZ $\{q_{s1},...,q_{sk-1}\}$ 一个信息包 $[m_{Esi}!D\text{-}AZ] \in U$
   - 1.7：S-AZ 的各节点以自己的身份发送 $[m_{Esi}!D\text{-}AZ]$ 给感知网络PN进行通信
2. 感知网络（PN）：
   - 2.1：接收信息包
   - 2.2：通过报文的路径进行传递过程
3. 接收方（r）：
   - 3.1：从PN接收各信息包 $[m_{Esi}!D\text{-}AZ]$
   - 3.2：D-AZ空间中广播，使得每一个 $q_{ri} \in$ D-AZ 都能收到信息包
   - 3.3：真正接收方qr将信息包合并与解密，得到原始信息M：

$$D_s[\sum_{i=0}^{k} m_{E_s i}] = D_s[M'] = D_s[E_s(M)] = M$$

另外，在从 p_r 传递返回的信息 N 给 p_s 过程中，因为发送方和接收方都是进行了 k 匿名，所以利用已计算好的 S-AZ 和 D-AZ，进行相似的逆向信息传递过程。

3.3 匿名空间

物联网感知层中节点（设备，工具等）是属于动态或者静态的状态，各节点可以往不确定的方向移动，也就是可以移动进或出匿名空间。这种不规则的运动会使得在不同的时间有不同的匿名空间结构而引起“交叉攻击”。例如在不同时间 t_1 和 t_2 匿名空间 AZ 有不同的节点集 set_{t_1} 和 set_{t_2}，它们之间的相同节点会大大减少安全保护的效率，因为真正节点所在的范围现在缩小，只为 set_{t_1} 和 set_{t_2} 的交叉结点，被泄露真实节点的概率从而加倍很多。

因此，怎么确定选取匿名空间，初始时候和后来不同时间段的时候，是匿名空间算法在感知层中的一个关键问题。这部分将提供确定匿名空间的两种解决方案：

固定 AZ：使用大型的空间（半径 R_{AZ} 大），让在 AZ 中有多于 k 个节点，随着时间过去，虽然会有移出一些节点，但仍然有足够 k 个节点在空间中（而移入节点也会使得节点密度变动不多）。

动态 AZ：初始时，AZ 根据节点密度来确定半径 R_{AZ}，所以初始节点数为 k。当时间过去，AZ 的半径变大，曾经移出小的空间 AZ 的节点仍然可以落在更大的空间 AZ。另一种方法来说，当需要时匿名空间的大小可以改变满足总有 k 个节点在 AZ 里边，而这些节点是固定的。它依赖于匿名程度要求，节点密度，移动速度和方向等因素。

3.4 其他模块

以下介绍是k-匿名交互机制的四个主要构成模块，包括：Cipher, AZ_Def, DS_Gen与 Es_Con：

Cipher 模块是专门负责管理密钥，加密，解密功能：KeyGen() 生成两方通信的对称密钥 s；Encrypt() 对所要发送的信息进行加密；$E_s\left(m_i\right)$：已用 s 为密钥的加密信息；Decrypt() 对所收到的信息进行解密。AZ_Def 模块负责确定匿名空间，包括发送和接收两方：选择适当隐私安全级别 k；根据节点密度 d，设备平均负载能力 e，隐私安全级别 k，选择策略（固定/动态）x 来计算适当的匿名空间的半径 $R=f\left(d,e,k,x\right)$。DS_Gen 模块实现信息分割（信息集）：收取各节点负载能力的数据 C；不同设备（节点）有不同负载能力，因此 $E_s\left(M\right)$ 根据 C 分成不同大小的信息段：$U_C\left(E_s\left(M\right)\right)=\left\{m_{1E_s},m_{2E_s},...,m_{k_sE_s}\right\}$。Es_Con 模块参加连接生成发送信心包：将匿名空间 AZ 接上原有的加密数据 m_{iE_s} 得到 $\left[m_{iE_s}\,!D-AZ\right]$。

第 4 章 性能分析与评估

因为 SKA 算法是在发送和接收两方都进行匿名空间，每一方的环境都有可能不同，例如节点密度不同，静动态不同，运动速度和方向不同等。这里先对一般的匿名空间过程进行分析，再对每个情况进行具体应用。

另一方面来看，物联网中的每一个节点（设备，工具等）都可能处于不同的状态，如静态（固定），慢行移动或者快速运动（不同的速度），但是总体来看，固定节点也可以抽象化看成运动节点，只是速度等于零而已。因此后面的分析会以平均速度为$\bar{v}$ 运动的节点进行分析和测试。

首先命名所需要使用的各变量：

- R_{AZ} ：匿名空间的半径
- $\bar{v}$ ：平均移动速度
- S ：匿名空间的面积
- L ：匿名空间的周长
- ρ ：节点密度
- t ：当前时间
- $\bar{t}$ ：平均时间
- t_k ：节点在匿名空间的时间
- $f_{t_k}(t_k)$ ：一个节点还在匿名空间 AZ 的密度概率

后面对固定匿名空间和动态匿名空间进行数学推断、仿真测试以及结果分析，最后对比功能和提出选择实现策略。

4.1 固定匿名空间

4.1.1 数学模式

这部分对采用固定空间的匿名空间算法进行分析。

借助于文献 [11] 的分析和结果，平均时间值可以通过匿名空间的面积、周长和半径来计算：

$$\bar{t} = \frac{\pi S}{\bar{v} L} = \frac{\pi\left(\pi R_{AZ}^2\right)}{\bar{v}\left(2\pi R_{AZ}\right)} = \frac{\pi R_{AZ}}{2\bar{v}}$$

在时间 t 过后，一个节点还在匿名空间 AZ 的概率是：

$$p = p\{t_k > t\} = \int_t^{\infty} f_{t_k}(t_k)\, dt_k = \int_t^{\infty} \frac{1}{t_k} e^{-\frac{t_k}{\bar{t}}} dt_k = e^{-\frac{t}{\bar{t}}}$$

初始时节点数处于匿名空间是 $n_0 = \rho\pi R_{AZ}^2$，在时间 t 过后，除了真实节点还有 i 个节点在匿名空间 AZ 里边的概率是：

$$P\{n = i\} = C_i^{n_0 - 1} p^i (1-p)^{n_0 - i - 1}$$

在时间 t 过后，匿名空间 AZ 还有 k 个节点（包括真实节点）的概率是：

$$P\{n \geq k-1\} = p\left(1 - \sum_{i=1}^{k-1} P\{n = i\}\right)$$

这也是在当前时间 t 匿名空间还能保证 k 匿名条件的可能性。

应用这个结果在 SKA 算法的发送匿名空间 S-AZ 和接收匿名空间 R-AZ 得到这两个匿名空间到当前时间 t 还能满足 k 匿名条件的概率是：

$$P_s\{n_s \geq k_s - 1\} = p_s\left(1 - \sum_{i=1}^{k_s - 1} P_s\{n_s = i\}\right)$$

$$P_r\{n_r \geq k_r - 1\} = p_r\left(1 - \sum_{i=1}^{k_r - 1} P_r\{n_r = i\}\right)$$

由于双方能安全地维持通信当且仅当发送匿名空间和接收匿名空间都还能满足 k 匿名条件，即具有 k_s 和 k_r 个节点在空间里边，所以在时间 t 过后，两个节点之间的通信还能成立的概率是：

$$P\{t\} = P_s\{n_s \geq k_s - 1\} \times P_r\{n_r \geq k_r - 1\} = p_s\left(1 - \sum_{i=1}^{k_s - 1} P_s\{n_s = i\}\right) \times p_r\left(1 - \sum_{i=1}^{k_r - 1} P_r\{n_r = i\}\right)$$

现在使用这个结果在 Matlab 7.10 进行仿真和分析。

4.1.2 安全级别测试

假设感知网络测试时间：$t=500s$；

发送和接收匿名空间中节点平均速度为：$\overline{v_s}=\overline{v_r}=1m/s$；

发送和接收匿名空间中节点密度为：$\rho_s=\rho_r=50\,nodes/km^2$；

发送和接收匿名空间的半径为：$R_{S-AZ}=R_{R-AZ}=300m$；

发送匿名空间的安全级别要求为：$k_s=\{4,8,12\}$；

接收匿名空间的安全级别要求为：$k_r=\{2,10\}$；

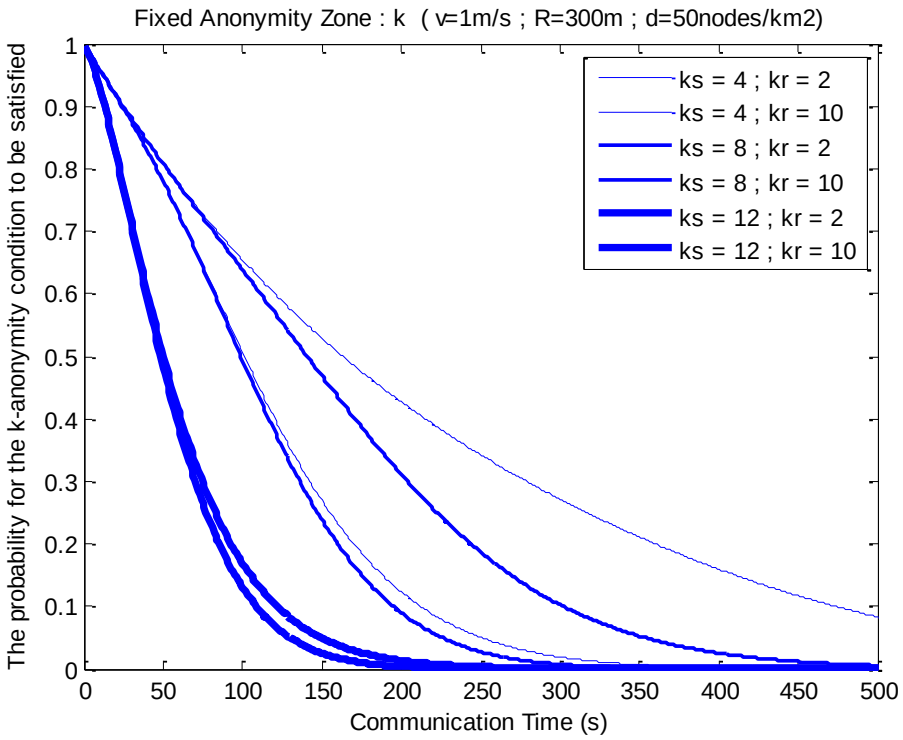

图5：固定匿名空间的安全级别测试

测试得到结果如图5所示。很明显，当安全级别需求越高，即匿名空间的系数 k 越大，那么满足匿名条件的要求更难，成立概率更小，匿名空间需要付出更多开销，如半径选取变大，计算过程变复杂。所以确定 k 匿名值时需要仔细地考虑不同设备和背景的不同安全需求，一般小型的和零散的设备（数量多）需要比较小的 k 安全值，而较大型的和服务器型的设备、工作站（数量少）需要更高保护级别，所以 k 值也随着定高。

另外还需要注意因为匿名空间是选择空间中的原有节点的 k 个节点出来实现你名化，所以 k 的值必须小于初始空间的节点数，即 $k \le n_0 = \rho\pi R_{AZ}^2$。

4.1.3 节点密度测试

假设感知网络测试时间：$t=500s$；

发送和接收匿名空间中节点平均速度为：$\overline{v_s}=\overline{v_r}=1m/s$；

发送和接收匿名空间的安全级别要求为：$k_s=k_r=8$；

发送和接收匿名空间的半径为：$R_{S-AZ}=R_{R-AZ}=300m$；

发送匿名空间中节点密度测试集为：$\rho_s=\{30,40,100\}\,nodes/km^2$；

接收匿名空间中节点密度测试集为：$\rho_r=\{50,200\}\,nodes/km^2$；

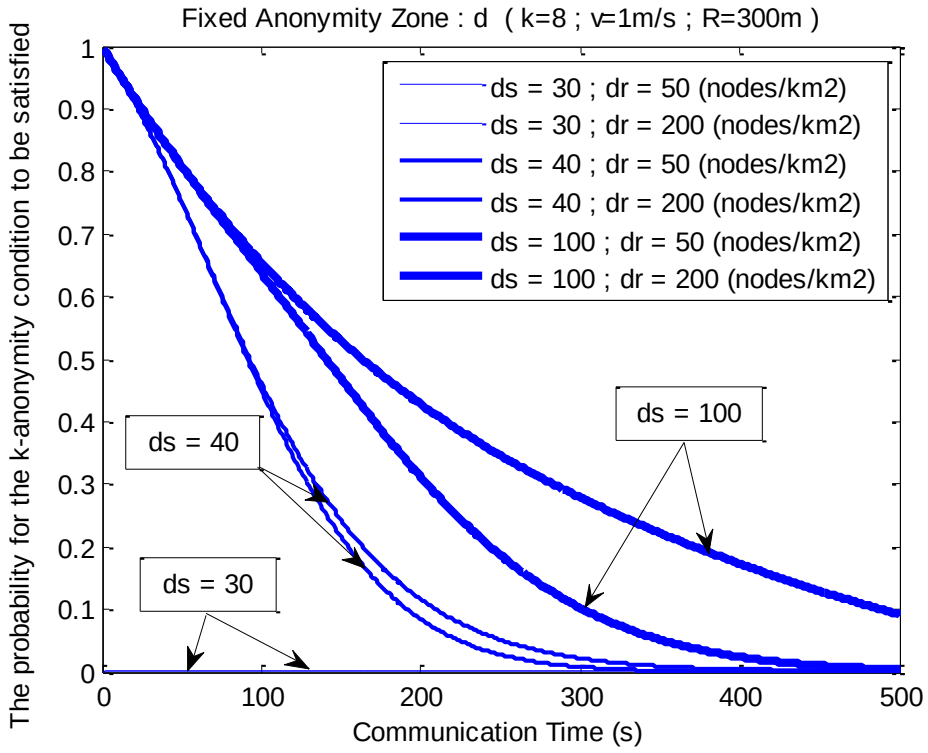

图6：固定匿名空间的节点密度测试

仿真得到结果如图6。此次仿真对一方匿名空间进行三组节点密度测验，分为30，40 和 100 nodes/km^2。从仿真图可见，在给定的环境下每一组测验给的结果不同，很明显，节点密度越大，表明在固定空间中节点数越多，那么通信算法更容易实现。当节点密度为 30 nodes/km^{2}以下，匿名空间成立的概率几乎等于零，通信问题是不可能的。但一超越这个最小值 $\rho_{\min}$，比如节点密度为 40 nodes/km^2，匿名空间成立的概率大大地上升，算法可以实现，而当节点密度是 100 nodes/km^2 时这个概率更为理想，匿名空间成立的可能性更加提高。因此可见，感知层中节点密度对匿名空间具有一定的影响。

4.1.4 平均速度测试

假设感知网络测试时间：$t=500s$；

发送和接收匿名空间的安全级别要求为：$k_s=k_r=8$；

发送和接收匿名空间中节点密度为：$\rho_s=\rho_r=100\,nodes/km^2$；

发送和接收匿名空间的半径为：$R_{S-AZ}=R_{R-AZ}=300m$；

发送匿名空间中节点平均速度为：$\overline{v_s}=\{0,0.5,1\}\,m/s$；

接收匿名空间中节点平均速度为：$\overline{v_r}=\{0,2\}\,m/s$；

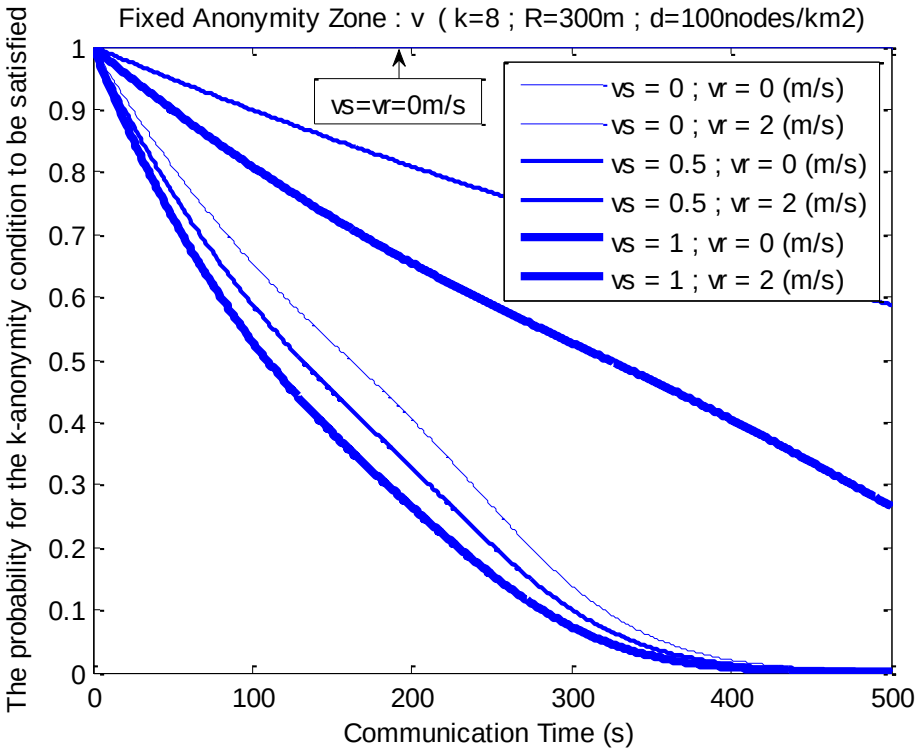

图7：固定匿名空间的节点平均速度测试

测试得到结果如图 7 所示。仿真图中表示，平均速度越小，匿名条件越好成立，它的概率越高。在这里，为了满足物联网中“物”的静动态特性，即又能处于固定又可以不确定方向的移动，测试中设置了具有速度为零的节点，它们相当于固定节点，而速度非零的就是那些移动节点。对于那些固定节点，匿名条件更容易实现满足，特别的是当$\overline{v_s}=\overline{v_r}=0\,m/s$，即网络中双方匿名空间的所有节点都是静态的，将没有节点移出或移入初始匿名空间所以 k 匿名条件总是成立，概率恒等于1（图 7 的最上面一条横线）。换句话说，当物联网感知层所有节点都是固定（静态）的，匿名空间算法总可以实现。

4.1.5 半径测试

假设感知网络测试时间：$t=500s$；

发送和接收匿名空间中节点平均速度为：$\overline{v_s}=\overline{v_r}=1m/s$；

发送和接收匿名空间的安全级别要求为：$k_s=k_r=8$；

发送和接收匿名空间中节点密度为：$\rho_s=\rho_r=100\,nodes/km^2$；

发送匿名空间的半径测试集为：$R_{S-AZ}=\{200,450\}m$；

接收匿名空间的半径测试集为：$R_{R-AZ}=\{250,500\}m$；

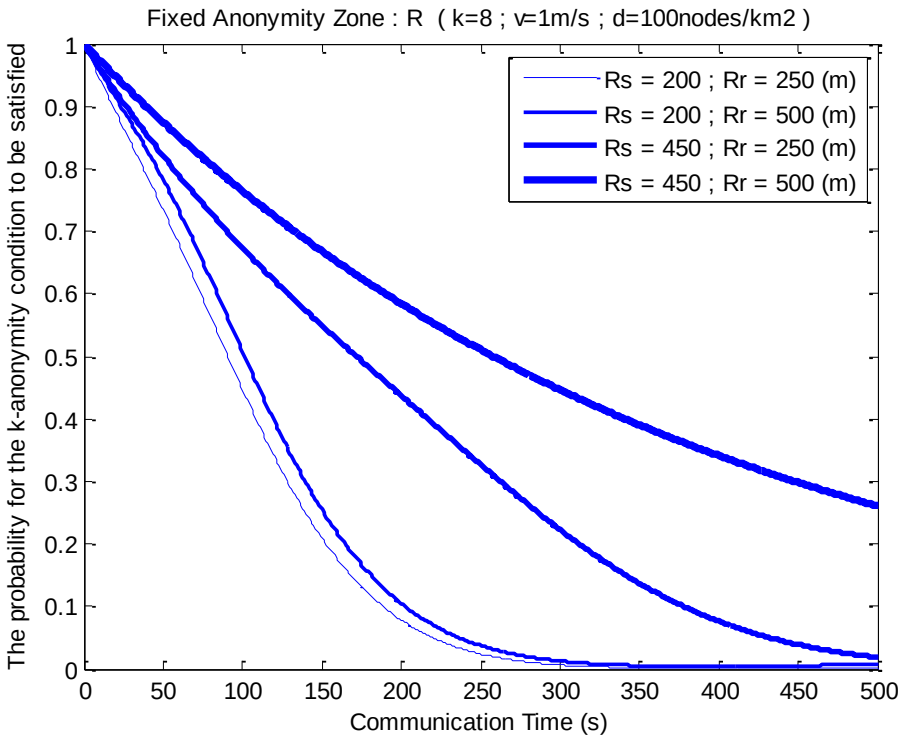

图8：固定匿名空间的半径测试

测试得到结果如图8所示。从图中容易可见当固定匿名空间的半径越大，同一个时间满足 k 匿名条件的概率越大，也就是能维持通信而不需要刷新修改的时间越长，减少了通信过程重新计算匿名空间的工作量，但也带来额外初始计算开销。选择适当的半径 R_{AZ} 会减少通信过程中的开销而初始计算量变动不大，使得系统总体性能提高，例如在上面情况双方空间$\overline{v}=1m/s$、$k=8$、$\rho=100\,nodes/km^2$，只要选择 R_{AZ} 大于 200m 就可以大大地提高维持匿名条件的概率而延长通信时间。

总的来看，在一定的环境下怎么选好匿名空间还是需要从事考虑这环境的各种具体属性。另一方面，通过分析这些方面的影响得出：在给定的节点密度、移动速度、安全级别要求等属性的情况下，如果需要到某一个时间 t 双方节点还能继续通信的概率为 P，那么可以从上述数学公式推导出来匿名空间所需要的半径计算公式，也就是实现了k-匿名空间算法最核心的思想和工作。

除此之外，像之前所描述的，在信息交互过程中，信息是来回传递，每一个节点都会担任发送和接收两个功能，所以每一个节点既是发送方又是接收方。对两方都进行k-匿名算法，当交互时间（次数）长，算法明显降低了计算和构建双方的匿名空间，使得系统的开销也随着减少。

4.2 动态匿名空间

4.2.1 数学模式

借助上面的理论和符号，这部分将对动态匿名空间仿真，测试和分析。

动态匿名空间的一个重要思想是匿名空间什么时候需要改变？改变时是怎么改变的方法？仿真测试之前先对这个问题解决。

匿名空间的初始半径为：

$$R_0 = \sqrt{\frac{k}{\pi\rho}} \qquad (1)$$

在时间 t 过后，原有的 k 个节点仍在匿名空间的概率是：

$$p_k(t) = e^{-k\frac{t}{t}} \qquad (2)$$

随着时间过去，移动节点可以往任何方向移动，有可能移出匿名空间，因此 $p_k(t)$ 会逐渐降低。假设给定一个概率 p_0，当概率 $p_k(t)$ 降低低于这个值，即 $p_k(t) \le p_0$，那么匿名空间需要扩大。

通过上面公式，概率 $p_k(t)$ 降低到 p_0 所用的时间是：

$$t_0 = -\frac{\bar{t}}{k}\ln(p_0) \qquad (3)$$

因此，当过了时间 t_0 半径 R_{AZ} 需要修改为更大，而在 $t = t_0$ 之前，概率 $p_k(t)$ 会按公式 (2) 从1逐渐下降到 p_0。当 $t > t_0$ 时，k 个节点在匿名空间的可能性是：

$$P_k(t) = e^{-\frac{2kvt}{\pi R_{AZ}(t)}}$$

在 $t > t_0$ 想概率 $P_k(t)$ 维持一个不变值，即不随时间改变，简单方法就是定下 $R_{AZ}(t)$ 的公式如下：

$$R_{AZ}(t) = at$$

其中 a 是可选变量。这样 $t > t_0$ 时匿名条件成立的概率为：

$$P_k(t) = e^{-\frac{2kvt}{\pi at}} = e^{-\frac{2kv}{\pi a}}$$

当给定一个所望的概率值 P_0，变量 a 确定如下：

$$a = -\frac{2kv}{\pi \ln(P_0)}$$

那么过了时间 t_0 匿名空间的半径 $R_{AZ}(t)$ 是：

$$R_{AZ}(t) = -\frac{2kv}{\pi \ln(P_0)} t \quad (3)$$

匿名空间的节点数为：

$$n_{AZ}(t) = \rho \pi R_{AZ}^2(t) = \frac{4\rho k^2 v^2}{\pi \ln^2(P_0)} t^2$$

现在对这些结果进行仿真。

实际上 P_0 和 p_0 可以取不同的值。在实验中为了更容易分析所以选择了 $P_0 = p_0$。另外，很明显，动态匿名空间算法不像固定匿名空间算法那样要考虑半径对概率的影响，因为动态算法中初始半径值由给定概率 p_0 根据公式(2)来决定，而在 $t > t_0$ 时它又随时间变化，如公式(3)。但是反过来，动态匿名算法还需要考虑选择期望概率值 p_0 对算法各方面所构成的影响。

4.2.2 安全级别测试

假设感知网络测试时间：$t=50s$；

发送和接收匿名空间中节点平均速度为：$\overline{v_s}=\overline{v_r}=1m/s$；

发送和接收匿名空间中节点密度为：$\rho_s=\rho_r=100\,nodes/km^2$；

发送和接收匿名空间的期望概率值为：$P_{0s}=P_{0r}=0.8$；

发送匿名空间的安全级别要求为：$k_s=\{4,12\}$；

接收匿名空间的安全级别要求为：$k_r=\{6,8\}$；

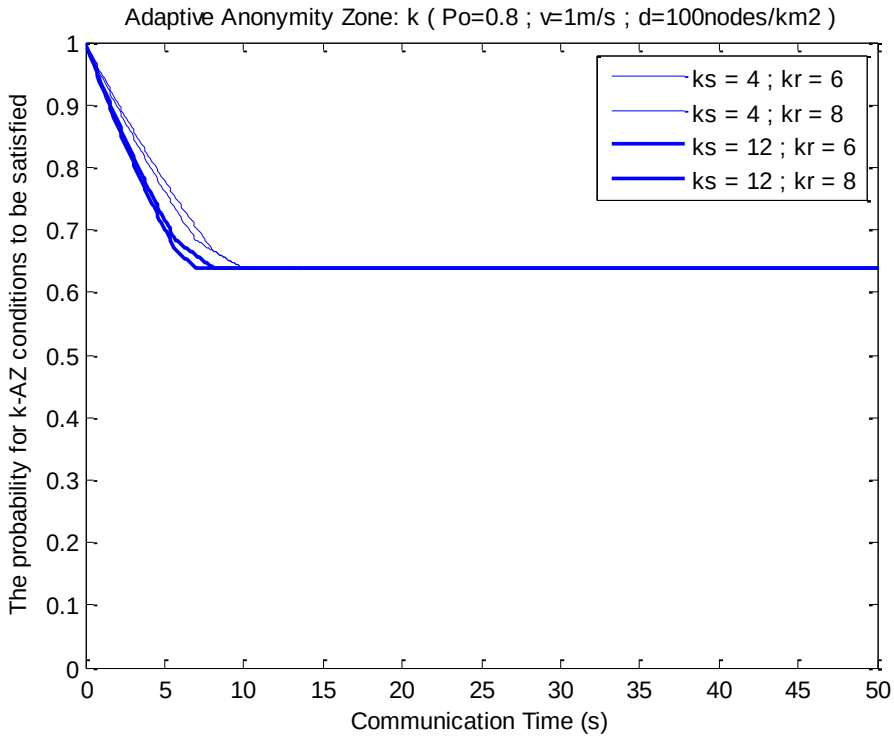

图9：动态匿名空间的安全级别测试

测试得到结果如图9所示。从图可见，匿名条件成立的概率$P_k(t)$随时间分为两个阶段，一开始从概率为1下降，过了时间t_0就进入维持P_0值的阶段。这个时间t_0直接收k选值的影响，当安全级别要求越高，即 k 要求越大，t_0就越小，$P_k(t)$更迅速地从1值下降。如在这测试环境假设中，t_0就取值从7s到10s左右，过这段时间匿名条件成立的概率就固定不变为给定的P_0值。实际上，在变动小范围内，安全级别k不会影响太多动态匿名算法的性能，特别是在长时间情况下。

这里注意的是，每一个匿名空间的期望概率为$P_{0s}=P_{0r}=0.8$，所以整个算法（包括两个匿名空间）的期望概率值为：$P_0=P_{0s}\times P_{0r}=0.8\times 0.8=0.64$

4.2.3 节点密度测试

假设感知网络测试时间： $t=50s$；
发送和接收匿名空间的安全级别要求为： $k_s=k_r=8$；
发送和接收匿名空间中节点平均速度为： $\overline{v_s}=\overline{v_r}=1m/s$；
发送和接收匿名空间的期望概率值为： $P_{0s}=P_{0r}=0.8$；
发送匿名空间中节点密度为： $\rho_s=\{10,100\}\,nodes/km^2$；
接收匿名空间中节点密度为： $\rho_r=\{40,60\}\,nodes/km^2$；

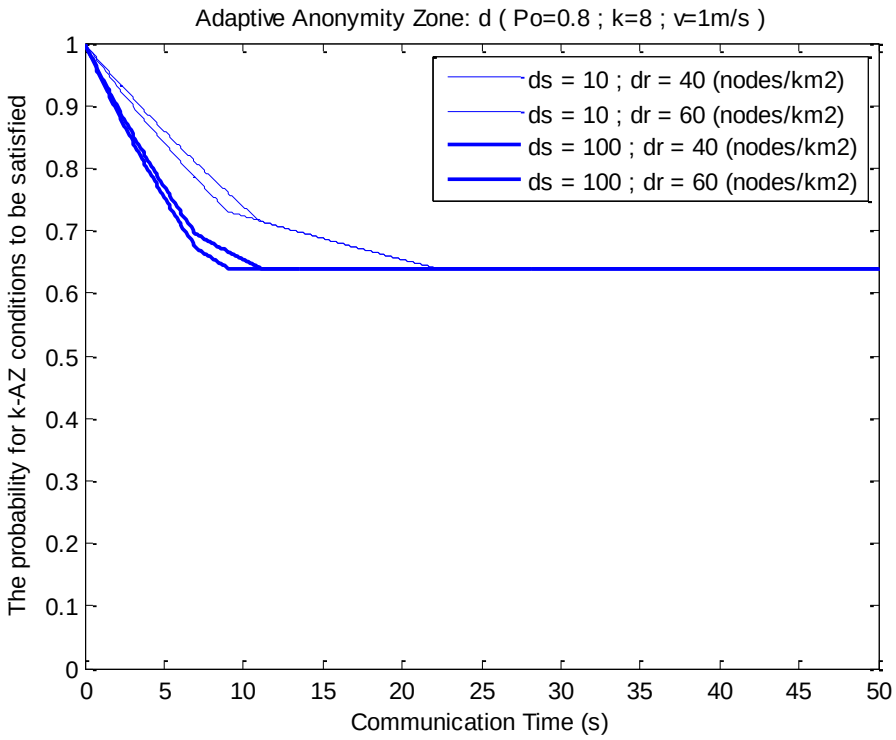

图10：动态匿名空间的节点密度测试

测试得到结果如图10所示。对于节点密度来说，它所对动态匿名算法的影响跟对固定匿名算法的影响完全不同。在固定算法当节点密度越高越有利，节点数多使得提高满足匿名条件的概率。但是在动态算法中半径是动态的，而初始半径又依赖于节点密度。当节点密度大，节点数多，所要选满足匿名条件的初始半径只要小的，如公式(1)所示。因此时间过后小半径的空间能满足匿名条件的概率比起那些大半径空间的更加迅速地降低，更快收敛到所望概率 P_0。可见节点密度在两个算法中体现完全相反的影响。

4.2.4 平均速度测试

假设感知网络测试时间：$t=50s$；
发送和接收匿名空间的安全级别要求为：$k_s=k_r=8$；
发送和接收匿名空间中节点密度为：$\rho_s=\rho_r=100\,nodes/km^2$；
发送和接收匿名空间的期望概率值为：$P_{0s}=P_{0r}=0.8$；
发送匿名空间中节点平均速度为：$\overline{v_s}=\{0,0.5,1\}$；
接收匿名空间中节点平均速度为：$\overline{v_r}=\{0,2\}$；

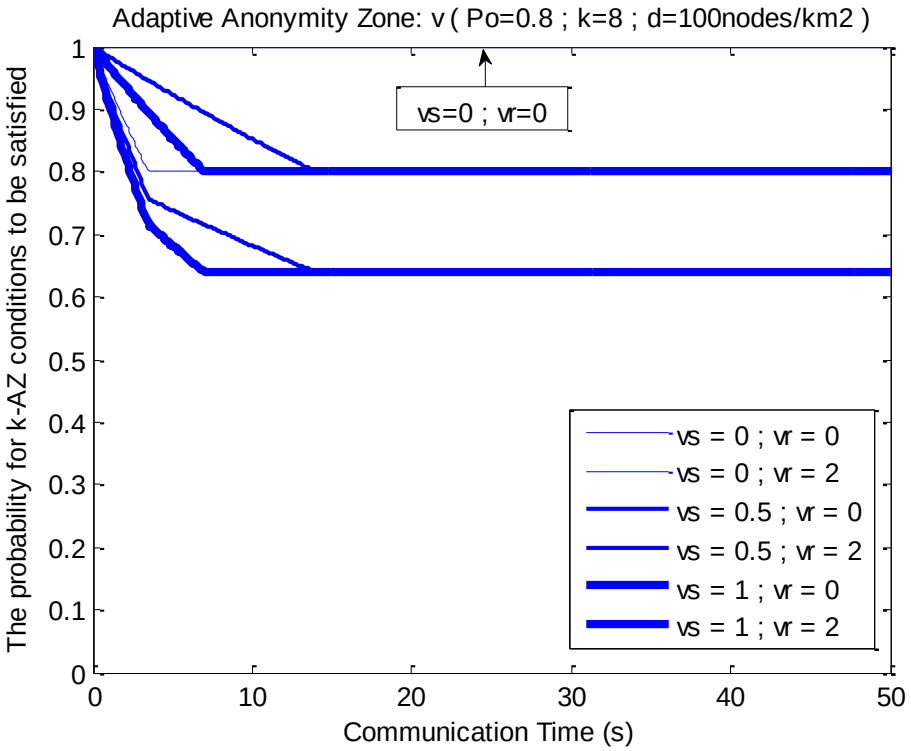

图11：动态匿名空间的节点平均速度测试

测试得到结果如图11所示。当节点平局速度越大，匿名条件成立的概率 $P_k(t)$ 越迅速地回到期望概率值 P_0。跟固定匿名算法相同，当平均速度为零，节点算为固定节点，不会自动移出匿名空间，匿名算法在这些节点上更为容易实现，特别是当 $\overline{v_s}=\overline{v_r}=0\,m/s$，所有节点都是静态的，匿名条件总成立，概率在任何时间都为1，如仿真图中最上方的水平线。

4.2.5 期望概率值测试

假设感知网络测试时间： $t=50s$；
发送和接收匿名空间的安全级别要求为： $k_s=k_r=8$；
发送和接收匿名空间中节点平均速度为： $\overline{v_s}=\overline{v_r}=1m/s$；
发送和接收匿名空间中节点密度为： $\rho_s=\rho_r=100\,nodes/km^2$；
发送匿名空间的期望概率值为： $P_{0s}=\{0.1,0.5,0.9\}$；
接收匿名空间的期望概率值为： $P_{0r}=\{0.3,0.7\}$；

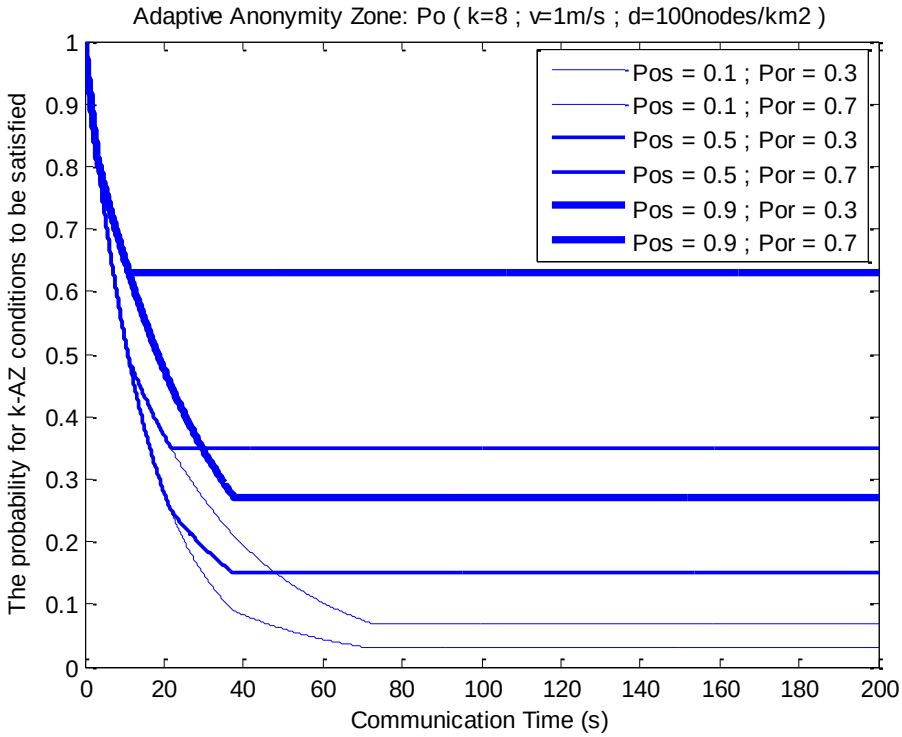

图12：动态匿名空间的期望概率值测试

测试得到结果如图12所示。从图中可见所望概率越小，虽然真实概率随时间迅速降低越快，但是 $P_k(t)$ 从1降低到 P_0 过程所用的时间更宽，因此转变时间 t_0 会更大。这可以用另外方式来表述：在数学公式 (3) 中当 P_0 变小（更接近于0）那么 $\ln(P_0)$ 么值是负值，而且变得更小，所以在其他要素不变的情况下有转变时间 $t_0=-\frac{\bar{t}}{k}\ln(p_0)$ 更大。

总之动态匿名空间法主要分为两个阶段，第一个阶段是匿名条件成立的概率从1快速降低到 P_0 过程，第二阶段是概率保持在给定的期望概率值 P_0 不变。它的特点是在长时间会稳定在某个概率上可持续。但是动态匿名算法最大缺点是需要开销大，在空间和计算复杂度上，这分析将在于文章接下来的比较性能部分更加详解。

4.3 算法性能对比

这部分评估比较上述两种动静态算法的性能。假设感知网的条件如下：
发送和接收匿名空间中节点平均速度为：$\overline{v_s}=\overline{v_r}=1m/s$；
发送和接收匿名空间中节点密度为：$\rho_s=\rho_r=100\,nodes/km^2$；
发送和接收匿名空间的安全级别要求为：$k_s=k_r=8$；
静态算法中发送和接收匿名空间的半径为：$R_{S-AZ}=R_{R-AZ}=300m$；
动态算法中发送和接收匿名空间的期望概率值为：$P_{0s}=P_{0r}=0.8$；

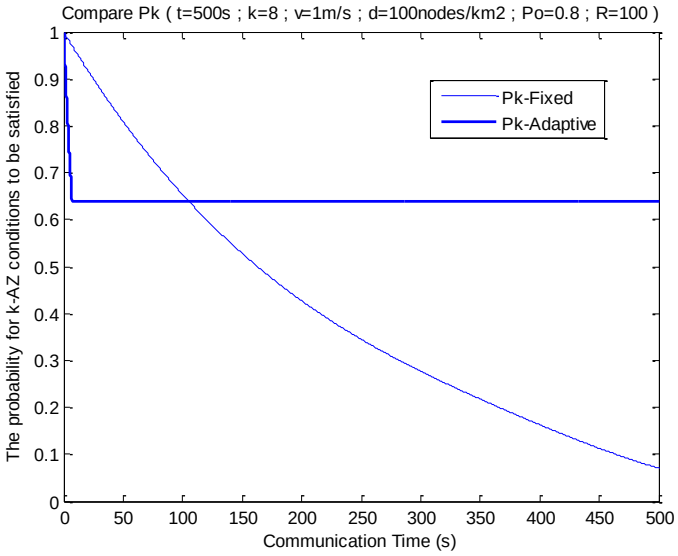

图13：动态和静态匿名算法的匿名条件成立的概率对比

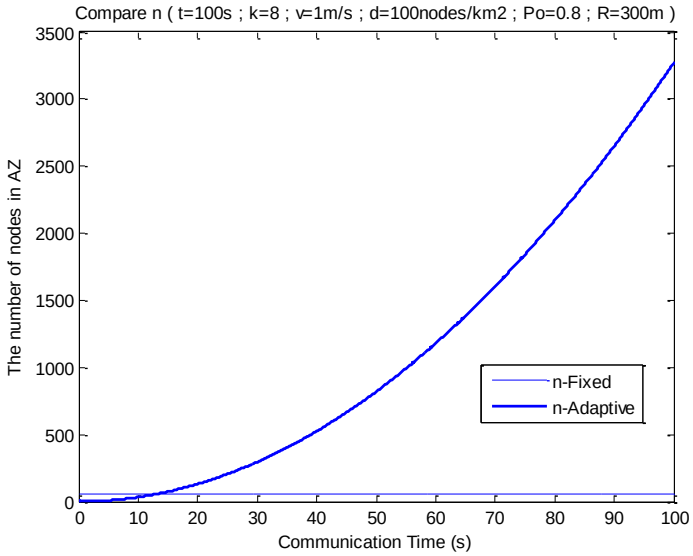

图14：动态和静态匿名算法的节点数对比

表4：算法性能对比

要素	固定匿名空间算法 (Fixed AZ Scheme)	动态匿名空间算法 (Adaptive AZ Scheme)
安全级别 k	k 越大， $P_k(t)$ 值越小	k 越大， $P_k(t)$ 值越小，越迅速从1降低到 P_0
节点密度 ρ	ρ 越大， $P_k(t)$ 值越大	ρ 越大， $P_k(t)$ 值越小，越迅速从1降低到 P_0
平均移动速度 $\bar{v}$	$\bar{v}$ 越大， $P_k(t)$ 值越小	$\bar{v}$ 越大， $P_k(t)$ 值越小，越迅速从1降低到 P_0
半径 R_{AZ}	R_{AZ} 越大， $P_k(t)$ 值越大	无（半径随时间改变）
期望概率值 P_0	无（概率随时间改变）	P_0 越大，转变时间 t_0 越短，阶段一越快结束。
可行概率 $P_k(t)$	随时间降低： $P\{t\}=p_s\left(1-\sum_{i=1}^{k_s-1}P_s\{n_s=i\}\right)$ $\times p_r\left(1-\sum_{i=1}^{k_r-1}P_r\{n_r=i\}\right)$ 收敛从1到0。	阶段一：$t<t_0$，随时间降低（比较短） $P_k(t)=e^{-\frac{2kvt}{\pi R_{AZ}(t)}}$ 阶段二： $t>t_0$，维持在给定期望值 $P_k(t)=P_0$
节点数 n	最大节点数不取决于时间 由初始匿名半径确定： $n_0=\rho\pi R_{AZ}^2$ 一般不太大。	随着时间迅速增加： $n_{AZ}(t)=\rho\pi R_{AZ}^2(t)=\frac{4\rho k^2v^2}{\pi\ln^2(P_0)}t^2$ 越来越大。

上面仿真图13和图14两种仿真结果完全体现出来两个匿名空间算法的主要思想和差别：静态匿名算法使用的是固定半径，随时间半径不会伸缩所以移出空间的那些节点使得满足匿名条件的可能性也随时间降低，过一段时间收敛为零 $P\{t\}=p_s\left(1-\sum_{i=1}^{k_s-1}P_s\{n_s=i\}\right)\times p_r\left(1-\sum_{i=1}^{k_r-1}P_r\{n_r=i\}\right)$；而也因这些移出节点使得静态匿名空间最多要处理的节点保持固定为 $n_0=\rho\pi R_{AZ}^2$，这个值依赖于 R_{AZ} 的选取，一般不会很大。与此相比，动态匿名算法有不同的启发性：在长时间内动态算法成立的概率值会稳定在一个期望值 $P_0=e^{-\frac{2kv}{\pi a}}$，可调大调小根据不同的目的，但是期望值越大，所要付出的开销也越高；另外从图14可看动态算法所要处理的节点数是

随着匿名空间的半径的膨胀而迅速提高 $n_{AZ}(t)=\rho\pi R_{AZ}^2(t)=\frac{4\rho k^2 v^2}{\pi \ln^2(P_0)}t^2$，像上述的例子在 30 秒通讯已经要对 250 个节点左右打交道，可比静态算法的节点数目多得多，解决方法就是可以使用刷新策略来减少节点数，即每时间隔进行刷新节点记录，重新计算初始半径并开始新的通信阶段。

另外从上部分进行综合，根据匿名空间的不同特性可以给出一个对比如表4所示。可见每个算法都有自己的优缺点以及适用场合。在物联网感知层中选择什么算法，是静态还是动态，和算法中怎么选择适当参数同时还满足安全要求，这需要仔细考虑具体网路的本质和特点，选择不同的高效和重点功能。比如在节点密度比较高，而且通讯时间相当长，可以采用固定匿名空间算法；而当节点密度少，时间通信也比较短，特别是强制要求某一个特定成功率，应该选取动态匿名空间算法来实现。另外，可以使用刷新手段来确保算法的性能和开销（如上面所陈述过），特别的是在刷新期间内完全可以切换两种方法之间，用来弥补它们之间的缺点，这也是动静态混合算法的思想，能满足提高和平衡整个系统的安全方面与其他功能。

第 5 章 结论

物联网的出现带动了人类中一个新潮流、新体验，同时也带来了不少挑战，而最大的挑战之一就是物联网的安全与隐私问题。只当安全与隐私方面都得到足够的对待和发展，物联网才能得到广泛的应用。处于现在的初级发展阶段，物联网安全隐私方面有一定的利益基于因特网已有的安全隐私基础，其中 k 匿名模型是可迁移思想进去物联网安全隐私保护领域的一个有效机制。

通过介绍匿名空间算法的背景、详细描述算法的流程和不同策略以及深入进行算法的性能分析与评估，本论文已经对物联网感知层的元信息交互安全隐私保护机制在给定的假设条件下提出了新算法的一个基准框架。

后面工作将着重于扩展到网络更一般的条件，它们对信息交互有何影响、影响的程度以及怎么有效地改进运用算法。例如：周围节点不是所有都可信，出现转发情况（因为隐私问题，有节点不愿意直接帮助在感知网络中发送信息，但是可以中转任务给其他节点发送，相当于构造匿名空间 AZ 时使用 multihop 寻找可信相邻节点）；感知层中局部网络的鲁棒性比较低，即信息传递不能确保可以百分之百到达目的设备，网络还会带有延迟，那么在不同传递情况发生会如何对待处理；或者物联网中设备位置比较分散，也即节点密度没有足够大，会降低匿名算法的效率和完整性，等等。因此，进一步扩展研究是一个重要有益工作，有助于匿名空间算法对物联网应用场景更加完善和灵活有效。

插图索引

图 1. 物联网信息交互模型 6
图 2. 阶段 1 - 发送匿名空间 10
图 3. 阶段 2 - 物联网传递过程 11
图 4. 阶段 3 - 接收集名空间 11
图 5. 固定匿名空间的安全级别测试 17
图 6. 固定匿名空间的节点密度测试 18
图 7. 固定匿名空间的节点平均速度测试 19
图 8. 固定匿名空间的半径测试 20
图 9. 动态匿名空间的安全级别测试 23
图 10. 动态匿名空间的节点密度测试 24
图 11. 动态匿名空间的节点平均速度测试 25
图 12. 动态匿名空间的期望概率值测试 26
图 13. 动态和静态匿名算法的匿名条件成立的概率对比 27
图 14. 动态和静态匿名算法的节点数对比 27

表格索引

表 1. 原始数据表 3
表 2. 匿名化之后的数据表 3
表 3. 符号解释 7
表 4. 算法性能对比 28

参考文献

[1] Mayer C. Security and Privacy Challenges in the Internet of Things. Electronic Communications of the EASST, 2009, 17.

[2] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. In Proc. of the Seventeenth ACM SIGACTSIGMODSIGART Symposium on Principles of Database Systems, 1998, page 188.

[3] Sweeney L. K-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570.

[4] Ciriana V, De Capitani S, Foresti S, and Samarati P. k-Anonymity. Springer US, Advances in Information Security, 2007.

[5] Klinov P, Taylor J M, Mazlack L J. Formal treatment of imprecision in the Semantic Web on tologies. http://citeseerx.ist.psu.edu/viewdoc, 2006.

[6] Zadeh L. Fuzzy sets [J]. Information and Control, 1965, 8(3) : 338-353.

[7] 刘向宇, 杨晓春, 于戈. 一种基于特征类的高精度隐私保护数据发布方法[J]. 计算机科学, 2005, 32(7): 368-373.

[8] 宋金玲, 赵威, 刘欣, 黄立明, 李金才, 刘国华. K-匿名数据集的增量更新算法.计算机科学, 2010, 37(4).

[9] 邓京璟, 叶晓俊. 基于R 树多维K-匿名算法[J]. 计算机工程, 2008, 34(1): 80-82.

[10] 李金才, 吕艳丽, 赵威, 刘国华, 李宏佳. 基于多维桶的-匿名表增量更新算法. 燕山大学学报, 2009, 33(5).

[11] R. Thomas, H. Gilbert, and G. Mazziotto. Influence of the moving of the mobile stations on the performance of a radio cellular network. In Proceedings of Third Nordic Seminar, 1988.

[12] J. Li, J. Jannotti, D. S. J. D. Couto, D. R. Karger, and R. Morris. A scalable location service for geographic ad hoc routing. In ACM Mobicom, 2000.

[13] X. Wu. Vpds: Virtual home region based distributed position service in mobilead hoc networks. In Proc. of ICDCS, 2005.

[14] B. Karp and H. T. Kung. Gpsr: Greedy perimeters stateless routing for wireless network. In Proc. of MOBICOM, 2000.

[15] Sweeney L. Achieving K-anonymity Privacy Protection Using Generalization snd Suppression [J]. Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 2002, 10(5): 571-588.

[16] LeFevre K, DeWitt D, Ramakrishnan R. Mondrian Multidimensional K-anonymity [C]. In Proc. of the 22nd International Symposium on Data Engineering. USA: [s. n.], 2006

[17] Mehta K，Liu D，Wright M.Loeation Privacy in Sensor NetWorks Against a Global Eavesdropper[C]. In Proc.of IEEEICNP’07.2007

[18] Shao M，Yang Y，Zhu S，et al.Towards Statistically Strong Source Anonymity for Sensor Networks[C]. In Proc.of IEEE INFOCOM'08.Phoenix AZ，USA:2008.

[19] Yang Y，Shao M，Zhu S，et al.Towards event source unobservability with minimum network traffic in sensor networks[C]. In Proc.of ACM WISec，08.NewYork，NY，USA:2008.

[20] Xiaoxin W, Elisa B. Achieving K-anonymity in Mobile Ad Hoc Networks. In Proc. of First Workshop on Secure Network Protocols, 2005, page 42.

[21] Ardagna C.A, Stavrou A, Jajodia S, Samarati P, Martin R. A multi-path approach for k-anonymity in mobile hybrid networks, 2008.

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：_____ 日 期：_____

附录 A 外文资料的调研阅读报告

THE PRIVACY PROTECTION MECHANISM BASED ON K-ANONYMITY OF THE COMMUNICATION IN THE INTERNET OF THINGS

Abstract: Based on the internet, the Internet of Things uses RFID, wireless data communications and many other technologies to cover everything in the world. In this network, the goods (commodities) can "exchange" to each other without human intervention. It uses radio frequency identification (RFID) technology through Internet to identify goods and the interconnection of information automatically. However, it also has a huge problem: these information could be leaked to other people, such as competitors. Therefore, in the perception, transmission, application process, how can we do to make information available just for ourselves, but not be used by others, especially not to be used by competitors? It, of course, requires more efforts to form a strong security system. This paper introduces basic concepts of privacy protection of IoT and one of its privacy security algorithms: k-anonymity model.

Keywords: Internet of Things, Security, Privacy, RFID, K-anonymity

1. The Internet of Things and Privacy Policy

The Internet of Things, IoT, is known as a self-configuring wireless network of sensors which interconnect all things on the world. The concept of Internet of Things was founded in 1999 at the Massachusetts Institute of Technology (MIT).

The Internet of Things can be sorted out eight topics: communication, sensors, actuators, storage, devices, processing, localization and tracking, identification. C.Mayer in his paper[1] analyzed sensitivity of those topics in the IoT to different security and privacy properties as below:

Topic \ Property	Integrity	Authenticity	Confidentiality	Privacy	Availability	Regulation
Communication	+++	+++	+++	++	+++	+
Sensors	+++	++	+	+++	+	+++
Actuators	+	+	+		+	++
Storage	+++	++	+++	+++	+	+++
Devices	+++	+	+	++	++	++
Processing	++	+	+	+++	+	+++
Localization/Tracking	+	+	+++	+++	+++	+++
Identification	++	+	+++	+++	+++	+++

(+ is low sensitivity, ++ is middle sensitivity, +++ is high sensitivity)

The key technologies of protecting privacy mechanism of the Internet of Things includes: Radio-frequency identification (RFID) and secure multi-party computation (SMPC).

Radio-frequency identification (RFID), which purposes to identify and track, is a technology that uses communication via radio waves to exchange data between a reader and an electronic tag attached to an object.

Secure multi-party computation (SMPC), also known as multi-party computation (MPC), is a problem that was initially defined by Andrew C. Yao in a 1982 paper. Yao suggested the millionaire problem in that paper: how to find out who is richer without between two millionaires, Alice and Bob, without revealing how much they have. Yao gave a solution named secure multi-party computation to satisfy those guys' curiosity while still keeps the constraints.

There are some other technologies, for example, password management, secure routing protocols, authentication and access control, intrusion detection and fault-tolerant technologies, etc. The next part of paper introduces one of them, k-anonymity model, which can prevent linking attacks.

2. k-anonymity model

k-anonymity model[2][3] was initially suggested by P.Samarati and L.Sweeney in the paper "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" submitted to IEEE Security and Privacy 1998.

At first, we take an overview of what linking-attack is. When information is more commonly shared, exchanged and sold, the shared information can be sensitive to individuals or organizations. Therefore, we need to find out how to not only release information freely but also make the identity of any individual included in the data

unavailable to be recognized. If it successes, information could be shared freely with more purposes. Attackers, however, could still retrieve private information by linking to external data.

For example: the hospital publishs a medical data released as anonymous, like table1.

Table1: Medical data released as anonymous

Birthday	Gender	Zipcode	Disease
02-05-1955	Male	64312	Broken Arm
11-01-1958	***Female***	***64354***	***Flu***
19-10-1958	Female	64303	AIDS
27-07-1959	Female	64312	Hepatitis
12-12-1957	Male	64354	Bronchitis
…	…	…	…

In the other hand, attackers can also get information from the voter list, like table 2.

Table2: Voter list

Name	Birthday	Gender	Zipcode
Tyson	12-12-1957	Male	64354
Allen	02-05-1955	Male	64312
Miley	***11-01-1958***	***Female***	***64354***
Serena	27-07-1959	Female	64312
Sofia	19-10-1958	Female	64303
…	…	…	…

The attackers use birthday, gender, zipcode to re-identify individuals, and result is: Miley, whose birthday is 11-01-1958, gender is female and zipcode is 64354, would be recognized with “flu” disease! So, k-anonymity model is known as privacy protection which can pretent linking attacks.

Two basic concepts of k-anonymity model are:

- quasi-identifier (QI): is the set of attributes which links private data set with external public data set.

- sensitive attribute (SA): is the set of attributes which need to be protected privacy.

In this case, Birthday, Gender, Zipcode are quasi-identifiers; Diease is sensitive attribute.

Two key techniques of k-anonymity model are:

- generalization: generalize values of attributes

  For example:

  Zipcode={64303, 64312, 64354, 64356}

  => Zipcode={6430*, 6431*, 6435*}

  => Zipcode={643**}

- suppression: ignore some attributes (such as Birthday attribute) in released data.

V.Ciriana and P.Samarati in a 2007 paper[4] defined the constraint conditions of k-anonymity as below:

Definition 1 (k-anonymity requirement): Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents.

Definition 2 (k-anonymity): Let T (A1, …, Am) be a table, and QI be a quasi-identifier associated with it. T is said to satisfy k-anonymity with respect to QI iff each sequence of values in T [QI] appears at least with k occurrences in T [QI].

After using these constraints of k-anonymity model to change original data set in table 1, we get results as below table 3:

Table3: k-anonymity model data set

Birthday	Gender	Zipcode	Disease
--195*	Male	643**	Broken Arm
--195*	Female	643**	Flu
--195*	Female	643**	AIDS
--195*	Female	643**	Hepatitis
--195*	Male	643**	Bronchitis
…	…	…	…

It is easy to be aware that the new data set is satisfied the constraint conditions of 2-anomynity model (k=2).

The k-anonymity privacy protection model is more and more important and public. There are many implementations of k-anomynity, as Datafly[5], m-Argus[6] and Classfly[7], etc. From structural point of view, howerver, the disadvantage of most implementations of k-anomynity so far is that the data set is set to static without considering the released data is actually tend to change: the addition of new data, the deletion of old data, change attributes, etc.

One solution is to re-implement k-anonymity on original data set to get new data set which meets the k-anonymity constraint. However, in the case of large amount of data updates, this approach does not only increase overhead , but also creates multi-version of the data output which causing leakage of private information. Therefore, the k-anonymity model would lose effectiveness and practicality of security purposes.

In these more complex conditions, it is worth to explore what measures taken to maintain privacy of dynamic data. Up to now, there are some solutions solved data insertion, data deletion, and data updating, such as: the k-anonymous incremental updating algorithm[8] of Song Jinling and Zhao Wei; k-anonymous algorithm based on multi-dimensional R tree[9] of Deng Jingjing and Ye Xiaojun ; or k-anonymous incremental updating algorithm based on multi-dimensional barrel of Li Jinyuan.

For example, the k-anonymous incremental updating algorithm of Song Jinling and Zhao Wei gets result as below:

DELETE(R, (Age<25 & Zipcode<64320))

Figure1: original data set

ID	Age	Gender	Zipcode	Disease
t1	24	Male	64312	Broken Arm
t2	35	Female	64354	Flu
t3	33	Female	64333	AIDS
t4	40	Female	64333	Hepatitis
t5	25	Male	64312	Bronchitis
…	…	…	…	…

Figure2: released data set before deleting

ID	QG	Age	Gender	Zipcode	Disease
t1*	1	[20, 30]	Male	[64300, 64320]	Broken Arm
t2*	2	[30, 50]	Female	[64330, 64360]	Flu
t3*	2	[30, 50]	Female	[64330, 64360]	AIDS
t4*	2	[30, 50]	Female	[64330, 64360]	Hepatitis
t5*	1	[20, 30]	Male	[64300, 64320]	Bronchitis
…	…	…	…	…	…

Figure3: released data set after deleting

ID	QG	Age	Gender	Zipcode	Disease
t2*	2	[20, 50]	*	[64300, 64360]	Flu
t3*	2	[20, 50]	*	[64300, 64360]	AIDS
t4*	2	[20, 50]	*	[64300, 64360]	Hepatitis
t5*	2	[20, 50]	*	[64300, 64360]	Bronchitis
…	…	…	…	…	…

3. Conclusions

Internet of Things is not fantasy technology, but is another technological revolution. After the computer and the Internet, IoT is known as the third wave in the development of information industry with the intelligent perception, identification technology and pervasive computing. The security and privacy protection, in fact, is the key for IoT services to apply in large-scale. As an effective method of security and privacy protection, k-anonymity model has become a new bright spot which is worth exploring, studing and proper applying in Internet of Things field.

References:

[1] Mayer C. Security and Privacy Challenges in the Internet of Things. Electronic Communications of the EASST, 2009, 17.

[2] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. In Proc. of the Seventeenth ACM SIGACTSIGMODSIGART Symposium on Principles of Database Systems, 1998, page 188.

[3] Sweeney L. K-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570.

[4] Ciriana V, De Capitani S, Foresti S, and Samarati P. k-Anonymity. Springer US, Advances in Information Security, 2007.

[5] Klinov P, Taylor J M, Mazlack L J. Formal treatment of imprecision in the Semantic Web on tologies. http://citeseerx.ist.psu.edu/viewdoc, 2006.

[6] Zadeh L. Fuzzy sets [J]. Information and Control, 1965, 8(3) : 338-353.

[7] 刘向宇, 杨晓春, 于戈. 一种基于特征类的高精度隐私保护数据发布方法[J]. 计算机科学, 2005, 32(7): 368-373.

[8] 宋金玲, 赵威, 刘欣, 黄立明, 李金才, 刘国华. K-匿名数据集的增量更新算法.计算机科学, 2010, 37(4).

[9] 邓京璟, 叶晓俊. 基于R 树多维K-匿名算法[J]. 计算机工程, 2008, 34(1): 80-82.

[10] 李金才, 吕艳丽, 赵威, 刘国华, 李宏佳. 基于多维桶的-匿名表增量更新算法. 燕山大学学报, 2009, 33(5).