2017 年第四届中国可视化与可视分析大会

数据可视分析挑战赛-挑战 1

（ChinaVis Data Challenge 2017 - mini challenge 1）

答　卷

参赛队名称：　东北师范大学-陈斌-挑战 1

团队成员：　陈斌，东北师范大学，82426409@qq.com，队长

栾剑，东北师范大学，2509394085@qq.com

徐劭斌，长春工业大学，2233935216@qq.com

汤雅兰，东北师范大学，1498926641@qq.com

张慧杰，东北师范大学，zhanghj167@nenu.edu.cn，指导老师

是否学生队（是或否）：是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）：D3，Excel，MySQL，Java，PBSViz（东北师范大学可视化小组开发的数据可视分析工具）

共计耗费时间（人天）：60 人天

本次比赛结束后，我们是否可以在网络上公布该答卷与视频（是或否）：是

挑战 1.1：伪基站常流动于人口密集的区域，以各种名义向一定范围内的手机发送垃圾短信，因此，了解掌握伪基站出行的时空模式，能够帮助执法人员尽早阻止和抓获不法分子，从而更好地维护社会秩序。然而仅仅从垃圾短信中很难确定其对应的伪基站，即无法确定来自同一台伪基站设备的垃圾短信，相同的垃圾短信有可能来自不同的伪基站，同一个伪基站可能不送不同的短信。请从宏观时空分析的角度出发，对垃圾短信数据进行可视分析，揭示伪基站的总体时空活动规律。

首先，我们将原始数据进行数据清理，并对错误数据和冗余数据进行删除处理。然后，利用 PBSViz 可视分析系统自带的数据初探工具进行分析。数据初探工具（如图 1）主要由查询控制栏和地图视图组成，通过约束数据查询条件筛选出符合要求的垃圾短信，最后，通过热力图呈现伪基站空间分布。我们通过主题模型将垃圾短信的内容分为四种类型，分别是发票骚扰、电信诈骗、生活服务骚扰和其他垃圾短信骚扰。可以使用联合查询方式得到每一种事件类型的伪基站随时间变化的空间分布(通过联合查询，能够得到每一种短信类型的伪基站随时间变化的空间分布情况)。通过热力效果，我们得知：伪基站主要分布在城市的主干道，并具有流动性强的特点。犯罪嫌疑人可能将“伪基站”设备放置在汽车内，驾车缓慢行驶或停靠在特定区域。另外，伪基站越来越灵活小巧，“背包客”成为安放“伪基站”的另一个选择，行李箱、电瓶车都可以作为伪基站的载体，从而流动性进一步增强。

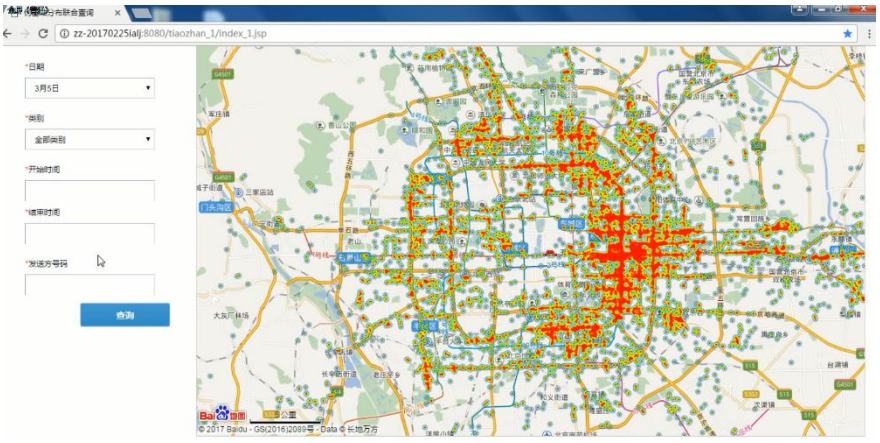


图 1 数据初探工具操作界面

PBSViz 可视分析系统的宏观时空分析界面（如图 2）主要由地图视图、日历矩阵图、趋势图以及气球图组成，其中地图视图展示伪基站影响分布的空间信息，通过控制面板可以呈现热力效果(如图 3a)和马赛克效果（如图 3b），其中每一个马赛克单元代表所在区域垃圾短信的数量，清楚地呈现其空间分布规律。另外，为了得到更美观的可视化效果，可以通过调整地图视图的底图选择用户更喜欢的可视化方案。

我们发现的伪基站空间分布规律如下：

1. 人流量大的地区，伪基站密集。例如在北京国际机场伪基站分布较多；

2. 二环、三环、四环和西大望路成为犯罪分子主要作案场所，尤其是东三环国贸商圈、大望桥附近成为重灾区，每天都有众多的伪基站在这里出现，发送大量的垃圾短信；

3. 国贸商城、北京 SPK、CBD 万达广场、北京银泰中心及华贸购物中心等著名商业地区人流量较大，成为伪基站犯罪的首选地区。

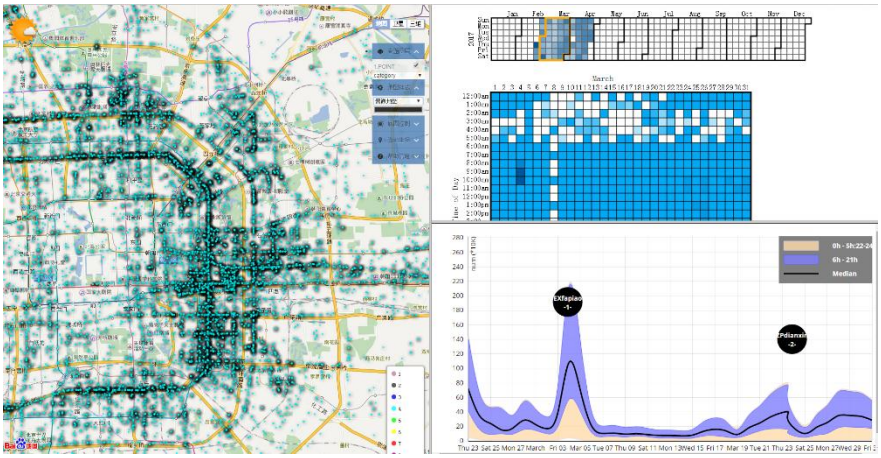
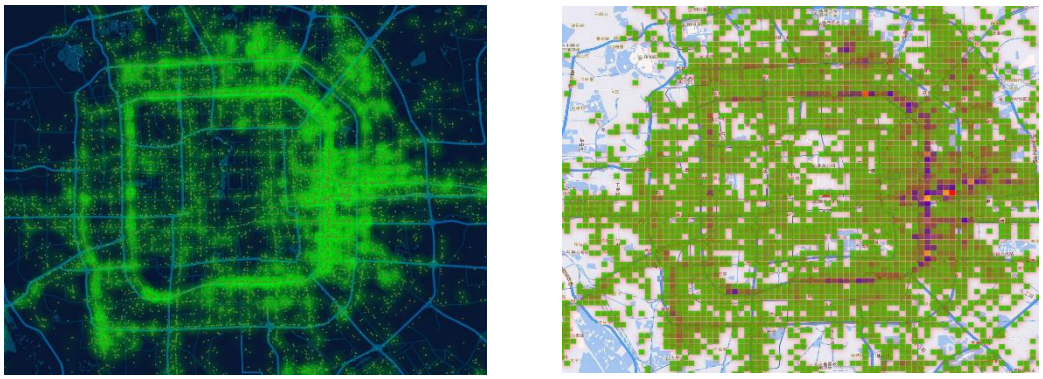


图 2 PBSViz 可视分析系统时空分析界面



（a）热力效果　　（b）马赛克效果

图 3 地图视图的多种可视化效果

日历矩阵视图（如图 4）：该视图是由日历图和矩阵热力图结合。日历图展现整体时间分布规律，该视图横轴为星期的某一天，纵轴为对应月份，每一个矩形单元代表某一天垃圾短信的数量，颜色深浅编码数量级的多少。通过点击对应月份交互联动矩阵热力图，能够细致展示全部垃圾短信的时间分布规律，矩阵热力图的横轴代表一个月的某一天，纵轴被划分为 24h 代表具体的时段。通过日历矩阵视图，我们发现了几个有趣的现象：

1. 伪基站的工作时间主要集中在工作时间段，每天 9h-11h 相对活跃；

2. 2 月 23 号、3 月 4 号、3 月 5 号、3 月 24 号和 4 月 17 号等天，有较多伪基站出现，发送了大量的垃圾短信。

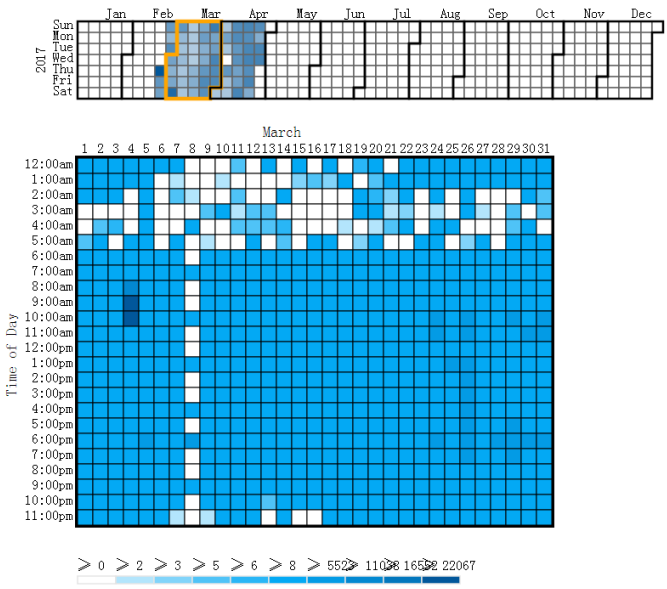


图 4 日历矩阵视图可视化

趋势图(如图5)：该视图主要展现垃圾短信数量随时间的变化趋势。为了更好呈现可视化效果，我们在数据初探后将原始数据每天的各个时段分为热门时段（6h-21h）和非热门时段（0h-6h；22h-24h），分别通过紫罗兰色和藕荷色编码。趋势图的横轴为具体的每一天，纵轴为垃圾短信的数量，条带的面积越大，表示伪基站越活跃，发送的垃圾短信数量越多。其中黑色趋势线为总数量的中值趋势线，中值趋势线落在哪个颜色条带中，表明伪基站在一天中哪个时间段内活跃。通过趋势图我们发现了如下信息：

1. 伪基站在热门时段（6h-21h）较为活跃；

2. 在热门时段，最大值趋势线与中值趋势线构成的区域面积较大，即垃圾短信数量比较多；

3. 对于 2 月 23 号和 3 月 4 号的非热门时段，最小值趋势线与中值趋势线构成的区域面积较大，即垃圾短信数量比较多。由此可以判断这两天存在异常现象。

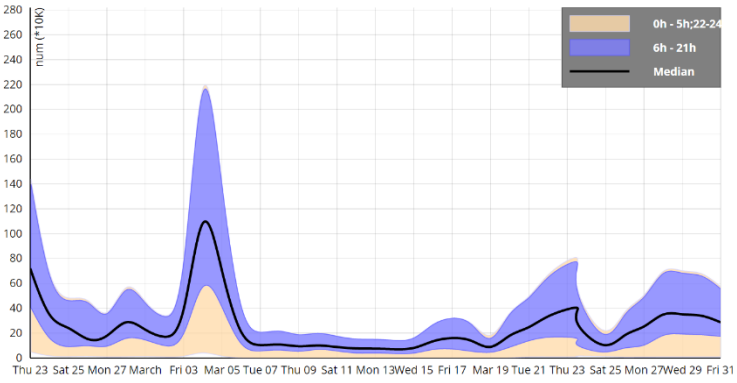


图 5 趋势图可视化

气球图（如图 6）：该视图主要呈现垃圾短信数量随时间段的变化趋势，横轴为每一天，纵轴为时间跨度，垃圾短信数量大小用圆的半径大小编码，半径越大颜色越深则说明该时间段垃圾短信数量越多。这里，我们分别对每一天和每个星期内的垃圾短信进行可视呈现，曲线的弯曲程度表示时间跨度大小。通过气球图我们可以发现：

1. 2 月 23 号、3 月 4 号、3 月 5 号、3 月 24 号等天，气球比较大且颜色较深，说明这些天的垃圾短息比较多；

2. 在 3 月 12-19 号和 3 月 19-26 号这两周内，伪基站发送了大量的垃圾短信，人们受到垃圾短息的严重骚扰。

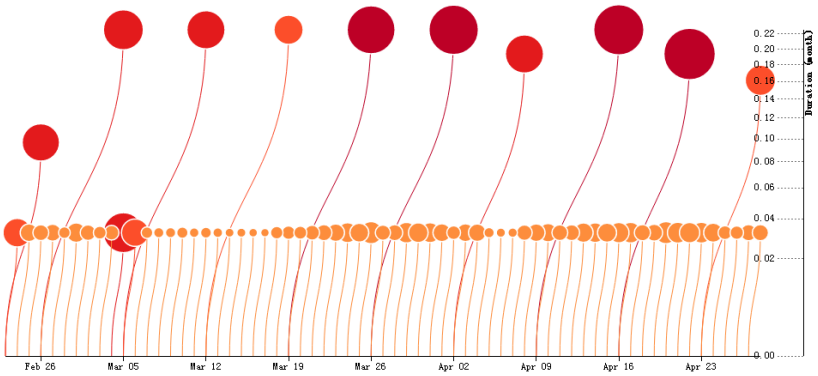


图 6 气球图可视化

挑战 1. 2：不法分子通过设置伪基站设备能够发送不同类型的垃圾短信，请尝试对垃圾短信的具体内容进行分类，分类标准不限，例如：按垃圾短信类型可以分为广告、诈骗等等，按垃圾短信对人们的人生经济危害程度可以分为一般、严重等等。请尝试在问题 1 的基础上进一步分析伪基站发送不同类型垃圾短信的时空分布规律。

我们通过 LDA 主题模型将垃圾短信按照其内容分为四种类型，分别是发票骚扰、电信诈骗、生活服务骚扰和其他垃圾短信骚扰。

堆叠鱼刺图（如图 7）：该视图主要呈现四种垃圾短信类型的数量变化趋势，包含展现每个垃圾短信类型总数量的鱼刺图和呈现每天垃圾短信总数量的堆叠图。其中该视图横轴代表某一天，纵轴为垃圾短信的数量，不同的颜色编码不同垃圾短信类型，其中，“A”为发票短信骚扰（开发票等信息），“B”为电信诈骗（银行卡等信息），“C”为生活服务骚扰（黄色短信等信息）和“D”其他垃圾短息骚扰。并通过堆叠效果（如图 8）清楚地展示每种类型垃圾短信的数量占全天垃圾短信总数量的比例。通过堆叠鱼刺图可视化效果，我们发现：

1. 每天都会有大量的开发票骚扰短信，骚扰人们正常生活；

2. 3 月 4 号全天垃圾短信数量比较多，且各种类型垃圾短信的数量都高于平时；

3. 3 月 27 号全天垃圾短信数量比较少。

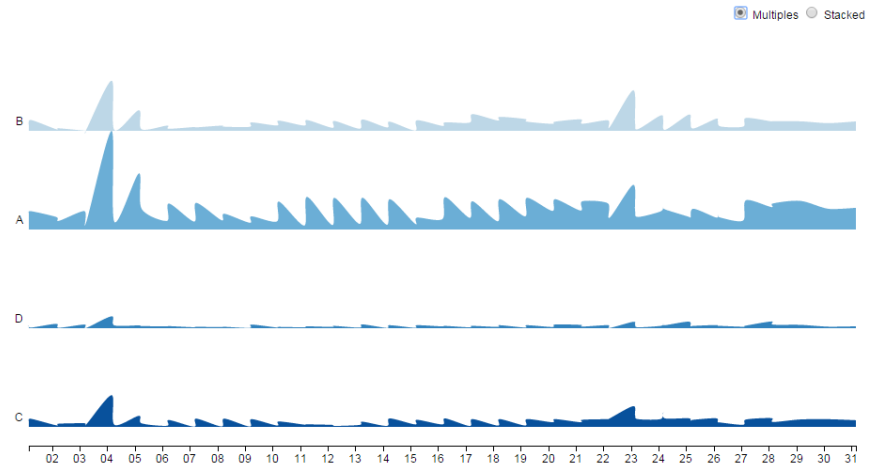


图 7 鱼刺图可视化

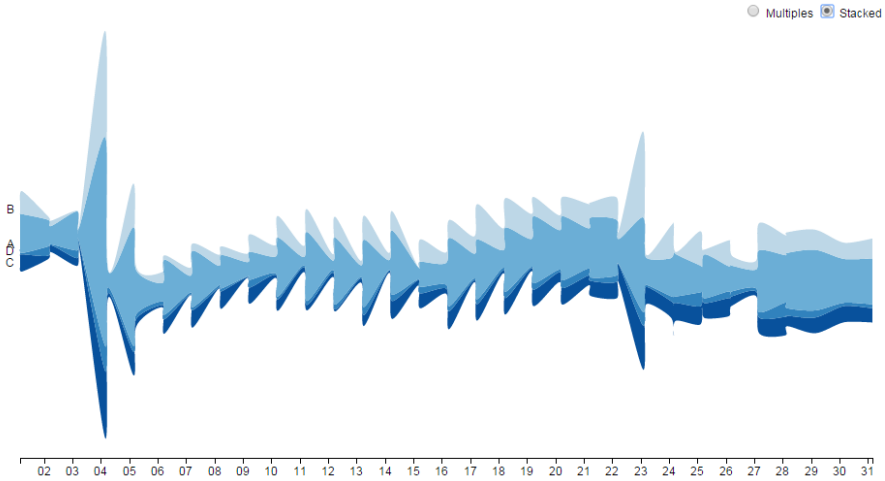


图 8 堆叠鱼刺图可视化

径向弹壳图：该视图主要呈现发票类短信占总数的比例。其中，日期按照径向分布，每一个弹壳的大小代表当天垃圾短信的总数量，圆环分为两个部分，一部分代表发票垃圾短信类型，另一部分代表其他垃圾短信类型。弹壳在哪个圆环的区域多，则表明所属短信的数量多，反之数量较少。颜色编码由蓝到深红代表短信与开发票垃圾短信类型的相关性，蓝色为正相关，深红色为负相关。从图中可以看出，每天存在大量的发票垃圾短信，3 月 6 号该类垃圾短信数量比较少。

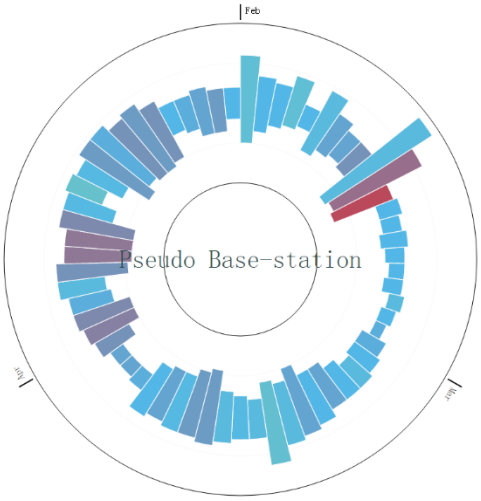


图 9 径向弹壳图可视化效果

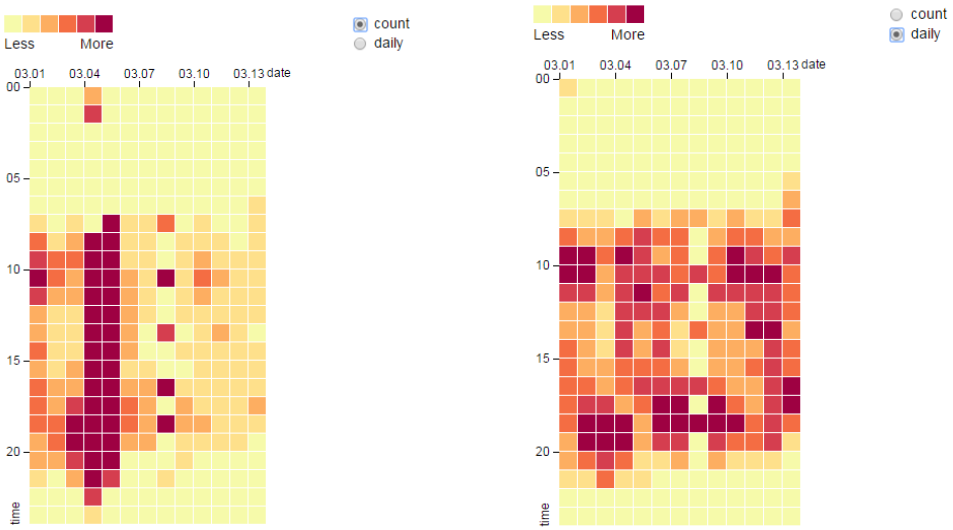
矩阵图（如图 9）：我们利用矩阵图，来发现各种类型垃圾短信数量随时间的变化情况。该视图横轴为每一天，纵轴为具体的时间段，每一个矩形单元代表该类垃圾短信数量的多少。另外，我们还可以利用该视图呈现发送此类型垃圾短信的频繁时间段。进一步结合该类型垃圾短信的地图视图、堆叠鱼刺图和径向弹壳图可视化的结果，我们发现：

1. 在 9h-20h 时间段内，发送开发票类短信的伪基站比较活跃；

2. 3 月 4-5 号，开发票类型的伪基站发送了大量的垃圾短信；

3. 二环、三环、四环、西大望路由于存在大量电信诈骗类的垃圾短信，成为犯罪分子主要作案场所，尤其是东三环国贸商圈、大望桥附近成为重灾区，每天都有众多的伪基站在这里出现；

4. 国贸商城、北京 SPK、CBD 万达广场和北京银泰中心、华贸购物中心等著名商业地区人流量较大，该区域伪基站产生大量开发票类骚扰短信和生活服务类垃圾短信。



（a）发票骚扰短信的矩阵图可视化　　（b）发票骚扰短信频繁的时间段

图 10 发票骚扰类型短信的矩阵图可视化效果

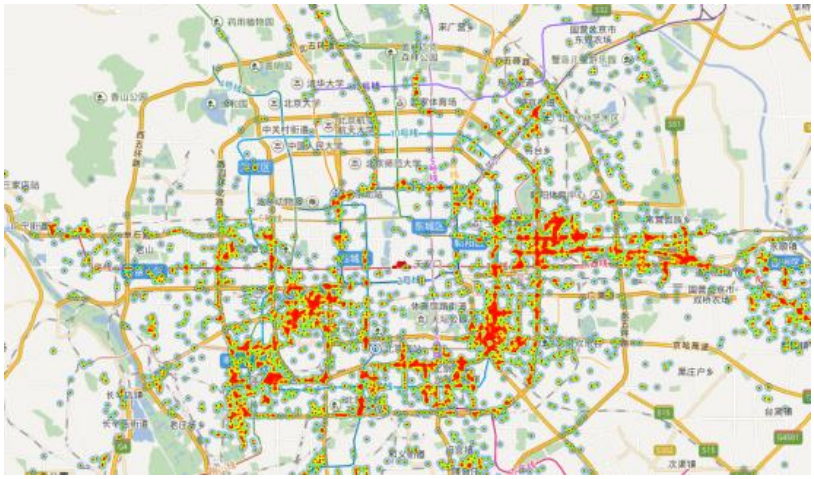


图 11 发票骚扰类型短信的地图热力可视化效果

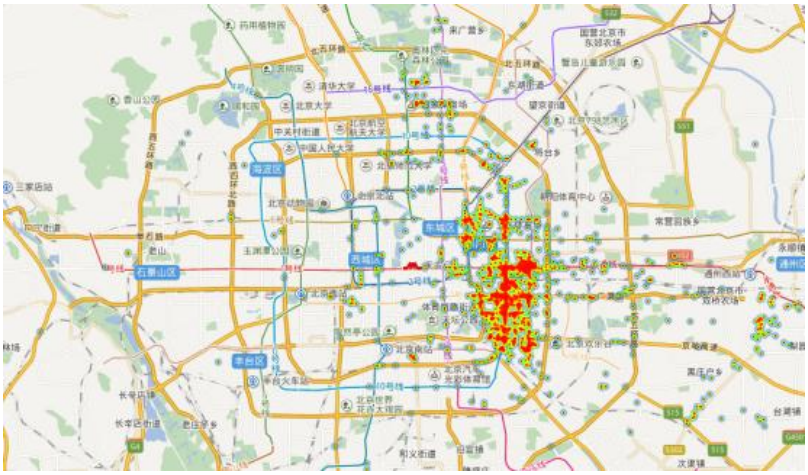


图 12 电信诈骗类短信的地图热力可视化效果

挑战 1.3：伪基站不仅破坏正常电信秩序，危害公共安全，扰乱市场秩序，而且严重损害群众财产权益，侵犯公民个人隐私，社会危害严重。据《人民网》统计，每年通过“伪基站”设备发送诈骗、赌博、推销、中奖等短信近千亿条，伪基站已成为社会一大公害。请结合以上两题中得到的伪基站行为模式，向执法人员提出打击整治伪基站的有效建议和方案，并结合数据分析结果进行说明。

伪基站对人们日常生活造成严重影响，危害公共安全，因此我们必须利用新方式、新手段实施“伪基站”打击治理工作。通过 PBSViz 可视分析系统，发现伪基站发送短信的空间分布主要集中在人流量大的地区（如图 13），并分布在城市的主干道，这也体现了伪基站流动性强的特点。由于伪基站众多，打击“伪基站”经常面临人手紧张的情况，可以尝试将监测点安放在北京市公交、私家车等移动载体上，同时在各大主干道，国贸商城、北京 SPK、CBD 万达广场、北京银泰中心、华贸购物中心等热门地区设立专门的固定监测点。另外，利用公安部门专业的“伪基站有害信息监测平台”，对手机用户“伪基站”信息接收情况进行实时监测，通过系统提供的实时地址定位信息，结合“伪基站侦测仪”手段，在路口进行拦截侦测，从而大大提升拦截捕获成功率，实现精准打击。

图 13 伪基站发送垃圾短信的空间分布

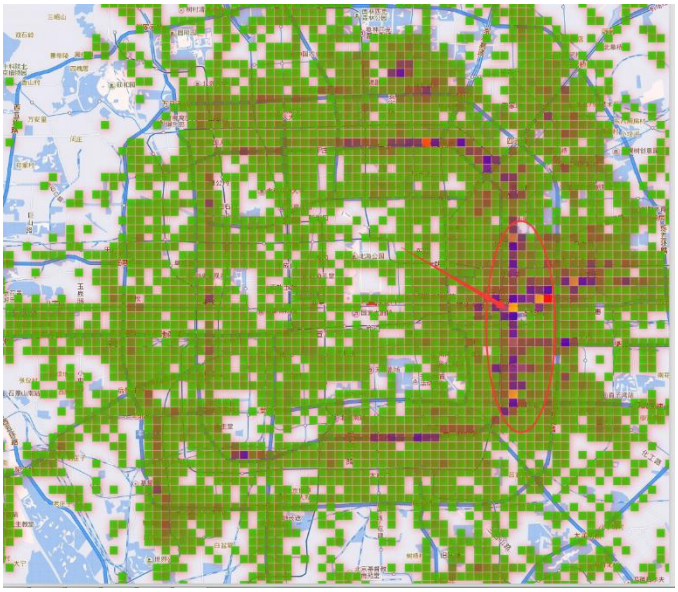


图 14 伪基站发送垃圾短信的马赛克热力效果

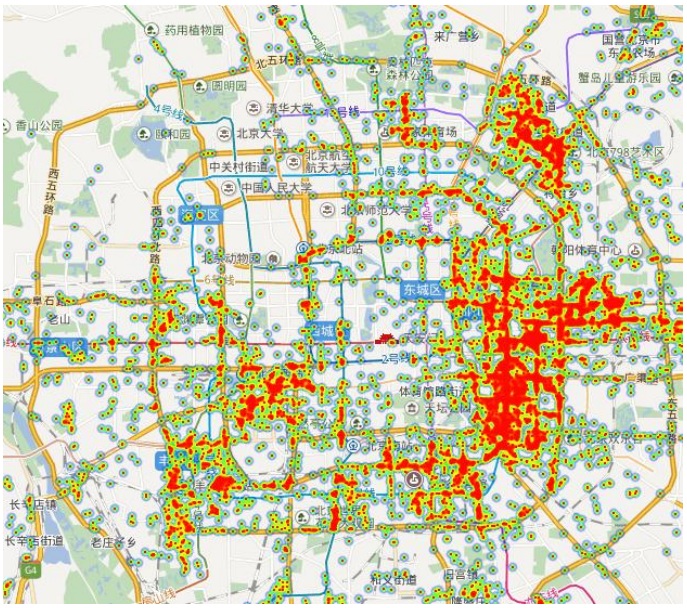


图 14 伪基站发送垃圾短信的热力图效果

另外，伪基站工作的活跃时间段一般集中于工作时间段 6h-21h（如图 14），通过在这个时间范围内，令运营商、无线电管理、通信、质检等多部门协同补强移动网络，让伪基站的覆盖尽可能不会强于合法移动网络。

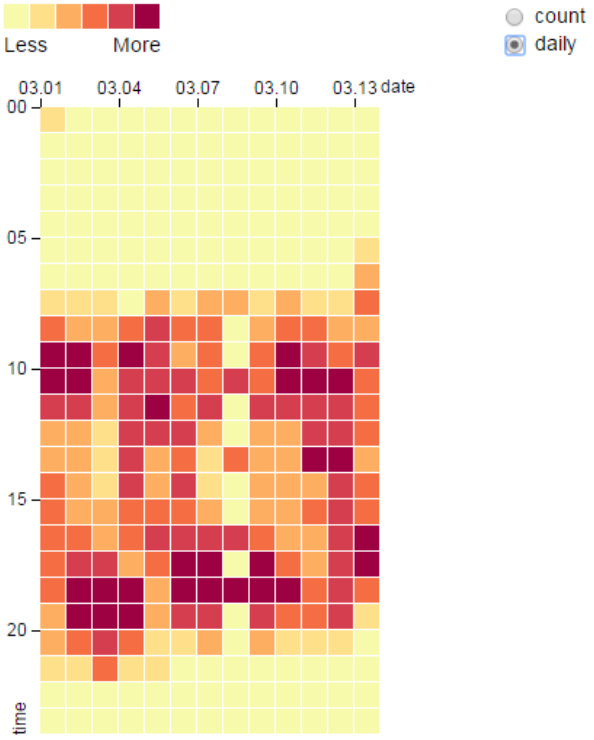


图 14 伪基站发送垃圾短信的时间分布

从伪基站的工作原理可以看出，伪基站只能获得手机的 IMSI，得不到手机的号码。也就是只知道身份证号码，不知道姓名。因此，我们可以要求在短消息中附加一个额外内容，也就是被叫号码。利用被叫号

码，手机可以校验是否是发给自己的，对不符合的短信一律主动删除，这样垃圾短信就没有现形的机会了。这种方法牵涉短信中心的改造，并且需要在手机的短消息功能上增加一个判断，如果是智能手机，实施起来应该非常方便。

公安、运营商、无线电管理、通信、质检、工商等相关部门都参与打击伪基站，才能形成高效的联动机制，对伪基站背后利益链的各环节进行打击。从源头入手，从产业链入手，进一步加大打击整治“伪基站”的工作力度，工商和质检等部门，同样应切实加强对“伪基站”设备生产、流通环节的严格监管。同时，继续做好互联网上相关违法信息的监测和清理整治工作，可以进一步压缩“伪基站”的生存空间。