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：　安全橙子-吕博-挑战 1

团队成员：　吕博，中国电子科学研究院创新中心，blyu@csdslab.net，队长

贾月，中国电子科学研究院创新中心，yjia@csdslab.net

吴兆胜，中国科学技术大学，mega_victor@163.com

是否学生队（是或否）：　否

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）：Processing，Echarts

共计耗费时间（人天）：　40 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

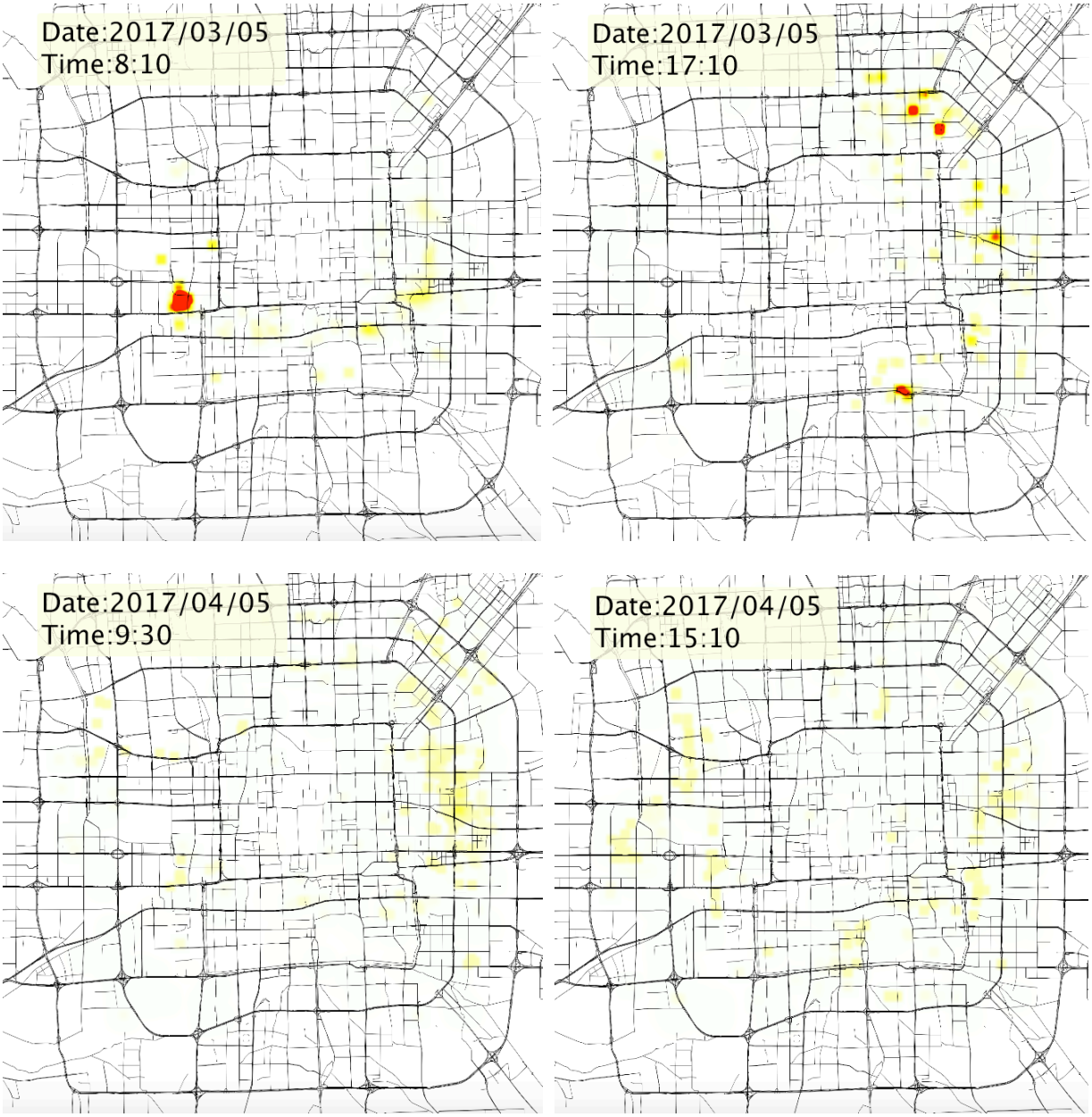
（灰色字为参赛信息填写模板，请参赛者在提交时参照模板填写）

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。（请将回答尽量控制在 1500 字和 8 张图片内）

（下面是答题区域）

（1）伪基站在北京分布的热力图

为了能够宏观展示出北京伪基站分布的时空规律，使用 Processing 绘制伪基站时变的热力图。例如下图分别展示了 3 月 5 日与 4 月 5 日上下午某时刻的热力图。



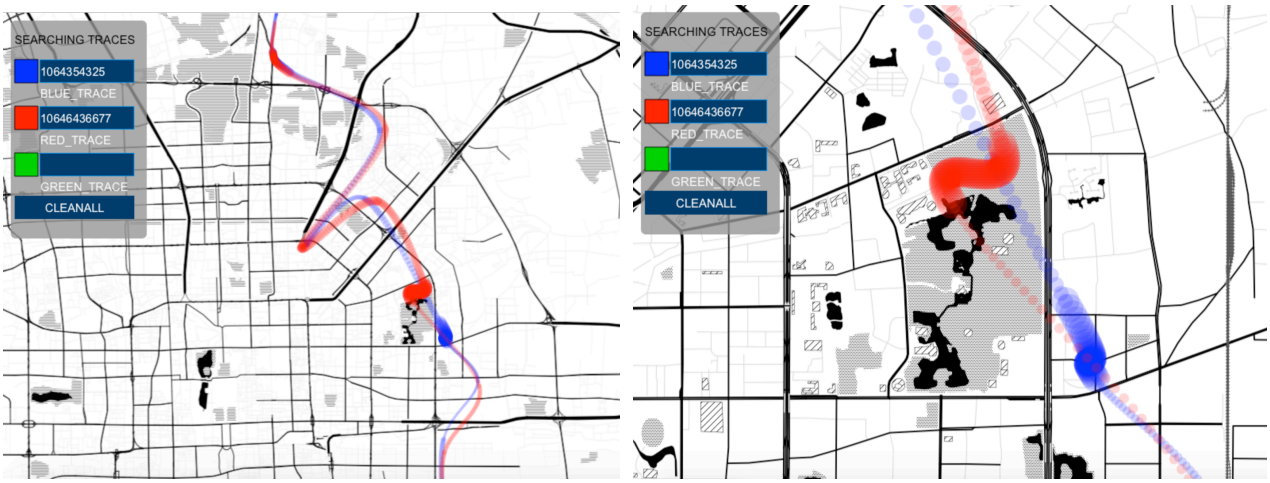
从图中可以宏观的看出伪基站分布总体的时空趋势，例如 3 月 5 日伪基站发送短信数据比较多，而且比较集中在少数几个热点区域（例如写字楼、商圈等），反之 4 月 5 日为清明长假，由于市民大多休息在家或出城旅游，伪基站发送短信数明显减少，而且分布更加均匀。从热力图中的规律可以总结出：伪基站在选择发送地点和发送时间时会非常谨慎和有针对性，尤其是和市民的日常行为有非常大的关联。热力图能够很直观的看出伪基站的分布，但是对于伪基站的轨迹模式还需要更细粒度分析手段进行分析。

（2）移动伪基站的移动轨迹模式分析

在原始数据中，伪基站伪装的发送方电话号码是识别一条时空轨迹的标识，但是有两个问题使我们不能用其作为识别伪基站轨迹的标志。问题一是有多个伪基站可能伪装成了同一个号码（例如 95580、10086 等），问题二是同一个伪基站可能同时使用了多个伪装号码。

为了对伪基站的轨迹进行可视化，首先将每一条用电话号标记的时空轨迹按照空间上的欧氏距离进行聚类，鉴别同一时段同一个号码背后的多条轨迹，从而解决问题一。然后利用轨迹的时空相似度将多条时空规律都极为接近的轨迹聚合为同一个伪基站，从而解决问题二。

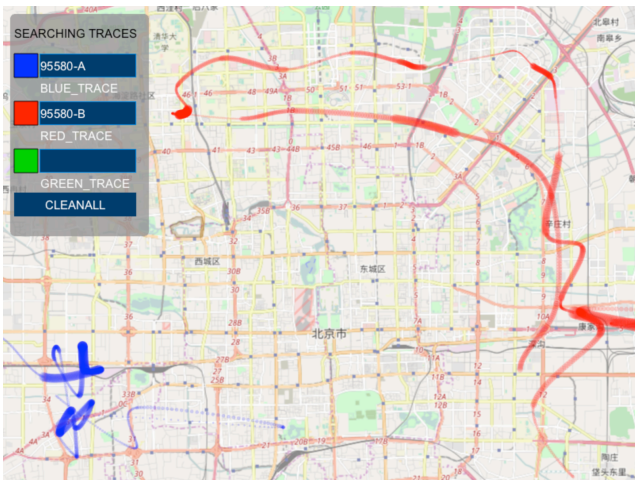
通过以上时空数据分析方法，可以有效的提取出伪基站的行动轨迹，在此基础上，本团队利用 Processing 搭建了时空轨迹查询与可视化系统，该系统支持缩小放大、轨迹检索、轨迹可视化等基本功能，同时支持地图风格选取、鼠标取经纬度等辅助功能。使用者可以在交互界面中输入手机号对轨迹进行检索，并指定轨迹的可视化颜色。在可视化轨迹中，轨迹的粗细与饱和度代表了该伪基站发送短信的数量和频率。

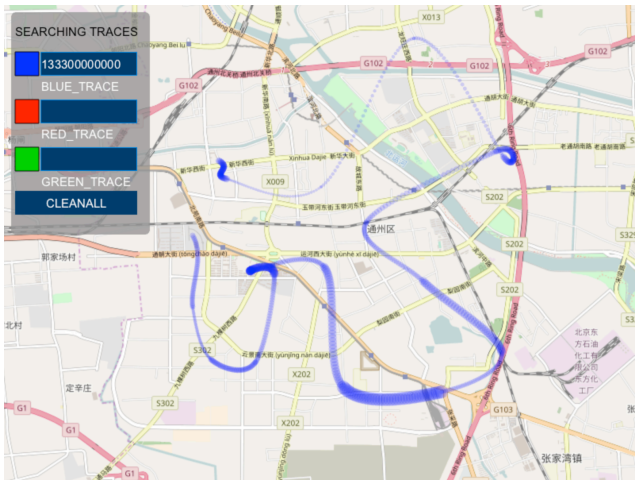


本团队通过分析发现，如上左图所示，很多轨迹的时空相似度很高，可以从图中直观的看出是同一个伪基站运营者使用了不同的手机号造成的，将地图放大后可以发现（如上右图所示），该伪基站运营者分别在公园的北门与南门的停车布设了伪基站，利用两个手机号集中发送一段时间之后，迅速的离开了该地。

在数据分析过程中团队发现，很多伪基站伪造相同的号码发送短信，尤其是冒充身份类的短信喜欢用 10086、95580、95560 等服务类号码，利用时空聚类的手段可以将同一时间内距离较远的轨迹区分开，如下左图所示，区分了朝阳区和海淀区的两条 95580 的轨迹。

不同伪基站隐藏行踪的手段也有所差异，通过计算某条轨迹的上坐标点的分布规律可以对轨迹进行分类，例如有的轨迹习惯围绕某一地进行活动（如下右图所示）甚至静止不动，并集中的在某几个具体的时间段（例如上下班高峰）发送短信，其余时间则隐藏行踪。而另外一些伪基站则时刻保持移动，通常沿着四环或者五环沿线发送短信（如下左图中的红色轨迹）。





通过对轨迹进行细粒度的分析可以发现伪基站移动的空间轨迹，对于伪基站时间模式和发送内容规律的分析，将在挑战 1.2 的解答中进行叙述。

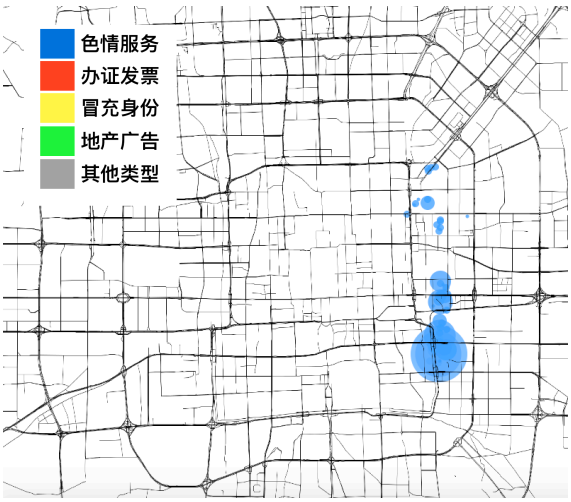
挑战 1.2：不法分子通过设置伪基站设备能够发送不同类型的垃圾短信，请尝试对垃圾短信的具体内容进行分类，分类标准不限，例如：按垃圾短信类型可以分为广告、诈骗等等，按垃圾短信对人们的人生经济危害程度可以分为一般、严重等等。请尝试在问题 1 的基础上进一步分析伪基站发送不同类型垃圾短信的时空分布规律。（建议参赛者回答此题文字不多于 1500 字，图片不多于 8 张）（请将回答尽量控制在 1500 个字和 8 张图片内）

（下面是答题区域）

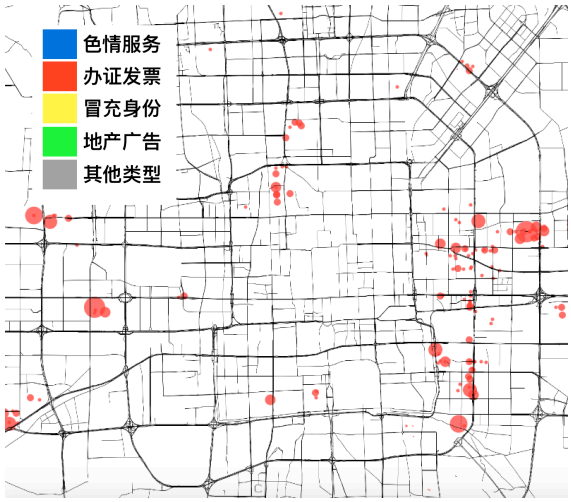
（1）内容的整体分布

为了分析伪基站发送短信内容的时空规律，使用分词库对短信内容进行分词和语义分析，提取短信内容的关键词，根据关键词的距离对短信内容进行分类聚类，参考专家对垃圾短信的分类方法，将短信分为办证发票、冒充身份、色情服务、地产广告、违禁推销、招聘广告、非法博彩、打折促销、教育移民和其他类型等十大类，并对每一类短信的危害程度进行量化评估。

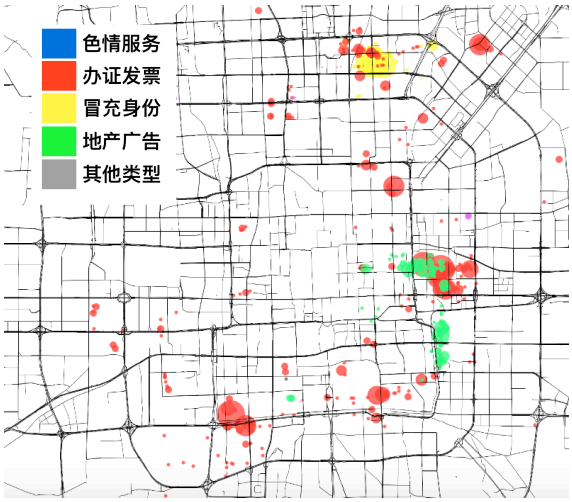
本文使用 Processing 将各类短信内容的宏观态势进行直观的展现，散点图的颜色代表短信类型，散点的大小代表短信数量。以 2 月 23 日为例，将一天分为四个时段，每个时段各取一个有代表性的时间点，如下图所示，凌晨色情类短信为伪基站发送的主要短信，而且主要集中在三环内比较繁华的商圈附近。办证发票类短信分布比较均匀，且从上午到晚上都存在。2 月 23 日为工作日，下班高峰同样是伪基站发送短信的高峰期，地产广告和办证发票类短信集中在金融商圈附近发送，发送时间短暂但发送数量很大。到了晚上，潜伏了一整日的色情服务类短信死灰复燃。



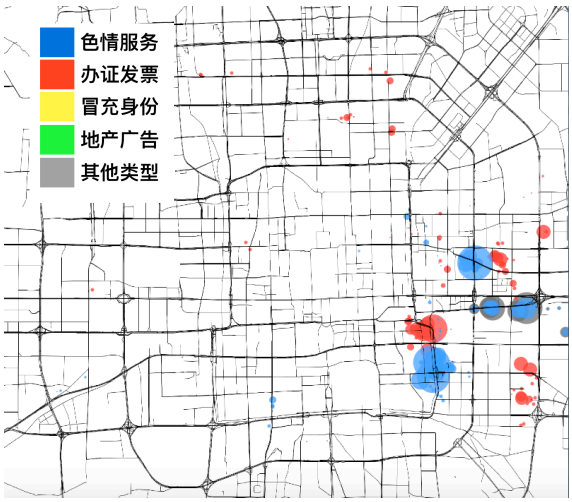
2 月 23 日 0:00—6:00 短信分布



2 月 23 日 6:00—12:00 短信分布



2 月 23 日 12:00—18:00 短信分布

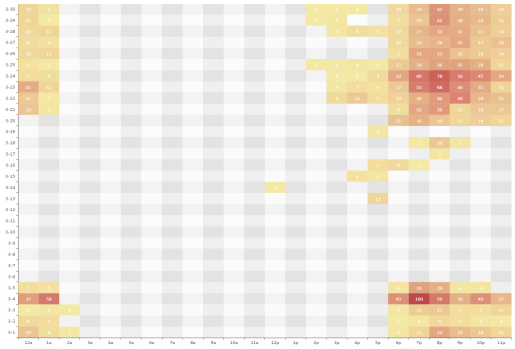


2 月 23 日 18:00—24:00 短信分布

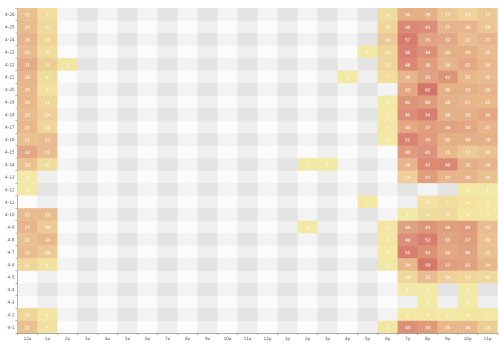
（2）伪基站发送短信的时间规律

为寻找伪基站短信发送的时间规律，本团队通过统计各类型短信每天不同时段的数量，构建了以下热力图，横坐标为一天的 24 小时，纵坐标分别是三月份的 31 天及四月份的 26 天，由于二月份数据量较少及篇幅所限，没有对其数据进行展示。主要几类信息的时间分布规律如下：

■ 色情服务类短信



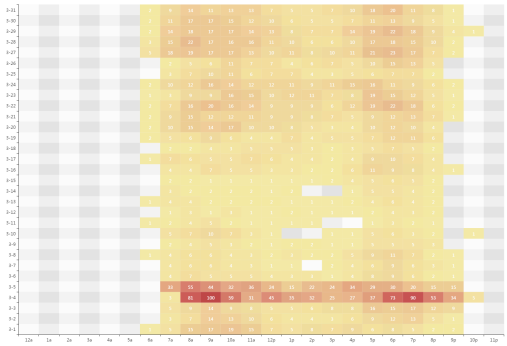
17 年 3 月色情服务类短信分布



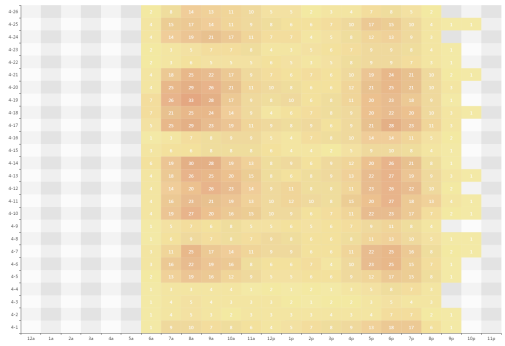
17 年 4 月色情服务类短信分布

色情信息主要活跃于晚上 18 时至凌晨 3 时，4 月 3 号、4 号、5 号是清明小长假，这个时段色情服务信息只有少量活跃且都分布在白天；另外一反常时段是 3 月 6 日至 3 月 20 日，这个时段几乎没有色情服务信息，本团队分析自国务院第 666 号令《国务院关于修改部分行政法规的决定》3 月 1 日公布并施行后，全国文化市场执法部门都加大了执法力度，此时段北京侦破几起“扫黄打非”大案，相应的色情服务类信息也大大减少，但之后的时间里色情类短信有明显的反弹趋势。

■ 办证发票类短信



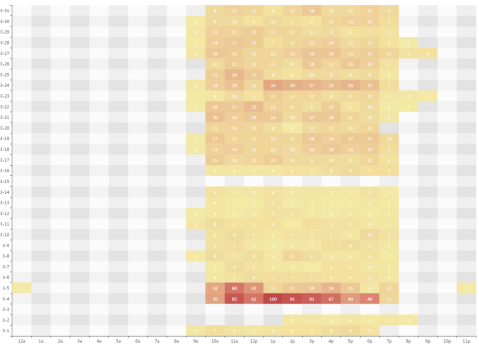
17 年 3 月办证发票类短信分布



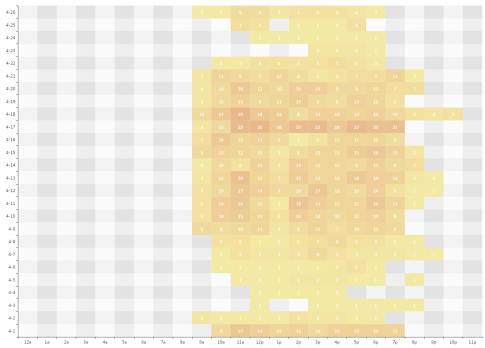
17 年 4 月办证发票类短信分布

通过分析图可以清楚的发现，办证发票类信息主要活跃于 6 时至 21 时，并且有极强的“周期性”—工作日活跃，双休日较收敛；另外三月份大部分日期都不如四月份活跃。3 月 4-5 日办证发票类短信存在异常，本团队分析认为 3 月 4-5 日两天的总体信息量极大，各类型短信数据都同比增长，远超过其他天的数据，对于这样异常的数据量激增，在分析行为规律的时候不予过多考虑。

■ 冒充身份类短信



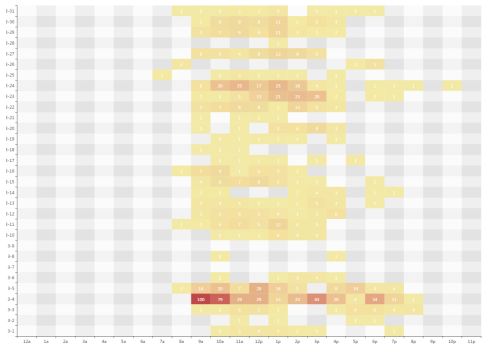
17 年 3 月冒充身份类短信分布



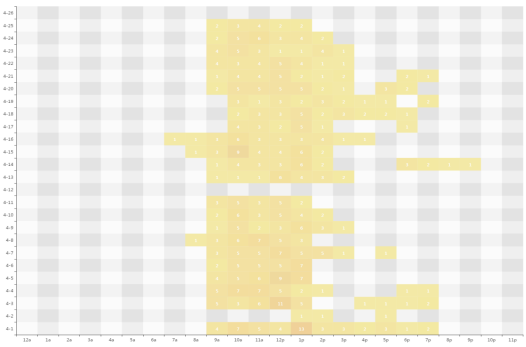
17 年 4 月冒充身份类短信分布

冒充身份类信息主要在白天活跃，并且活跃程度较为平均，比较有趣的是在 3 月 3 日及 3 月 15 日这两天这类短信没有出现，3 月 15 号是消费者权益保护日，3 月 3 号是政协第十二届全国委员会第五次会议，管控较为严格；清明小长假冒充身份类短信活跃也较少，分析缘由同 1.1 中第一小节。

■ 地产广告类短信



17 年 3 月地产广告类短信分布

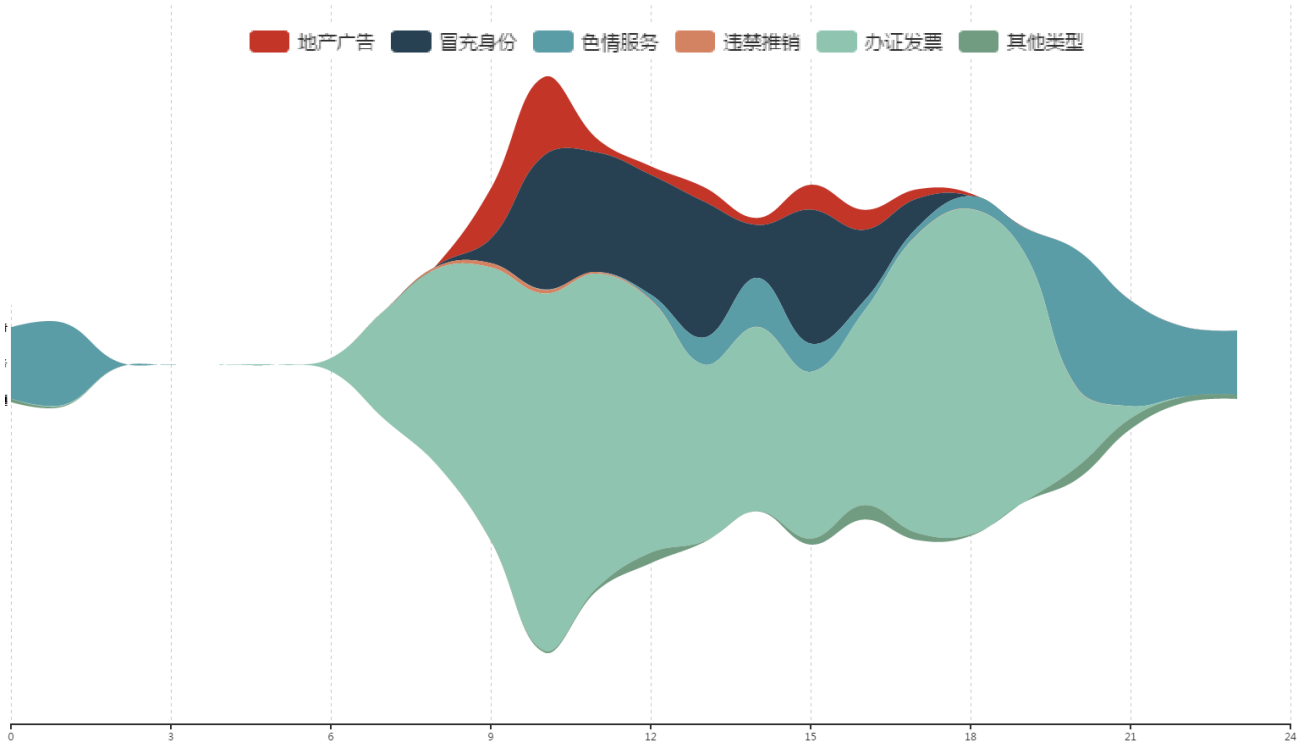


17 年 4 月地产广告类短信分布

通过分析图可以清楚的观察到地产广告主要活跃于 9 时至 13 时，在清明小长假时活跃时段相对较少。

综合以上分析可初步看出各类型短信都有自己的时间分布规律，总体来看在清明小长假、3.15 消费者保护日等节、假日以及重要活动与政策发布的日期，各类短信出现的都较少，有些类型的信息甚至全天都没有出现，应与警务人员的打击力度有很大的关系，然而一旦打击力度下降，各类伪基站短信都有反弹，色情类信息表现尤为明显。

所有短信一天内大致的时间分布及总量规律如下图：

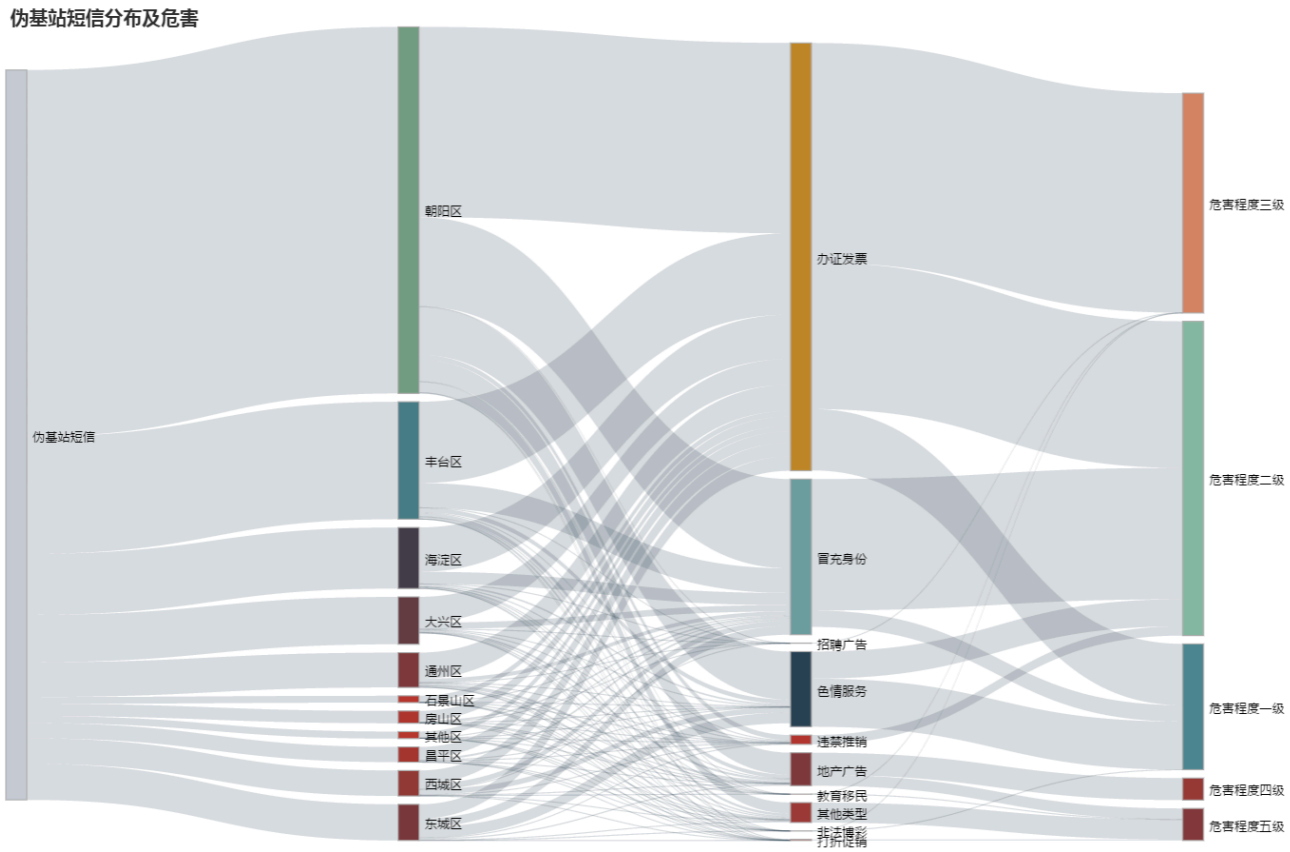


总体上办证发票短信的数量最多，主要分布在白天；其次数量较多的色情类短信主要活跃于夜间及午间；之后冒充身份及地产类广告类短信活跃于上班期间；剩余少量的其他类型短信零星分布于全天。

（3）内容与地点的关系

为更加充分的把握伪基站短信的时空规律与影响度之间的关系，本团队使用百度 LBS 提供的 API 将原数据的经纬度映射到北京市行政区中，根据伪基站短信的分布以北京市行政区划为单位进行统计，并将 10 类短信的危害程度划分为 5 级，“危害程度一级”为最恶劣以此类推直到“影响程度五级”，根据短信内

容、伪装号码、分布时间、分布地域等不同因素，将十类短信有所区别划分到 5 级危害程度体系中，具体内容展示如下：



伪基站短信主要分布于经济发达、写字楼密集的城六区，以朝阳区为首，其次是丰台区、海淀区、东城区西城区与石景山区，但不与人口与土地面积成正相关。从桑基图中可以直观的看出伪基站短信的流向，例如色情服务类短线主要来自朝阳区，但是东城区色情服务类短信所占的比例同样不可忽视，而海淀区色情服务类短信所占比例最少；石景山区没有教育移民类短信。

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。（请将回答尽量限制在 1000 个字和 5 张图片内）

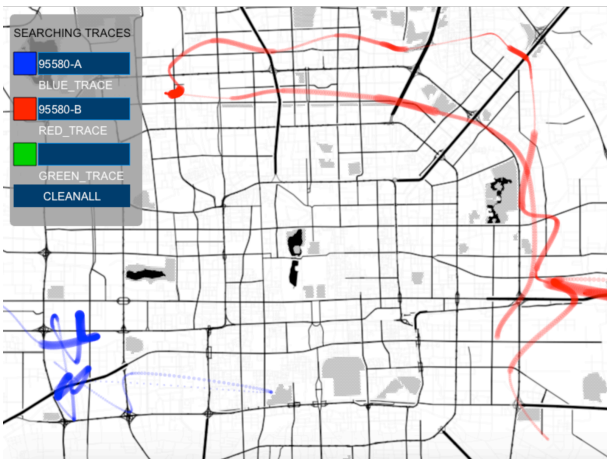
（下面是答题区域）

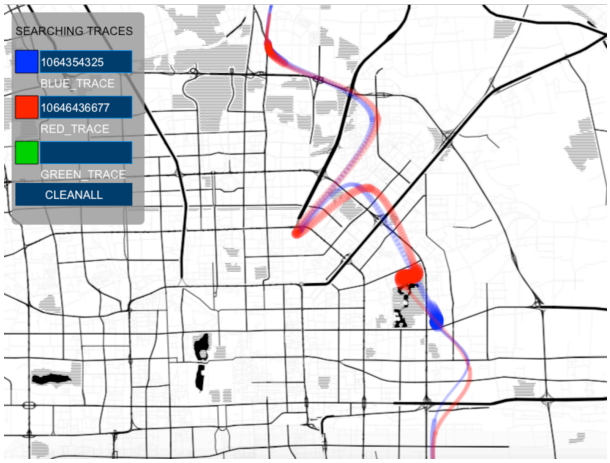
通过对伪基站数据进行可视化分析，本团队对伪基站管理提出以下建议。

（1）伪基站轨迹跟踪与交通摄像头联动实现伪基站溯源

在现实中伪基站通常采用车载的形式，车辆沿着公路行驶，途中定点投放短信。但是在现有的伪基站系统中不包含车牌号信息，这对伪基站溯源和实施抓捕带来了挑战。通过将伪基站数据于交通摄像头数据进行关联分析可以有效解决这个问题。现有的交通摄像头能够准确识别车牌号，通过对系统数据库中车牌号进行检索可以获得某个车牌号的移动轨迹，利用本团队提出的时空轨迹相似度关联的方法，可以准确锁

定伪基站车辆的车牌号。例如左下图中显示的 95580-B 轨迹，车辆从从东四环出发，在北四环正常行驶后停车一段时间，转而从北三环行驶回原地，经过的地点均为摄像头覆盖区域，很容易从车辆监控中建立车牌号与行驶轨迹的关系。



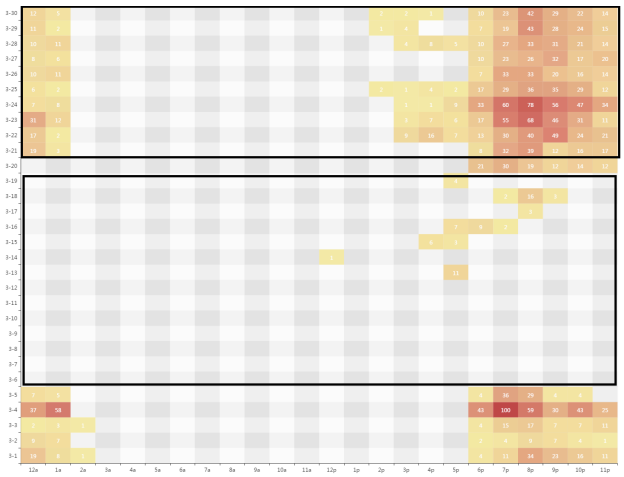


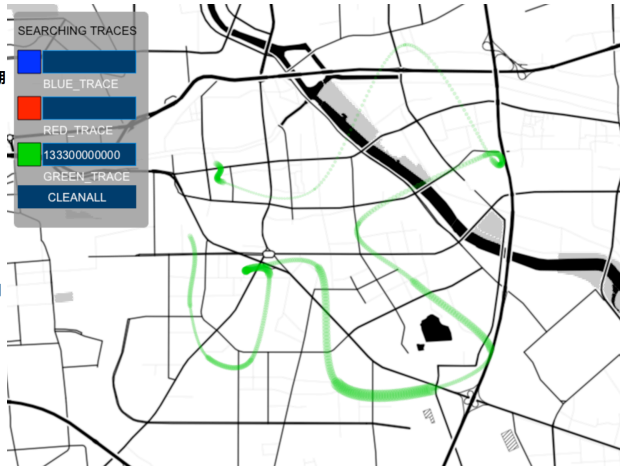
（2）完善伪基站短信上报机制，建立伪基站号码查询系统并向社会开放

本届挑战赛只开放了两个月的伪基站数据，从数据集中的地点来看，上报数据的用户集中在京东地区，数据量和数据覆盖的人群并不多。因此建议完善伪基站短信上报机制，汇聚数据资源，建立伪基站号码查询系统，具备伪基站号码查询、历史活动轨迹可视化、短信内容模糊检索等功能。当平台积累了一定量数据之后，可以建立伪基站上报与检索平台，面对社会公众开放提供服务。本团队在解决挑战问题的同时，开发了伪基站轨迹检索平台原型（如上右图所示）。

（3）基于分析预测伪基站行为模式实现伪基站精准管控

根据前两题的分析，我们可以初步确定出不同时段不同类型伪基站出现概率最高的地点，从而可以作为优先打击的依据。例如分析发现，在一段时间的严打之后，伪基站活跃程度会大幅反弹，根据以往数据总结得出，在大小节假日、重要活动、重要政策推行之后，在反弹期进行定时、定点抓捕，可以提高抓获伪基站的概率。例如，色情类短信在一段时间的严打之后，预计会有反弹期。根据从历史记录分析出的时空规律（如下两图所示），可以为执法人员提供若干定点抓捕区域，同时与道路监控联动，提高抓捕成功率。





（4）揭示伪基站的运营家族

在可视化分析的过程中团队发现，伪基站移动的时间、空间和内容都有很大的重复性，伪基站移动路线、发短信时间和短信内容都有固定的行为模式，这意味着大部分伪基站是由少数运营者掌控的。利用轨迹关联分析与模式分析的手段可以将伪基站聚类为几大运营者，从而揭示伪基站背后的运营家族，为伪基站管控提供依据。