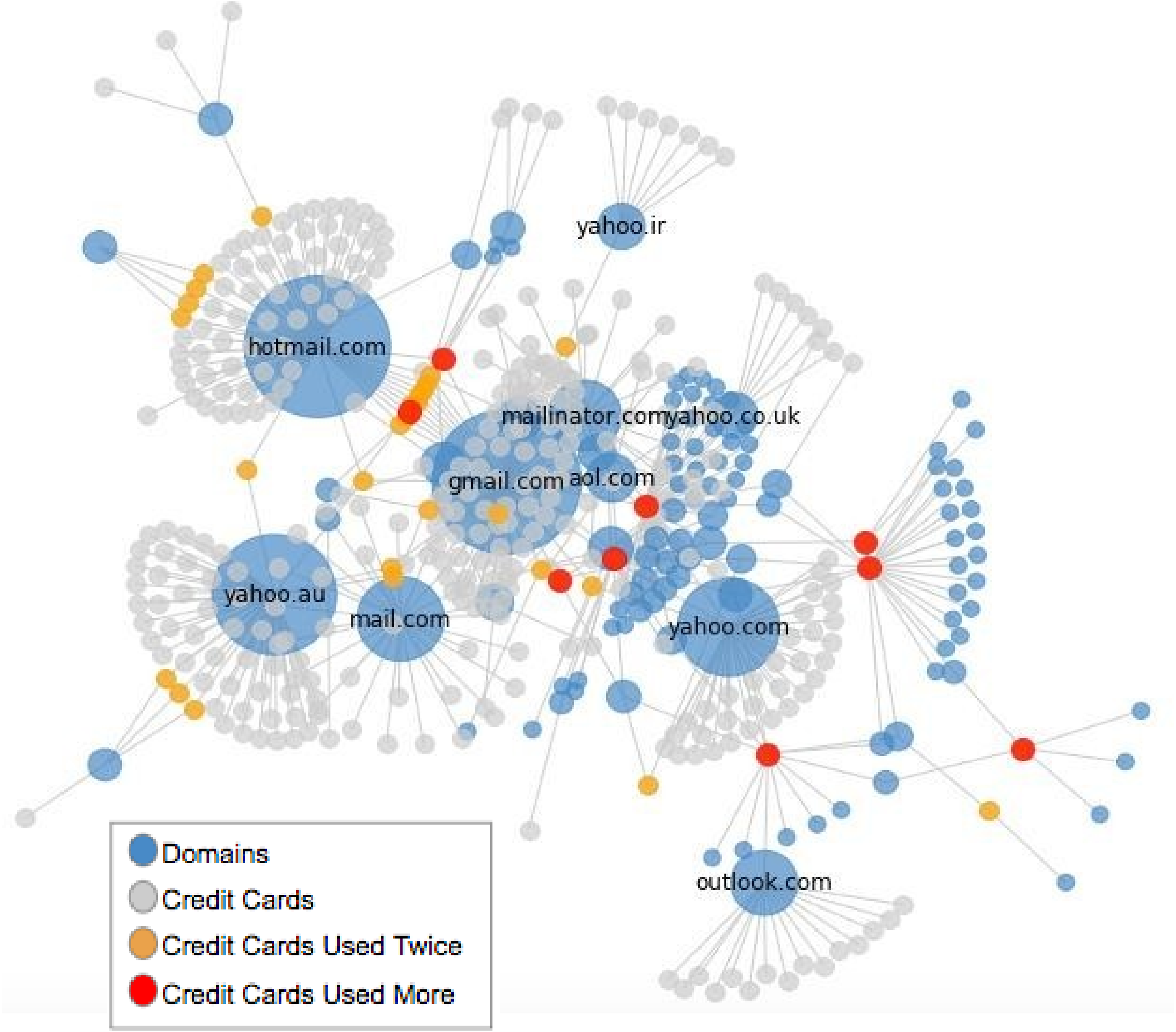
Analyzing Fraudulent Transactions

Network Analysis

- Visualizations

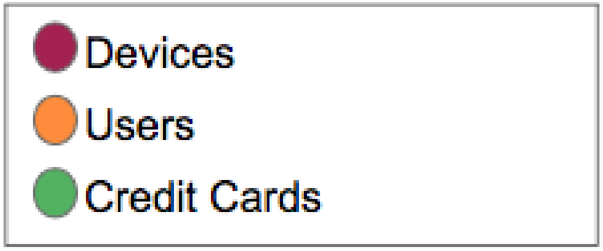
Credit Cards in Domains



The graph shows the relationship between credit cards and domains. Popular domains like Gmail and Yahoo receive a lot of credit cards. Some credit cards are used multiple times.

Interactive visualization with devices, users, and credit cards.

https://github.com/Italosayan/A2-D3.js/tree/master/ravelin



● Devices

● Users

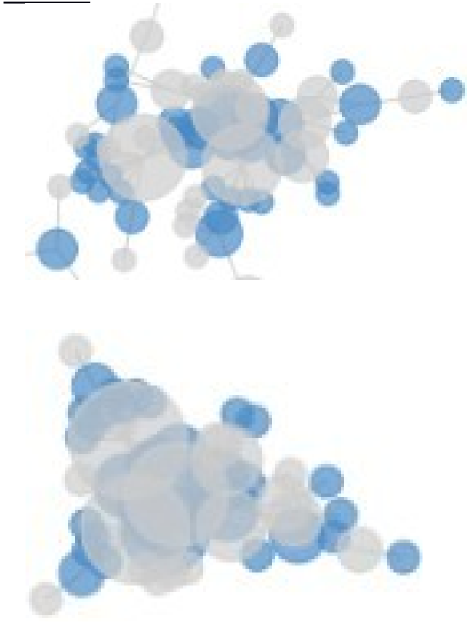
● Credit Cards

The central red circle is just a **pivot** where all devices come from. Devices are linked to users. The users are linked to credit cards. If nodes are clicked they are stored. This allows for further analysis.

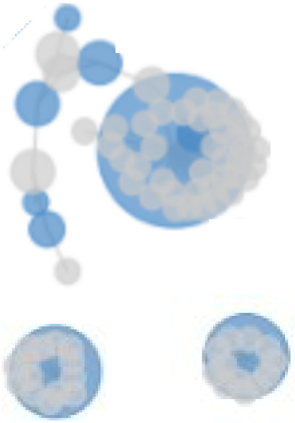
- Different types of fraudulent behavior

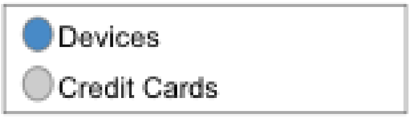
  The relationship between devices and credit cards was plotted and two behaviors can be identified. Two examples of each behavior displayed.

Same credit cards used on multiple devices



A device using multiple cards





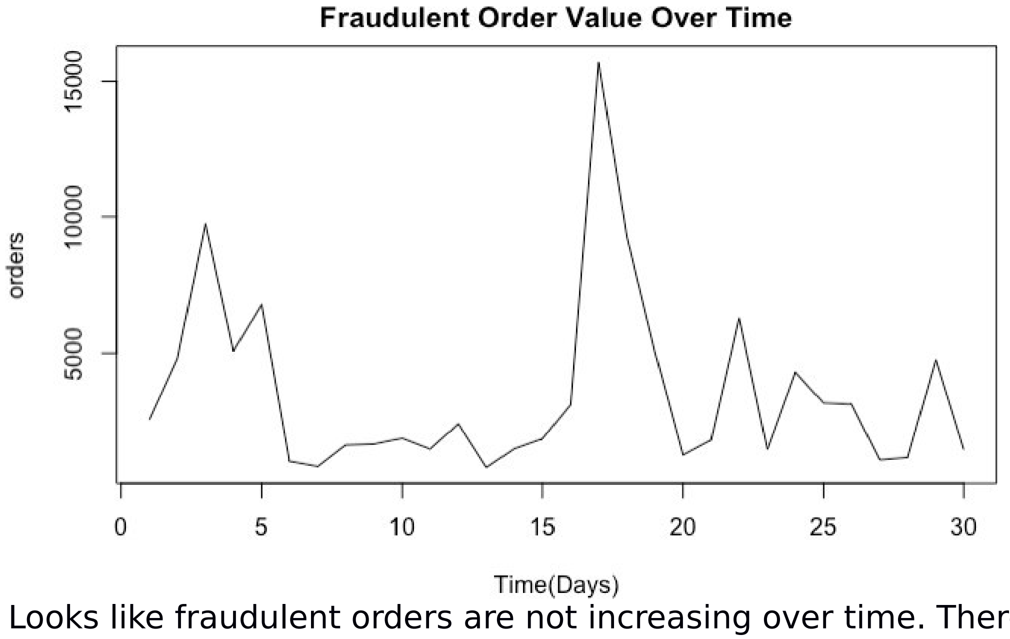
Possible explanation:
On the left side, a robber has acquired a few credit cards and is using multiple devices to make transactions. On the right side, a robber has acquired multiple credit cards.

Conclusions:

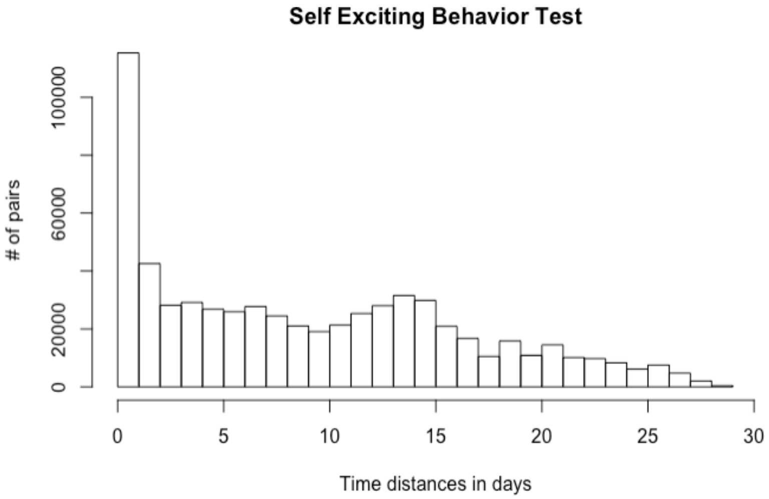
- ✓ Some domains are more popular than others.
- ✓ The relationship between credit cards, user ids, and devices is complex. Sometimes a group of devices uses a group of credit cards. Sometimes all transactions come from only one device.
- ✓ A user stamp can’t have two devices.
- ✓ There are 6 instances were a credit card is used by more than 5 devices. A single credit card has been used by as much as 16 devices. Multiple robbers could have access to the card. Could indicate the existence of some type of black market for cards.

Location Analysis

- Exploratory Analysis:



Looks like fraudulent orders are not increasing over time. There is not a trend component to the series.



Self-exciting behavior[7] is when the occurrence of an event increases the probability of subsequent nearby events. In order to investigate if fraudulent orders follow SEPP behavior, the space distance and days distance is calculated on all pair of events. Then events far from each other are removed (more than 20,000m apart). Finally, the time distance of the nearby events is plotted in a histogram. **Looks like there are more than 100,000 pairs of nearby delivery locations that were used on a short window of 2 days.** The SEPP behavior of processes like burglaries, earthquakes or tweets is more intense.
More data is necessary to make a final conclusion.

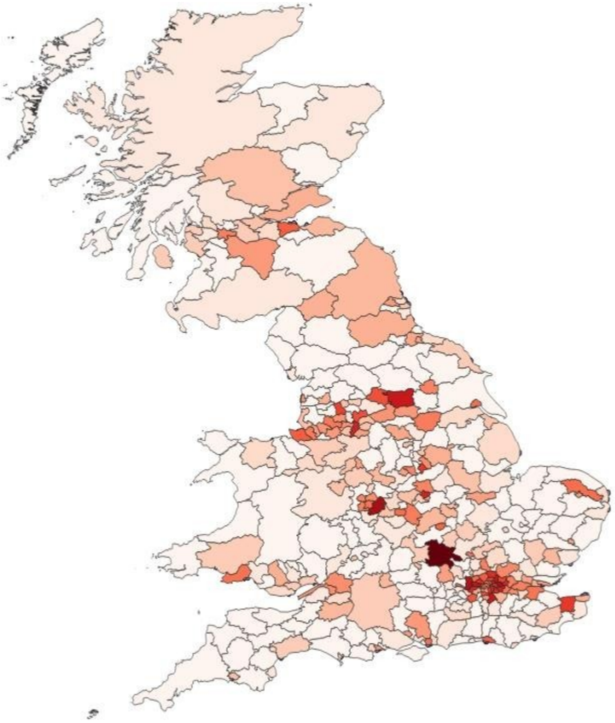
[7] More information https://github.com/Italosayan/P-P-P

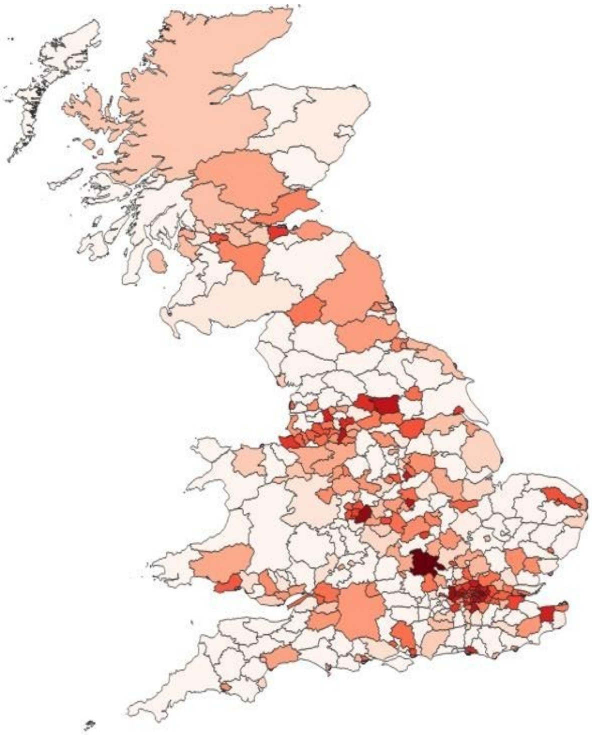
- Fraud detection team allocation

  For now, I would approach resource allocation using hotspots. A naïve approach because we assume previous delivery locations will be used again. It’s not a bad approach considering that 31% of delivery locations have been used more than once on different dates.

  The next step would be to fit spatiotemporal models[8] to the data and benchmark using the naïve approach as a starting point.

- Hotspot by number of orders Hotspot by value





The hotspots maps are similar. It makes sense considering that the distribution of order amounts doesn’t have many extreme outliers.

- Limitations

  A robber wouldn’t mail orders to his personal address.
  Whether an order is fraudulent could be determined several days after the event. Fraudulent events could never be discovered and would never be labeled as such.

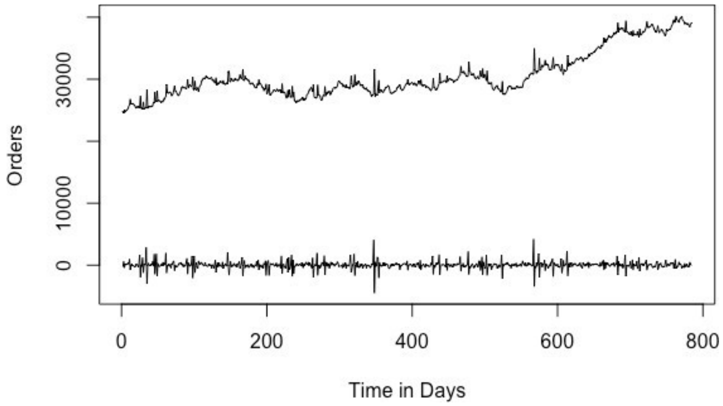
[8] More information: https://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx

Conclusions:

- ✓ 31% of delivery locations have been used more than once on different dates.
- ✓ Fraudulent order didn’t experience an upward trend on November.
- ✓ The hotspot resource allocation strategy would be a good starting point.
- ✓ Fraudulent deliveries are a spatiotemporal process. Multiple models could be fitted to the data.
- ✓ The limitations to the data also limit any model we fit the data to.

- Modeling: Predicting number of orders

<u>Simple model: Naïve prediction.</u>
Today’s orders will be the same as yesterday. The plot shows the regular order quantity time series and the errors of the naive method.

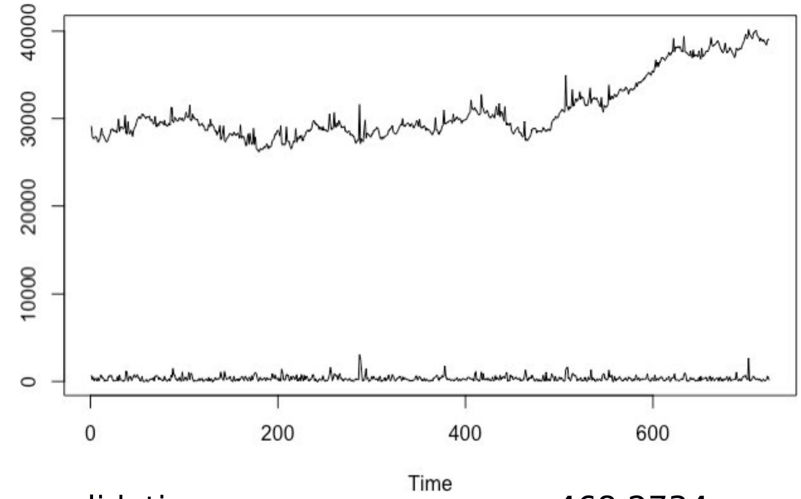


Cross-validation mean average error[4]: 647.65

[4] More information on time series cross validation. https://otexts.org/fpp2/accuracy.html

Multivariate Regression

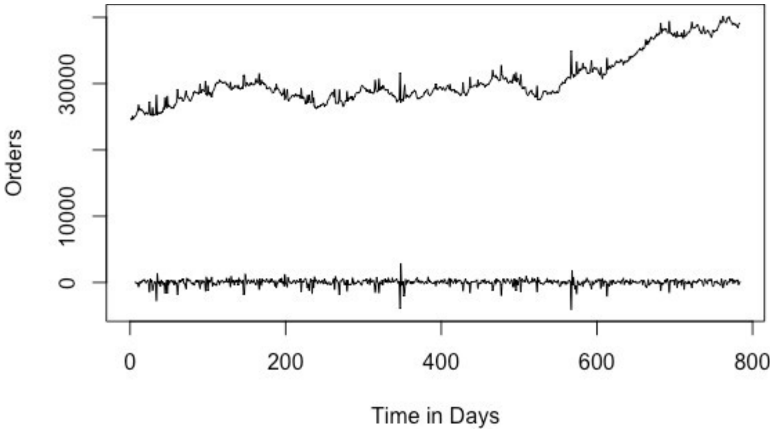
Drivers available, orders of the day before and weather circumstances were used to fit a multivariate linear regression. The predicted variable is the order difference between t-1 and t. Predicting the change instead of the raw quantity gives stationarity[5] to the series.



Cross-validation mean average error: 468.2734

Neural Network

A feed forward network with 5 input layers and 3 hidden was fitted. The inputs are the 5 past periods[6]. The plot shows the regular order quantity time series and the errors of our method. Similar to the dynamic regression, the change between t-1 and t is predicted.



Cross-validation mean average error: 520.5733

[5] https://en.wikipedia.org/wiki/Stationary_process

[6] Nnetar function of the forecast package in r.

- Important variables

  In order to find important variables, we can fit a lasso regression. The variables that don’t affect our predictions will have a low coefficient. The variables were normalized in order to use the coefficient as a measure of importance.

Orders of the previous period	3646.35
Strike Day	2614.53
Very Rainy Day	1102.52
Available drivers	260.23
Rainy Day	73.43

Conclusions:

✓ The multivariate regression had the best prediction performance. It’s the only model that takes into account the weather variable.

✓ The neutral network approach used previous order quantities and had a low error. Including weather circumstances will improve the model.

✓ Weather and lagged orders are the most significant predictors of future orders.