## 天津大学-侯伟婷

利用词云显示短信文本中出现频率高的关键词

1. 非法服务：**代开发票**，虚假办证，赌博，套现服务，**色情服务**
2. 其它：散播谣言，中介服务，不知所云
3. 诈骗：**积分兑换**，虚假中奖，冒充他人，**假冒机构**，借贷放贷
4. 广告推销：伪基站推广，股票，**房地产**，招聘，教育培训，医疗保险，汽车周边，信用卡办理，麻将/扑克牌，商城店铺

词云.jpg

分析垃圾短信具体正文，将垃圾短信分为四大类

使用圆环图展示

展示类别与时间的关系（使用时钟进行隐喻）

分析垃圾短信对人的经济危害程度，内容分类不一定等于经济分类

按不同行政区对着几种类别进行划分

时间-地点-类别

概览-过滤-细节

Overview-Filter-Detail

## 重庆大学-何凌君

对文本数据进行分词和词频统计，再根据关键词，将伪基站发送的短信分为5类：

1. 发票办证
2. 诈骗短信
3. 色情广告
4. 房产买卖
5. 其他

各类短信主要出现时间分布情况

发票办证：6:30-21:30

诈骗短信：9:00-21:00

色情广告：19:00-2:00

房产买卖：9:00-16:00

房产广告垃圾短信午前和午后出现较多，到晚饭前基本结束，与购房者的活动时间重合，按照地点分布，房产广告类主要在CBD区域，住在郊区，每天通勤去CBD上班的人群是潜在的购房者，所以也适合这类垃圾广告投放的目标人群。

发票办证是所有类型的垃圾短信中，分布区域最广，时间最宽泛，数量最多的垃圾短信类型，站到所有垃圾短信数量的一半左右。短信出现最多的时段集中在7:00-11:00、17:00-20:00。结合时空分布可以看出，发票办证类选择的是早晚高峰时段，常常出现在拥堵路段，拦截、发布短信。因为发票办证几乎对各类人群都有一定的需求，为了能够覆盖到更广泛的人群，伪基站的违法者经常在早高峰和晚高峰活动。

色情卖淫广告垃圾短信的特征

其活跃时间多在傍晚到凌晨，根据中国扫黄打非网站的信息，两会期间数量较少。

诈骗短信犯罪性质最为恶劣。从地点分布上来看，主要集中于大兴、望京区等较为富裕的区域。这些区域中生活的人群，手里有余钱，是诈骗的目标人群。

诈骗短信在一个时间点发送的短信内容完全相同，而发票类和色情类的短信在聚集点会出现不同的短信文本。

## 东北大学-吕伟朝

对数据中的短信内容content字段进行分词和词频统计，选取Top100制作成如图所示的垃圾短信词云，关键词越大表示词频越大。

1. 广告推销类：金融服务，实体店，房地产，教育培训
2. 违法信息类：**代开发票**，赌博，色情，办证刻章，谣言，药品
3. 诈骗短信：**打款诈骗**，电商网站，钓鱼诈骗
4. 骚扰信息类

词云：

代开发票：发票，代开，保真

打款诈骗：付款

假冒建行、工行号码，含有木马链接，对人们经济危害较为严重

圆环图.jpg

取一天的垃圾短信进行分析，沿地铁，主干线沿线较多。

横轴：时间

纵轴：数量

柱：四层堆叠图

代开发票、办证印章、打款诈骗多集中在工作时间，

色情一般集中于下午19:00至凌晨1点，

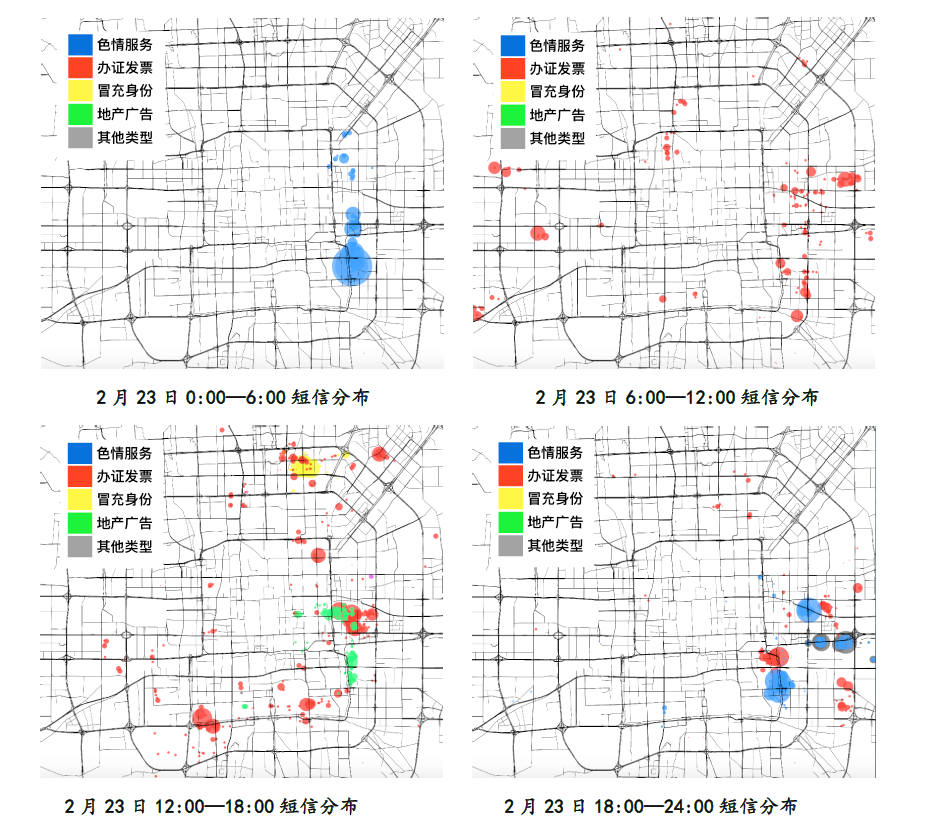
房地产集中于上午：适合看房

电商：早上和下班之后，放送时间适合网上购物

## 安全橙子-吕博

利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

1. 办证发票
2. 冒充身份
3. 色情服务
4. 地产广告
5. 违禁推销
6. 招聘广告
7. 非法博彩
8. 打折促销
9. 教育移民
10. 其他类型



热力图显示各类型短信每天不同时段的数量，

横坐标：一天24小时

纵坐标：三月份的31天和四月份的26天

清明小长假，色情服务信息少量活跃

3月6日-3月20日，两会，扫黄打非，加大执法力度

办证发票：

工作日活跃，双休日收敛

冒充身份：

两会，315，清明，较少

地产类垃圾短信：

9时-13时

办证发票短信的数量最多，主要分布于白天，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。

百度LBS提供的API将原数据的经纬度映射到北京的行政区中。按照危害程度，将其分为5级，危害程度一级到危害程度五级。

伪基站短信----行政区----类别----危害程度

桑基图.jpg，看出短信的流向。

伪基站短信主要分布于经济发达、写字楼密集的城区，以朝阳区为首，其次是丰台区、海淀区、东城区、西城区和石景山区，但不与人口与土地面积成正相关。