doi:10.3969/j.issn.1000-436x.2013.z2.022

校园无线网用户群体的移动行为聚集分析

周昌令[1,2]，钱群[1]，赵伊秋[1,2]，尚群[1]
（1. 北京大学 计算中心，北京 100871；2. 北京大学 信息科学技术学院，北京 100871）

摘 要：寻找更好更高效的计算用户之间相似度的方法是个难题，聚集结果对网络运维的帮助也较少被关注。提出了终端移动轨迹的稀疏链接区间（SLI, sparse linked intervals）概念，以此为基础使用社会网络分析的方法有效地分析了移动终端的聚集关系。主要采用了北京大学无线校园网真实的实际运行数据进行分析，并用公开数据集进行了验证。实验结果表明，提出的方法能够很好地发现用户群体。还分析了 3 种常见的聚集层次子图模式，以及它们的形成原因和与无线网络管理的联系。

关键词：无线网络；移动轨迹；稀疏链接区间；社交网络分析；相似性
中图分类号：TN925.93 **文献标识码**：A **文章编号**：1000-436X(2013)Z2-0111-06

Modularity analysisof users'mobile behavior in campus wireless network

ZHOU Chang-ling[1,2], QIAN Qun[1], ZHAO Yi-qiu[1,2],SHANG Qun[1]
(1. Computer Center, Peking University, Beijing 100871, China;
2. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

Abstract: Choosing a better and easily computed similarity metric is a challenge task, and the problem how the formed-modularity can help network operation attracts less attention. By introducing sparse linked interval(SLI) to represent wireless nodes' mobile trajectory, mobile nodes' modularity relations via social network analysis methods were effectively revealed. Using real operational datasets of Peking University campus wireless network and public dataset as validation, three common motifs in modularity level, and their formation reasonsand relations to network managements were demonstrated.

Key words: wireless network; mobile trajectory; sparse linked interval; social network analysis; similarity

1 引言

近年来无线网络技术在整个网络技术中发展得非常迅速，各种无线终端（例如笔记本、平板电脑以及智能手机等）层出不穷，先进的无线网技术让人们可以每时每刻使用互联网。飞速增长的用户群对无线网络的性能提出了更高的要求，无线网的管理和优化也成为人们关注的内容[1]。与有线网络不同的是，无线网络的用户具有明显的移动特性，同时无线连接也更容易受到环境的影响。例如，当很多用户节点都集中在某个接入点（AP）附近时，用户所体验到的网络性能就会明显下降。因此，研究无线用户的移动特性，对于无线网络的规划、网络管理等具有直接的指导意义。

同一时间段邻近位置无线网络的用户或终端节点之间往往存在某种关联关系。例如，在同一个教室上课或自习的学生、在同一个会场参加会议的研究者等，他们使用的无线终端往往会关联到邻近的无线接入点 AP 上。用户终端的位置关联属性往往与现实中的人与人之间的社交关系对应，移动终端的携带者是人，人本身的社交属性可以从轨迹信息中得到[2,5,7]。

收稿日期：2013-09-08
基金项目：国家发展改革委 2011 年国家信息安全专项基金资助项目
Foundation Item: 2011 National Information Security Special Project Funded of National Development and Reform Commission

本文将社交网络分析的方法应用到校园无线网用户的移动行为分析上，并以实际的真实运行数据进行验证。本文的主要贡献包括：提出了移动终端接入轨迹的稀疏链接区间（SLI, sparse linked interval）概念，并以此为基础进行移动终端之间的相似度量计算，此方法具有较好的时间复杂度；对比不同数据集的分析结果归纳了 3 种典型的聚类簇的特性和形成原因，以及它们对无线网络管理和规划的联系。

2 相关研究

Balachandran 等人[2]对 ACM 为期 3 天的会议中 195 名参会人员使用无线网的 trace 数据进行分析，研究了无线局域网的用户行为以及网络性能指标，其中，也指出了用户的移动性与使用习惯相关，在公共场所如参加会议情况下所体现出的用户移动特性与校园网的环境下会不同。KotzD 等[3,4]研究了用户的使用习惯，通过量化用户接入 AP 的数量和时长来判断用户是否拥有“驻留位置(home location)”，并籍此得出用户的移动性。基于移动模型(如文献[5~7])的研究主要关注于用户的轨迹变化，注重用户的位置估计和移动轨迹预测，所关注的主要是个体或统计平均的用户移动性。

Hsu W 等人的工作[8~10]与本文较为接近。他们提出以 TRACE 框架来研究无线用户接入和使用无线的模式，用户之间的关系通过关联矩阵（association matrix）来度量。与本文的区别在于“关联矩阵”方法主要关注位置信息，本文则同时考虑了时间和空间维度。例如关联矩阵法并不区分用户终端分别在 A、B 位置停留时间相同但先后顺序不同的情况，在本文中这就是不同的移动行为。同时 Hsu W 等人的工作主要关注用户移动性带来的信息扩散能力，帮助设计可以用在容延网络(DTN)中的网络协议。本文则更关注不同移动模式带来的无线网络管理和规划设计的挑战。

Newman 等人[11]提出了在大规模网络图中分离出类簇团体（community）的方法，Kairam S 等人[12]的方法可以更好地展示大规模网络图的特征。由于所涉及的网络图节点多、边关系复杂，与 Thakur G 等人[13]采用的采样方法不同，本文首先基于阈值提取关键边和关键节点，再对关键节点形成的团簇关系进行扩展。

3 相似性度量

为了理解校园无线用户的行为模式，本节分析了表示用户随时间变化的移动路径方法，并定义了稀疏链接区间（SLI, sparse linked interval），以此为基础计算不同用户终端间的关联程度。

3.1 基于位置的关联关系

接入同一个 AP 的用户终端位置邻近。如果某 2 个或多个终端经常接入在相同 AP 或邻近的 AP 上，它们之间应该会存在某种关联关系。例如 2 个经常一起自习和上课的同学，他们的携带移动终端就可能会体现出这个特征。以图 1 为例，假设图 1 中的 AP 之间有较远的物理距离。其中，在 T1 时刻(C1, C2, C3, C4)终端接入在 AP1 上，（C5, C6, C7）接入在 AP2 上。随着时间的推移，终端可能从一个 AP 移动到另一个。从 T1 到 T2 时刻，观测到 C2 从 AP1 移动到了 AP2，而（C5, C6）从 AP2 移动到了 AP1。于是可以推出(C1, C3, C4)、(C5, C6)分别存在一定的位置关联关系。

3.2 稀疏链接区间

当一个用户终端（唯一的 MAC 地址）移动使用无线网时，按时间顺序记录下 MAC 依次接入的 AP 及停留在每个 AP 的起始和结束时间，以此作为此终端移动轨迹的记录。每个 MAC 地址对应一系列这样的轨迹记录 (ap, startTime, endTime)。

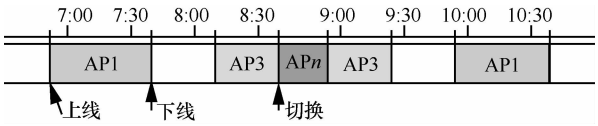


图 2 用户终端 MAC1 的移动向量示例

不同的终端 MAC 地址之间，如果在某一时段接入邻近或相同的 AP，称它们之间存在时间重合。

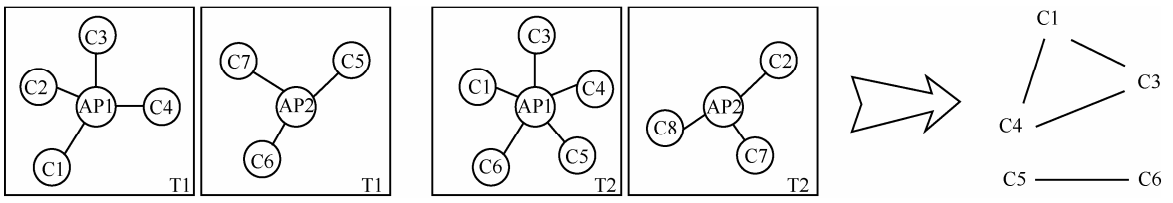


图 1 基于位置的关联关系

重合越多则相互的关联度越高。直接从原始的移动轨迹记录来获得时间重合值计算的复杂度较高[14]。作者设计实现了一种表示 MAC 接入轨迹的方法，它可以高效地计算出不同 MAC 之间的时间重合度。

如图 3 所示，对每个 MAC 地址的轨迹记录，重新按照其接入 AP 进行分类，然后把每个 AP 的接入时间区段按先后顺序排列，注意区段中的起始时间用负数表示。这种表示方法很容易用各种编程语言提供的散列结构和数组来实现。MAC 和 AP 接入时间区段的链接关系决定了哪些 MAC 地址之间需要计算重合关系。由于几乎所有的用户都只会接入到很小比例的部分 AP 上[14,17]，这种链接关系是很稀疏的。因此本文给它命名为稀疏链接区间。

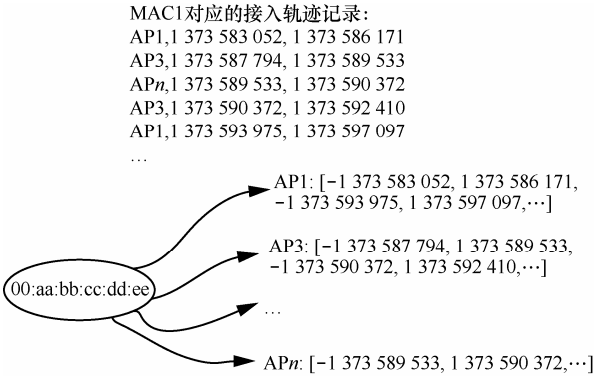


图 3　稀疏链接区间

采用稀疏链接区间来表示轨迹记录，可以很有效地计算某些值。例如区间元素之和即为 MAC 在该 AP 的在线总时长。使用稀疏链接区间表示方法后，计算某个地址 MAC1 与其他 MAC 的时间区间重合长度的方法如下。

1) 根据 MAC1 的接入 AP，获得曾经接入过此 AP 的 MAC 列表（AP 到 MAC 的列表可以事先计算好），只对此列表中出现过的 MAC 计算重合度。由于每个 MAC 接入的 AP 数很稀疏，此步骤相比所有 MAC 之间两两进行循环计算大大减少了计算量。

2) 对每个 MACx 取得其在对应 AP 上的时间区间，将此区间与 MAC1 对应的区间合并到一起，并按绝对值进行排序。举例来说，假设区间分别为[−1, 3, −5, 8, −9, 10]和[−2, 4, −5,7]，合并后排序的区间为[−1,−2,3,4,−5,−5,7,8,−9,10]。则重合的时间区间为连续的 2 个负数后面的区间，如例子中为[−2, 3, −5,7]。计算重合的时长只需对区间的元素求和即可，如例子中的长度为 3。

稀疏链接区间方法的计算复杂度为 $O(nkm\cdot\log m)$，其中，n 为节点总数，k 为节点平均接入的 AP 数量。对比直接计算的方法[14]，其计算复杂度为 $O(n^2m^2)$，其中，n 为总节点数，m 为节点的平均记录数量。

3.3　相似度计算

本文以重合时长占该 MAC 的在线总时长的比例作为 MAC 之间相似度的依据。2 个 MAC 相似度定义为

$$sim_{i,j}=\frac{\sum_{ap} len(overlap(I(i),I(j)))}{\sum_{ap} len(I(i))}$$

其中，$I(i)$,$I(j)$分别为 2 个 MAC 在不同 AP 上的接入时间区间。$overlap()$为计算重合区间，$len()$为计算区间的长度。注意不同的 MAC 总在线时长一般不同，此度量值是非对称的。为了更好地展示 MAC 地址间的关系，本文只对阈值超过预先设定值的才认为存在关联关系。

3.4　聚集关系度量

本文采用社交网络分析（SNA）的方法来理解无线终端间关联的关系。每个 MAC 地址对应社交网络图中的一个节点，存在关联关系的节点间具有边相连。由于相似度是非对称的，此图为有向图。

表 1 中列出了各个与分析节点聚集性质相关的指标[16]。除此之外，本文还使用了节点聚集系数(CC，cluster coefficient)指标。聚集系数的定义为

$$CC_i=\frac{\left|\{e_{jk}\}\right|}{k_i(k_i-1)}, v_j,v_k\in N_i; e_{jk}\in E$$

其中，N_i 为节点 i 的所有相邻节点。此系数计算节点 i 的相邻节点之间的连接数与它们所有可能存在连接的数量的比值，值越大说明它的相邻节点间聚集成团的可能性越大。全局聚集系数为所有节点聚集系数的平均。

表 1　　各中心度量指标信息

指标名称	涵义	用途
频度中心 (degree centrality)	连接数最多的节点	识别内部相连资源数
间接中心 (betweenness centrality)	链接不相通的群组	减少不相通群组之间的活动
亲近中心 (closeness centrality)	对于所有其他节点都很接近的节点	迅速获取或发布信息

4 数据分析

本文的主要数据来源是北京大学校园无线网的真实运行数据。作为对比验证，本文也选取了 USC 提供的校园网 trace 数据[18,19]进行分析。

北大的校园无线网采用的架构是瘦 AP 集中控制，目前在线使用的控制器 13 个，AP 接入点 1 407 个。无线覆盖了主要的教学区、办公区、图书馆以及部分学生宿舍。

4.1 数据处理

与 USC 采集 Syslog 日志的方式[19]不同，本文使用定时轮询方法，每 10 min 取一次。使用 SNMP 协议从无线控制器获取信息，与本文分析有关的主要是终端的接入信息。每次采集时检查终端 MAC 所接入的 AP 是否变化，接入 AP 更改或没有记录时产生一条新的轨迹记录（首次和末次发现时间为当前时刻），否则更新原记录的末次发现时间。同时此系统还会采集终端的 MAC 地址、IP 地址、RSSI、流量等信息，以及 AP 的配置信息等，保存到 MySql 数据库中。在本文后期解释聚集结果时，这些历史数据可以提供详细的回溯信息。

与 USC 数据的另一个不同之处在于本文的终端接入信息是准确到 AP 的，而 USC 的只提供区域级的信息。由于这个原因，处理 USC 的数据时可能一些节点间原本不存在关联关系，如一幢大楼的一层和五层的用户之间，也被算成了关联。因此 USC 数据集中得到的强关联节点数和边数都远大于北京大学的结果。

为了更好地获得有代表性的移动行为分析，本文对原始数据还进行了处理，主要针对以下 2 种情况：1) 在线总时长少于 1 天或出现天数少于 2 天的节点。这部分节点不是典型的校园网用户；2) 持续在线的节点。产生这个现象的原因可能是此终端是一个常驻节点，没有移动行为。此现象的另一个原因是 AP 本身的缓存故障，终端实际已经离线但在控制器中仍然可见。晚上 24:00~凌晨 6:00 仍然在线的 MAC 大都属于此类别，可以排除。

经过前面去噪处理后的数据，采用第 3 节中的方法计算出节点间的关联关系，依照预设的阈值导出强关联的边信息。本文取阈值为：在线总时长不小于 24 h，关联度不小于 50%。有了边信息后，导入到社交网络图分析软件 Gephi[20]中进行分析。

4.2 聚集模式

分析结果显示，不同的数据集产生的结果细节有差异，但整体来看具有类似的聚集模式。图 4 是北京大学 PKU2 数据集产生的社交网络图。以它为例可以看到，全局上形成了多个形式上分离的子图。子图大小各异，小的子图只有 2~4 个节点，大的子图有数十甚至上百个节点。注意这里只考虑了超过阈值的强关联，所以这些形式上分离的子图之间也可能存在弱一些的关联。USC 的数据集由于按区域进行粗粒度聚合，结果整体上产生的分离子图较少，但全局上仍表现为多个明显聚集的多节点子图，所形成的子图包含的节点也较多。

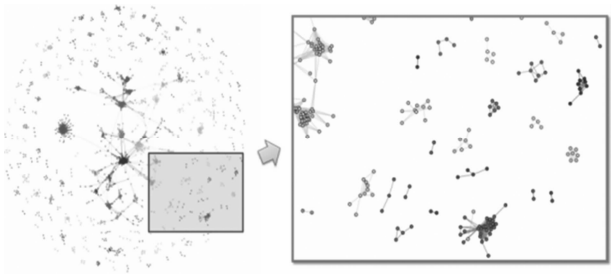


图 4 北京大学 PKU2 数据集的社交网络

对节点信息及历史记录的反查，可以推出它们之间产生关联背后真实的原因。例如，对只有 2~4 个节点的子图，通过分析节点 MAC 地址的 OUI 信息、流量记录、帐号登陆信息等，可以发掘出这种关联的原因：是一个用户拥有的多个终端，还是具有紧密关系的朋友们所携带的终端。相比较而言，节点较多的、大一些的子图的关系要复杂一些。表 3 列出了一些常见的多节点聚集模式。

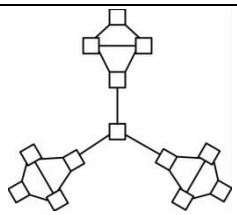
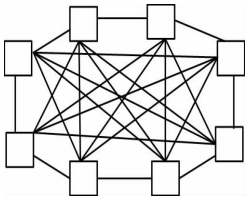
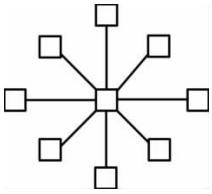
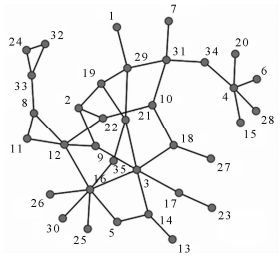
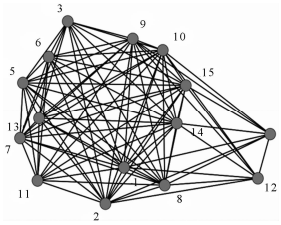
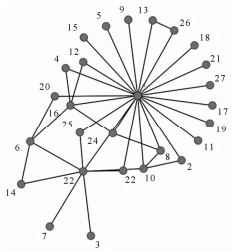
5 结束语

校园无线网的移动终端接入相同 AP 可以看作

表 2 各数据集的统计信息数据集

数据集	AP 数量	区域数	MAC 总数	强关联节点数	强关联边数	聚集系数	采集时间
北京大学 PKU1	1 388	—	52 297	4 060	83 350	0.225	2012.6.1~2012.6.30
北京大学 PKU2	1 134	—	37 776	2 072	29 601	0.218	2013.7.15~2013.7.31
南加州大学 USC[19]	—	137	25 381	20 484	362 368	0.260	2006.1.25~2006.4.28

表 3　　多节点的聚集模式

类别	均衡分散型	关系密集型	中心散射型
子图形状			
代表性的子图			
特征	节点数很多，节点的平均度数不大，3 种不同的中心度量指标差异较大。节点的聚集指数较小	节点数比较多，节点平均度数大，两两节点间几乎都互连，3 种中心度量指标值几乎相同。节点的聚集指数较大	节点数比较多，节点的度数不均衡，有明显的频度中心节点。中心节点度数最大，平均聚集指数较小
聚集产生的原因	用户节点有明显的移动行为，节点通常出现在多个位置，没有驻留位置或各个节点驻留位置不同。没有固定的出现区域，通常出现在教学区	用户节点倾向于较少的移动，选择了共同的驻留位置。终端用户通常是同一个实验室的或在学生宿舍区邻近的位置	中心节点与其他节点行为模式有差别，中心节点在线时间长，并且有驻留位置。而其他节点移动性较大。常出现在图书馆和大的自习室
对无线网络管理和规划的启示	用户之间关联度相对较弱，位于间接中心的节点往往对多个区域的网络都有体验，无线规划时可以考虑参考它的意见	用户关联度高，建议网管系统做节点故障诊断时把相关节点也列出。规划时应有针对性地对此类用户的需求进行优化	频度中心用户是代表性用户，排查故障和做规划应了解此节点的信息和此终端用户的需求

是物理位置和时间上的重合，这种重合越多，终端之间的关联关系越强。本文通过引入移动终端接入轨迹的稀疏链接区间（SLI）表示方法，降低了计算这种重合所引起的节点相似度的计算复杂度，并使用社交网络分析的方法来分析校园无线网用户的移动行为。在结合真实运行数据的基础上，本文整理了常见的聚集子图模式的特征和产生原因，并指出子图模式在无线网络管理和规划中可能的应用。

本文采用社交网络图来分析校园无线网络的用户移动模式，是校园无线网络管理的一个新的发展方向。目前，此种方法的局限性在于需要人去观察形成的图谱，对子图谱的理解也需要借助其他数据和经验。未来可以考虑自动提取出一些关键的关联信息和子图。另一个问题是目前只有接入相同 AP 的节点才认为是相关联的。由于无线信号本身往往有重叠，相邻位置的用户可能接入到 2 个不同的 AP。使用区域粗分的方法对聚集结果有影响，更好的办法是采集 AP 的无线信号邻居关系，把接入相邻 AP 的终端也关联起来。利用社交网络分析的方法进一步去挖掘用户的使用模式和移动模式，并把分析结果推广到无线网络优化、信息推送、协议设计等领域，也是未来的发展方向。

参考文献：

[1] MATHEWS J B. Why are wireless services important to state and education leaders:southern regional education board[EB/OL]. www.sreb.org.

[2] BALACHANDRAN A, VOELKER G M, VENKATRANGAN P, *et al*. Characterizing user behavior and network performance in a public wireless LAN[A]. Proc of ACMIGMETRICS'02[C]. 2002.195-205.

[3] KOTZ D, ESSIEN K. Analysis of a campus-wide wireless network: wireless networks11[A]. ACM MobiCom 2002[C]. 2002.115-133.

[4] KIM M, KOTZ D, KIM S. Extracting a mobility model from real user traces[A]. INFOCOM[C]. 2006.1-13.

[5] YU Z, XING X. Enable smart location-based services by mining user trajectories[J]. Communications of the CCF (in Chinese),2010, 6(6): 23-29.

[6] GHOSH J, BEAL M J, NGO H Q, *et al*. On Profiling Mobility and Predicting Locations of Campus-Wide Wireless Network Users[R]. State University of New York at Buffalo, 2005.

[7] LIN M, HSU W J. Mining GPS data for mobility patterns: a survey[EB/OL].http://www.sciencredirect.com/science/article/pii/s1574119213000825.

[8] HSU W, DUTTA D, HELMY A. Structural analysis of user association patterns in university campus wireless LANs[J]. IEEE Transactions on Mobile Computing (TMC), 2012,11(11):1734-1748.

[9] HSU W, DUTTA D, HELMY A. Mining behavioral groups in large wireless LANs[A]. Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking-MobiCom'07[C].

2007. 338-341.

[10] HSU W, HELMY A. On nodal encounter patterns in wireless LAN traces[J]. IEEE Transactions on Mobile Computing,2010,9(11):1563-1577.

[11] NEWMAN M, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2):26113.

[12] KAIRAM S, MACLEAN D, SAVVA M, *et al*. GraphPrism[A]. Proceedings of the International Working Conference on Advanced Visual Interfaces-AVI 12[C]. 2012. 498.

[13] THAKUR G S, HELMY A, HSU W J. Similarity analysis and modeling in mobile societies: the missing link[A]. Proceedings of the 5th ACM Workshop on Challenged networks-CHANTS 10[C]. 2010. 13.

[14] MOON S, HELMY A. Understanding periodicity and regularity of nodal encounters in mobile networks: a spectral analysis[A] Global Telecommunications Conference (GLOBECOM 2010)[C]. 2010. 1-5.

[15] JAIN A, REDDY B. Node centrality in wireless sensor networks: Importance, applications and advances[A]. Advance Computing Conference (IACC)[C]. 2013.127-131.

[16] HENDERSON T, KOTZ D, ABYZOV I. The changing usage of a mature campus-wide wireless network[J]. Computer Networks, 2008, 52(14):2690-2712.

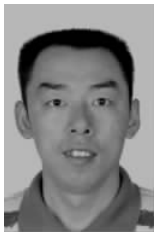
[17] PAPADOPOULI M, SHEN H. Characterizing the duration and association patterns of wireless access in a campus[A]. Wireless Conference 2005-Next Generation Wireless and Mobile Communications and Services(European Wireless)[C].2005.1-7.

[18] KOTZ D, HENDERSON T. CRAWDAD: a community resource for archiving wireless data at dartmouth[J]. IEEE Pervasive Computing, 2005, 4(4):12-14.

[19] Community resource for archiving wireless data at dartmouth[EB/OL]. http://crawdad.cs.dartmouth.edu/data.php.

[20] MATHIEU B, HEYMANN S, JACOMY M. Gephi: an open source software for exploring and manipulating networks[A]. ICWSM[C]. 2009.

作者简介：



周昌令（1977-），男，重庆人，北京大学工程师，主要研究方向为网络与信息安全、无线网络、网络流量分析及网络管理等。

钱群（1985-），男，辽宁营口人，北京大学工程师，主要研究方向为无线网络分析、网络管理等。

赵伊秋（1987-），女，北京人，北京大学硕士生，主要研究方向为无线网、数据挖掘等。

尚群（1972-），男，北京人，北京大学高级工程师，主要研究方向为无线网、网络管理、数据库等。