数据可视化期末报告

成员: 林恒旭 赵刚 李秉轩

December 30, 2021

1 问题一

1.1 信用卡-所有者推断

根据信用卡消费记录，会员卡消费记录和车辆 gps 定位记录，可以推断信用卡的所有者。我们发现，gps 只记录了部分时刻下车辆的位置信息。车辆在某些时间段内每隔几秒记录了一个 gps 经纬度坐标，而在其它的某些时间段内没有任何 gps 坐标信息。在这些没有 gps 坐标的时间段上，车辆很可能处于停止的状态，这意味着车主可能在附近的商店购物，并使用信用卡和会员卡消费。因此，我们遍历了一辆车所有记录了 gps 坐标的时刻，如果下个时刻和当前时刻的间隔超过一分钟，那么在两个时刻包括的时间段内，车停下了。所有这样车辆停止的时间段被获取保存下来。由于车辆停止意味着的可能消费行为，所以当一张信用卡的所有消费时间恰好都在车辆停止的时间段内，那么该信用卡可能属于该辆车的车主。因此，可以粗略地获得每辆车的车主可能持有的信用卡的集合。考虑到不可能存在两个不同地点的信用卡消费记录匹配到一辆车的同一个停止区间的情况，排除上述情况，可以过滤大部分的车辆-信用卡匹配结果。剩下的匹配结果中，每辆车平均可以匹配到约 2.1 个信用卡结果。最后，通过人工手动筛选，得到最终的信用卡-所有者匹配结果。

1.2 会员卡-所有者推断

根据会员卡消费记录，gps 定位记录和信用卡-所有者匹配结果，可以推断会员卡的所有者。在一辆车的停留时间段内，该车的坐标可以用这个时间段开始的 gps 坐标来代替。而该时间段内的信用卡消费记录中的地点的坐标可以自然地对应到上述坐标，由此得到了所有地点的经纬度坐标。接着，对每张会员卡，筛选其所有消费记录，将地点对应到具体坐标，利用坐标信息和日期信息，匹配所有在该日期停留过该坐标的车辆。最后，通过人工手动筛选，得到最终的会员卡-所有者匹配结果。

Name	CarID	Credit Card ID	Loyalty Card ID
Nils Calixto	1	7889	L5777
Lars Azada	2	1415	L7783
Felix Balas	3	9635	L3191
Ingrid Barranco	4	7688	L4164
Isak Baza	5	6899	L6267
Linnea Bergen	6	7253	L1682
Elsa Orilla	7	9241	L5947
Lucas Alcazar	8	7117 4948 9551	L6119
Gustav Cazar	9	1321	L3014
Ada Campo-Corrente	10	6691 8332	L2070
Axel Calzas	11	1877	L4149
Hideki Cocinaro	12	7108	L6544
Inga Ferro	13	7819	L5259
Lidelse Dedos	14	1874	L4424
Loreto Bodrogi	15	3853	L1485
Isia Vann	16	7354	L9254
Sven Flecha	17	7384	L3800
Birgitta Frente	18	9617	L5553
Vira Frente	19	6895	L3366

Name	CarID	Credit Card	Loyalty Card ID
Stenig Fusil	20	6816	L8148
Hennie Osvaldo	21	9405	L3259
Adra Nubarron	22	1286	L3572
Varja Lagos	23	3484	L2490
Minke Mies	24	4434	L2169
Kanon Herrero	25	214	L9637
Marin Onda	26	1310	L8012
Kare Orilla	27	3492	L7814
Isande Borrasca	28	5921	L3288
Bertrand Ovan	29	2681 3547	L3295 L9362
Felix Resumir	30	6901	L9363
Sten Sanjorge	31	5010	L2459
Orhan Strum	32	8156	L5224
Brand Tempestad	33	9683	L7291
EdvardVann	34	4795	L8566
Willem Vasco-Pais	35	2463	L6886
Albina Hafon	36	9220	L4063
Benito Hawelon	37	3506	L7761
Claudio Hawelon	38	9614	L5924
Henk Mies	39	8642	L2769
Valeria Morlun	40	7792	L5756
Adan Morlun	41	9152	L5485
Cecilia Morluniau	42	9735	L9633
Irene Nant	43	4530	L8477
Irene Scozzese	44	2276	L3317

1.3 可视化设计

1.3.1 可视化描述

图 1 展示了轨迹的可视化模块，右上角为全部40俩车的多选框，勾选特定车辆后，图上只会显示这些车辆的位置以及打印他们在一天中的行动轨迹。时钟是系统全局的虚拟时间，我们的系统建立一个虚拟的时间系统，当用户点击左下角的播放按键时（现在显示为暂停状态），系统会按照每100毫秒自增虚拟时间1秒，自增的同时会更新所有的视图。按照这种方法，我们的各个视图的数据时钟是对齐时间的，并且对于每个动画（比如多辆车运行），我们不是单独启动一个线程运行，而是按照每一帧计算所有车辆位置的方式，这样的效果就是勾选多辆车和勾选一辆车，系统的运行效率相差不大。图上的不同颜色圆圈分别代表不同的地点，这些位置是根据数据集确认的。我们的系统虽然是全动态的，但是因为添加上了暂停按钮，所以注意到任何异常情况就可以点击暂停，然后静态的观察当前的数据（悬浮鼠标，显示tooltip）。

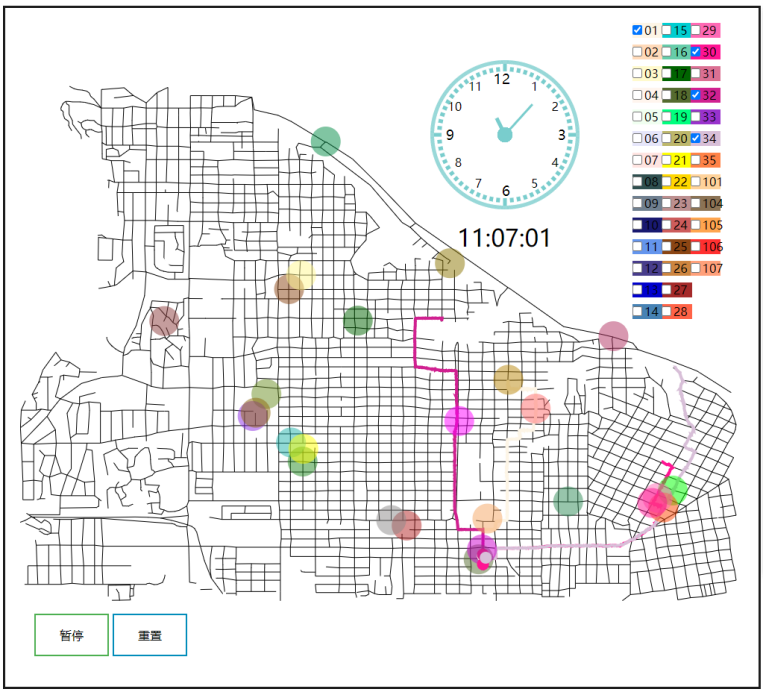

图 1

1.3.2 可视化使用

因为是全动态的过程，我们很容易可以根据路径模拟情况锁定一些异常情况，之后在基于数据细致分析落实。下面举一些分析的例子：

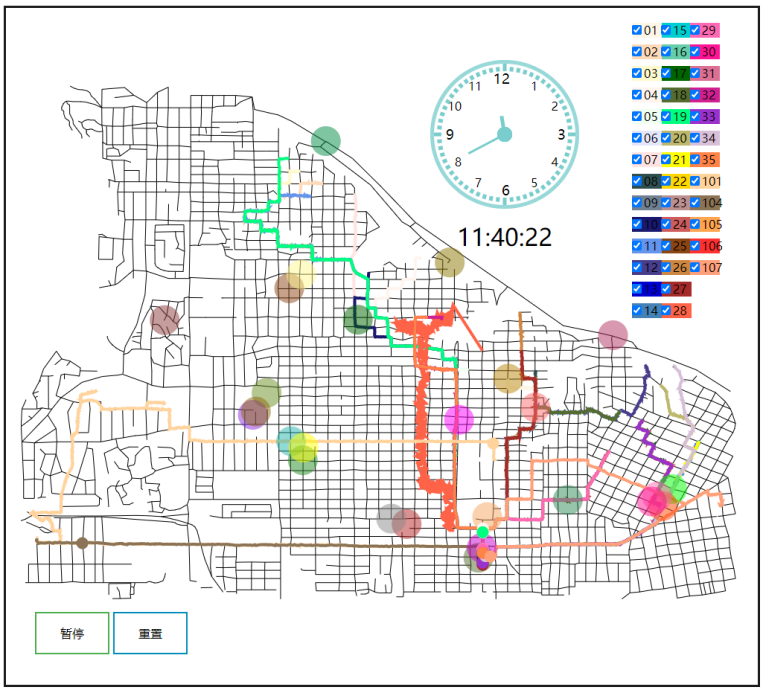

图 2

28号车的异常轨迹信息（图 2 橙色），说明该车司机存在扰乱GPS的可疑行为。

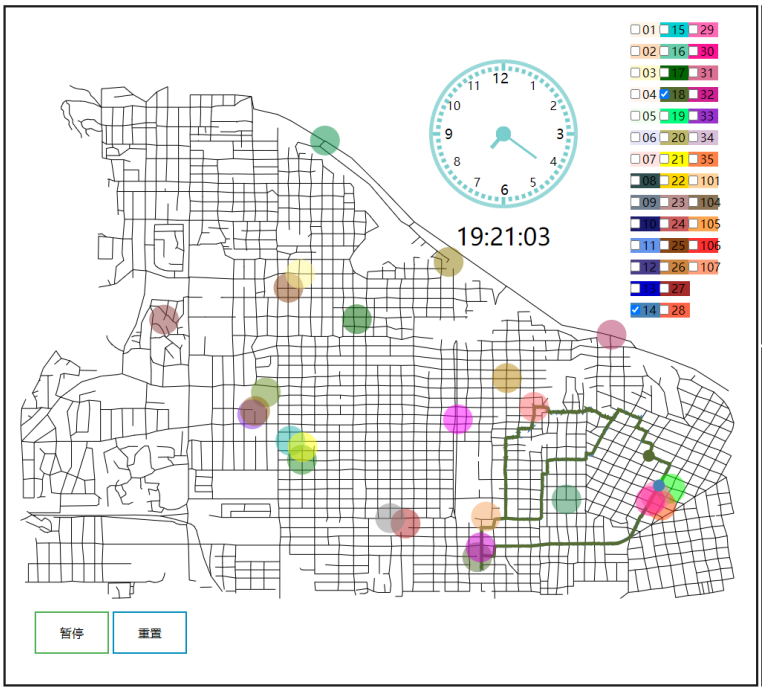

图 3

14号和18号全天近乎相同的运行轨迹，二者存在着非正式的关系。

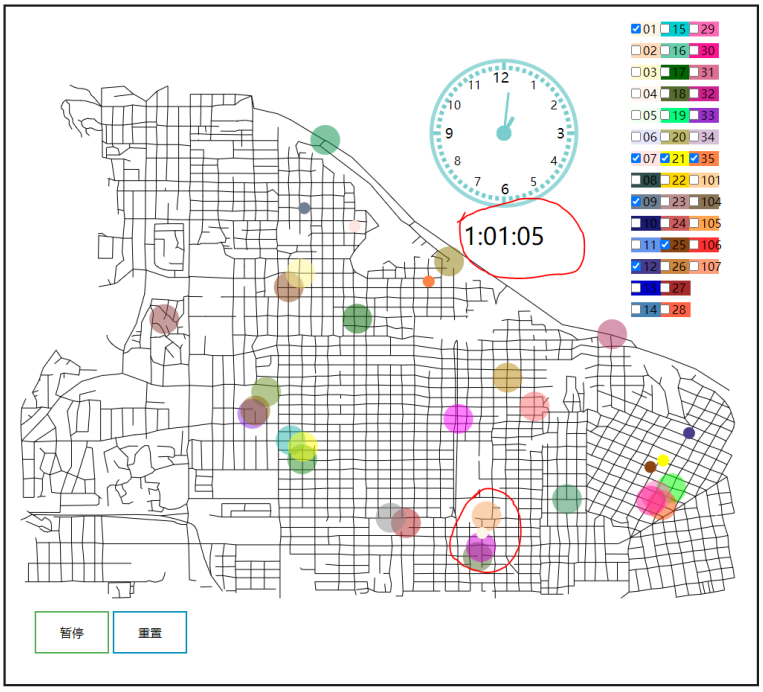

图 4

1号车凌晨时间去公司上班，说明可能拥有特殊身份。

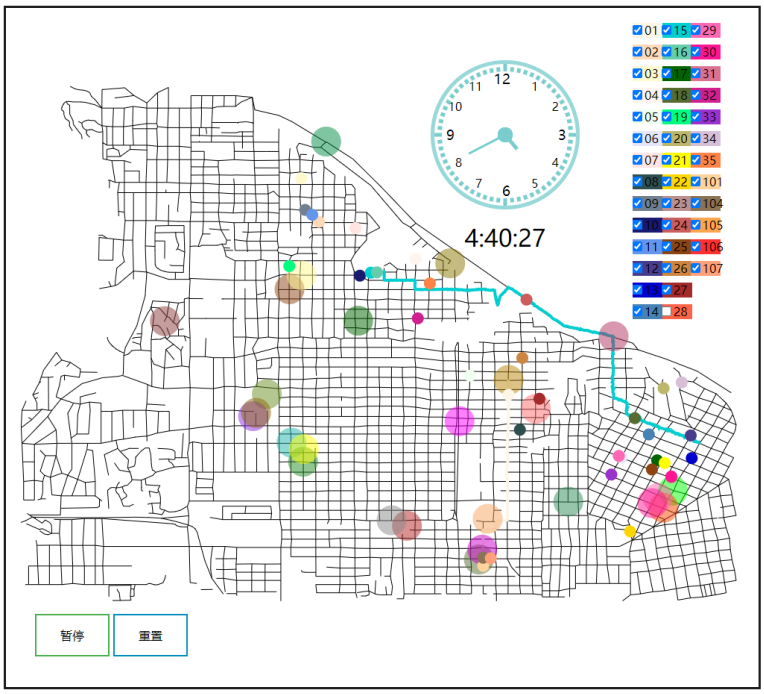

图 5

15号车半夜横跨大半个城市的诡异行踪，可能在从事非法交易。

1.4 热门时间和热门地点确认

热门时间和热门地点主要通过同一时段内，一个区域内停留车的数量来决定，以10分钟为间隔，程序计算每个时刻40辆车停留位置属于的区域，一个区域如果停留的车超过2俩（人为决定这个阈值）。该地点就视为热门地点，对应的时间为热门时间。通过可视化视图容易观察到热门地点和热门时间。

（1）图 6 中标红的两个区域在每天晚上 11 点到次日早上 9:30 成为热门地点，我们推测这是由于这两个地点是公司相关的员工社区，并且该时间为非工作时间。

（2）每天的 9:30 到 18:00，Jack's Magical Beans、Bean There Done That 两个地方成为热门时间，猜测公司就在这附近，是所有人正常办公的地方。

（3）每天正午 12:00 到 13:30，Roberts and Sons、Gelatogalore、Hippokampos、Abila Zacharo、Katerina's Cafe、Frydos Autosupply n' More、Brew've Been Served、Guy's Gyros、General Grocer 等地方人数增多，而公司区域人数降低。我们猜测这可能是中午的午休时间，此时吃饭，休闲等场所的人数增加，成为热门地点。

（4）每天的 18:00 下班以后，Katerina's Cafe、Frydos Autosupply n' More、Brew've Been Served、Guy's Gyros 等地方成为热门地点，可以看出这是大家下班后长时间休闲的场所。

（5）还观察到一个有趣的现象，在周末（也就是 12,19 号两天），Frydos Autosupply n' More、Brew've Been Served 等娱乐场所的热门时间持续时间变长。

1.5 异常情况

根据信用卡-所有人匹配结果，可以发现，总共三人持有多张信用卡：拥有 8 号车的 Lucas Alcazar 持有三张信用卡，卡号分别为 7117，4948 和 9551；拥有 10 号车的 Ada Campo-Corrente 持有两张信用卡，卡号

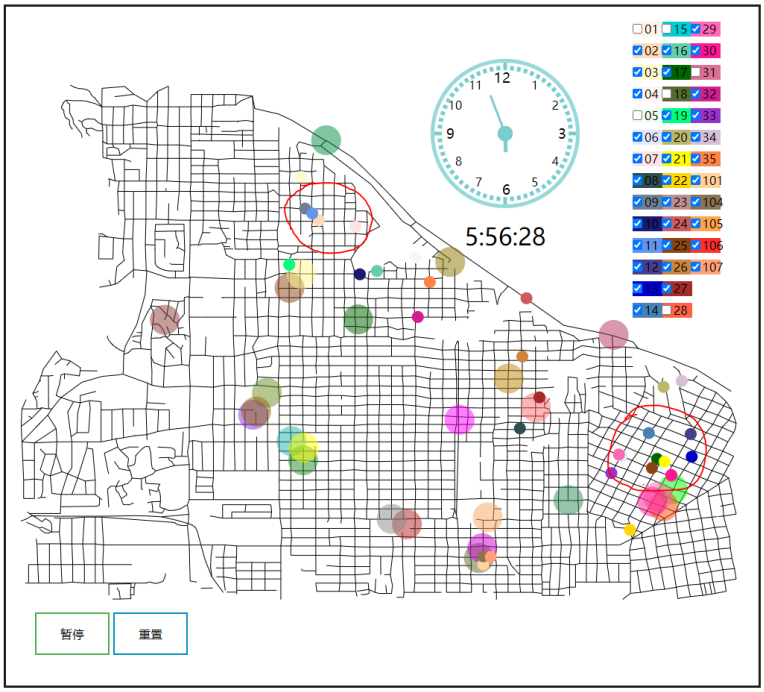

图 6

分别为 6691 和 8332；拥有 29 号车的 Bertrand Ovan 持有两张信用卡，卡号分别为 2681 和 3547。根据会员卡-所有人匹配结果，可以发现，只有一人持有多张会员卡：拥有 29 号车的 Bertrand Ovan 持有两张会员卡，卡号分别为 L3295 和 L9362。

2 问题二

2.1 问题分析

通过员工不同的雇佣类型，可以将所有员工划分为不同的组，包括 Information Technology，Engineering，Executive，Security 和 Facilities。同一组成员的关系表示官方关系，不同组成员的关系表示非官方关系。每隔半小时取一次时间，以及该时刻下的所有车的 gps 坐标，如果该时刻没有 gps 坐标记录，则用之前最近的时刻的 gps 坐标代替。对于每一时刻，利用所有车辆的 gps 坐标信息，计算两两之间的距离。该距离反映了两辆车对应的两位员工关系的强弱。距离越近，说明关系越强；距离越远，说明关系越弱。通过设置一个阈值，并过滤掉距离大于该阈值的所有关系，可以得到部分员工之间的关系，其中包括组内的员工官方关系和组间员工非官方关系。由于组内的成员由于雇佣形式的一致，因此所有组的组内成员之间都建立关系，而组间成员之间的关系根据强弱决定是否保留。按照上述方法，能够得到每一时刻下部分成员的关系网路。

2.2 可视化设计

我们通过力导向图来表示各个组和成员之间的关系远近，如图 7 所示。在力导向图中，每个结点代表了一个人。我们为力导向图设置了连接力，力的大小根据是否同组而不同，同组间的连接力较大，不同组的连接力较小，这样就可以保证力导向图同组的节点互相靠近。此外，我们保留了不同组之间的连接力，并且用line元素显式地表示不同组之间的连接力，力的大小会改变line元素的粗细，因此我们可以在力导向图中着重观察组间的连线，从而发现非官方关系。最后，我们为图中的结点都设置一个斥力，使得没有连接的结点相互远离，相连接的结点相互靠近。随着时间的变化，力导向图也会随之变化，这样我们通过不断地播放和暂停动画来查看人物间的距离动态变化情况。

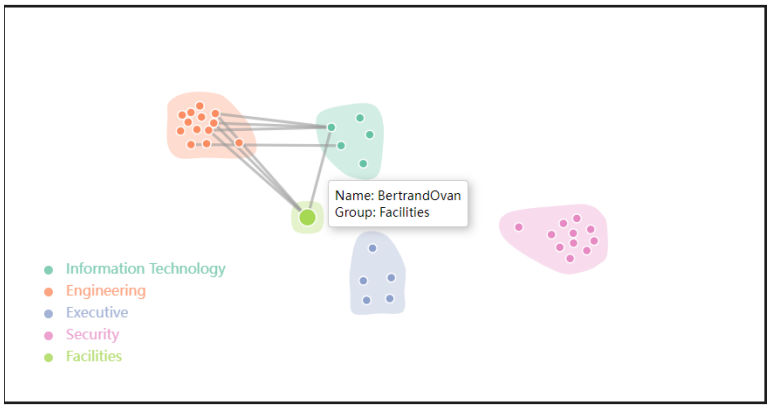

图 7

通过拉拽特定结点，我们可以分析该结点与其他结点的关系紧密程度，可以看到 BertAndOvan与Engineering 和 Information Technology 部门中的几位人物有着一些非官方关系。

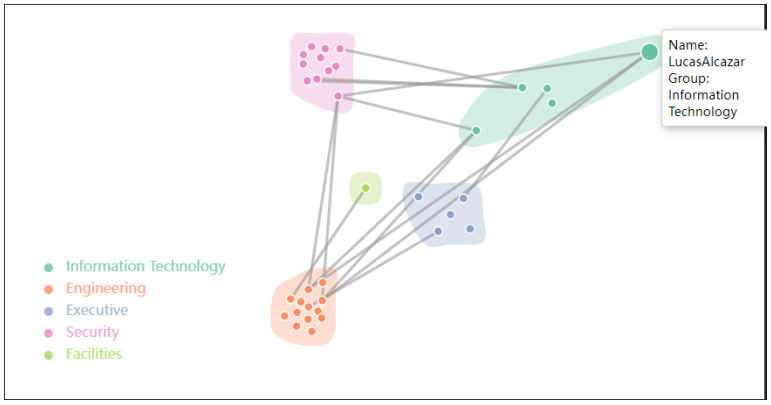

图 8

又例如在该图 8 中，我们可以发现LucasAlcazar可能与Engineering部门的几位人物有一些非官方关系。

2.3 调查结果

通过力导向图，可以发现存在一些值得注意的非官方关系，以及在这些非官方关系中高频出现的人物。下表按照关系从强到弱列出了部分非官方关系。首先，Nils Calixo 出现在较强的几个非官方关系中，这位员工值得注意；其次，Lucas Alcazar 最多地出现在非官方关系中，同时，根据之前的信用卡匹配结果可以看出，该员工持有三张信用卡，因此该员工应该被高度关注；最后，Bertrand Ovan 也比较频繁地出现在非官方关系中，而且也持有了两张信用卡，并且他是唯一一名有登记车辆的 Facilities 员工，他与其他员工之间的关系值得被深挖。

1	Source	Group	Target	Group
2	Lars Azada	Engineering	Ingrid Barranco	Executive
3	Nils Calixto	Information Technology	Lars Azada	Engineering
4	Nils Calixto	Information Technology	Ingrid Barranco	Executive
5	Nils Calixto	Information Technology	Hideki Cocinaro	Security
6	Ingrid Barranco	Executive	Hideki Cocinaro	Security
7	Lars Azada	Engineering	Bertrand Ovan	Facilities
8	Nils Calixto	Information Technology	Bertrand Ovan	Facilities
9	Lars Azada	Engineering	Hideki Cocinaro	Security
10	Ingrid Barranco	Executive	Bertrand Ovan	Facilities
11	Hideki Cocinaro	Security	Bertrand Ovan	Facilities
12	Lucas Alcazar	Information Technology	Lidelse Dedos	Engineering
13	Lucas Alcazar	Information Technology	Ada Campo-Corrente	Executive
14	Lucas Alcazar	Information Technology	Birgitta Frente	Engineering
15	Lucas Alcazar	Information Technology	Kanon Herrero	Security
16	Lucas Alcazar	Information Technology	Bertrand Ovan	Facilities
17	Lucas Alcazar	Information Technology	Varja Lagos	Security
18	Lucas Alcazar	Information Technology	Isia Vann	Security
19	Lucas Alcazar	Information Technology	Hideki Cocinaro	Security
20	Elsa Orilla	Engineering	Willem Vasco-Pais	Executive
21	Lucas Alcazar	Information Technology	Marin Onda	Engineering

3 问题三

（1）10号15号16号半夜的非法交易行为

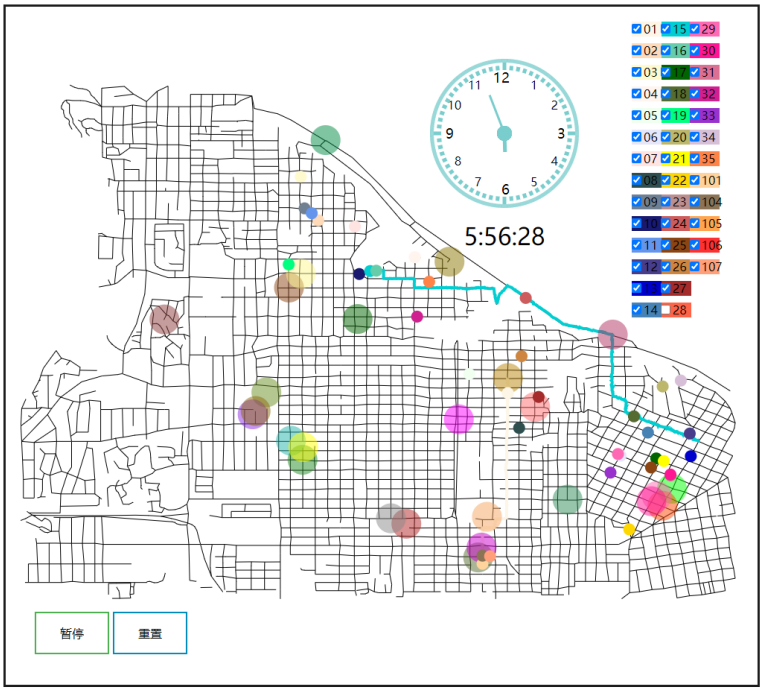

图 9

（2）9号车直接从公司路径跳跃至可疑地点（可能是11号的家），存在扰乱GPS隐匿行踪的行为，一晚上后，第二天和11号先后上班，并且次日起始点也存在偏移（红色标识）。我们怀疑9号与11号保持着不正当关系，并且不愿意暴露。

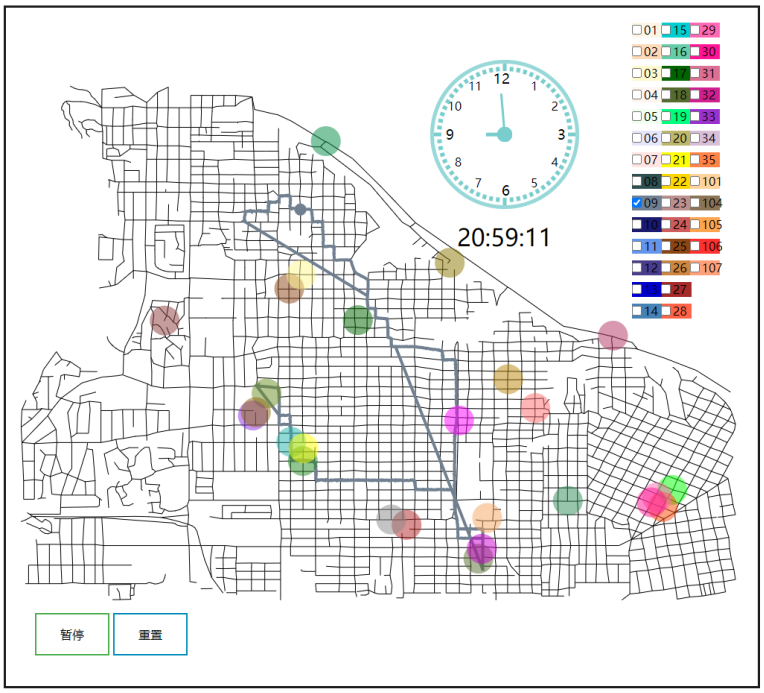

图 10

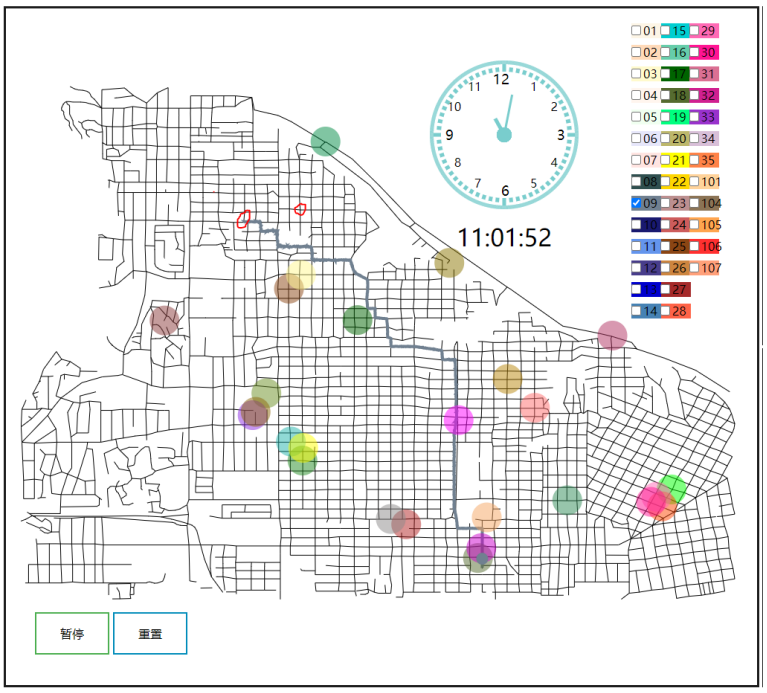

图 11

（3）15号车和24号车奇怪的关系，二人总是约在不同地点见面，然后以不同的路径返回。一天中相见的次数较多。

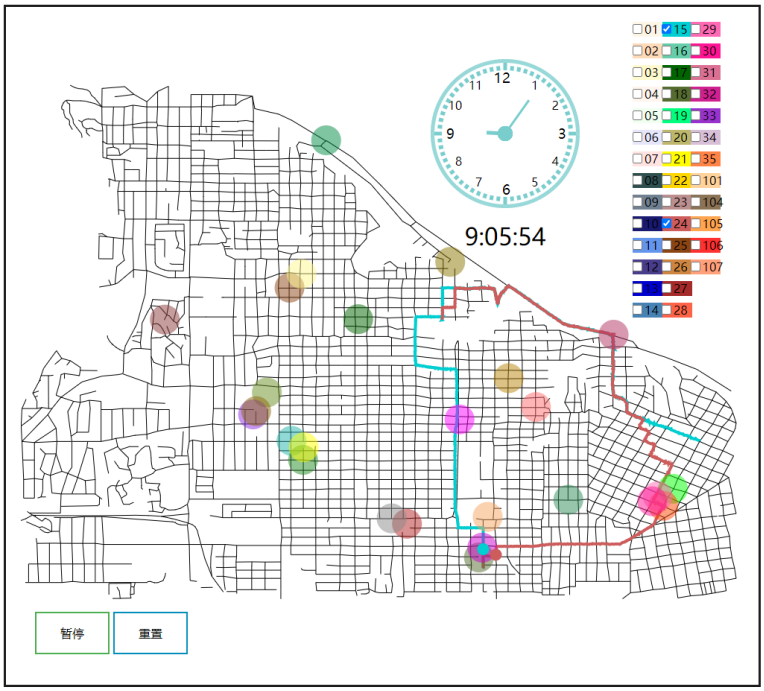

图 12

（4）4号，10号，35号几位大佬在高尔夫球场的会面，可能是谈论大事。

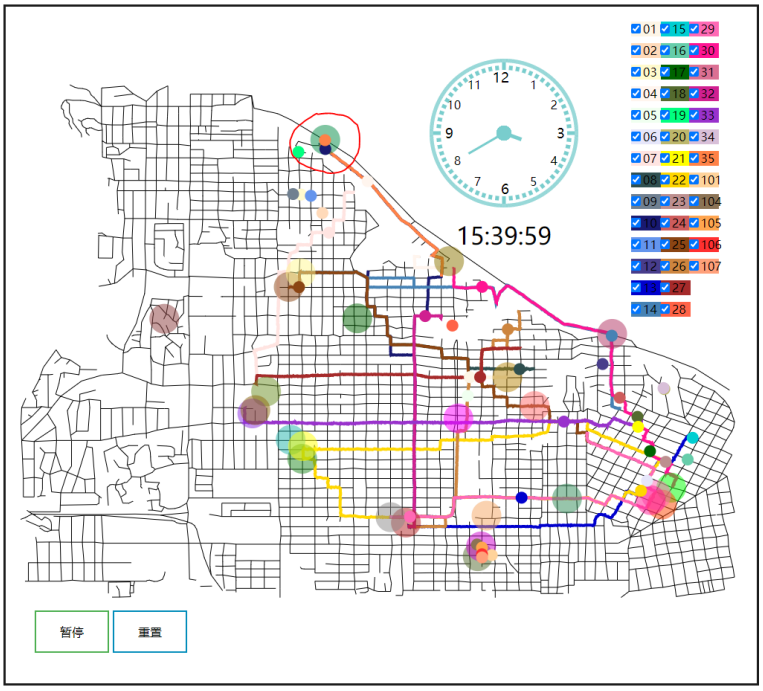

图 13

还有一些从消费数据上发现的异常

（1）1 月 13 日 19 时 20 分，在 Frydos Autosupply n' More的 1 号的信用卡上记录了一大笔消费。但是此时 1 号的车位于 Ouzeri Elian。

（2）一些凌晨时期的消费记录，比如 1、10、23、32 号员工，可能存在半夜的非法交易。

（3）异常的消费记录，比如 1 月 17 日 22 号在 Shoppers'Delight 的消费记录，积分卡记录了 269.33 美元，信用卡记录了 289.33 美元。可能涉嫌非法交易篡改消费记录。

小组成员贡献：

林恒旭：统筹规划，绘制可视界面，撰写报告

赵刚：数据分析，匹配关系破案，撰写报告

李秉轩：绘制可视界面，撰写报告，汇报展示