**Featurespace Interview Task: Identifying A Common Point Of Compromise**

**Background Information**

**Transaction Fraud**

Many of Featurespace’s clients are commercial banks, who issue debit and credit cards to customers. The customers then use these cards to make purchases at various merchants (for example, Amazon, Tesco etc). Information about each purchase or attempted purchase is stored as a transaction record by the commercial bank.

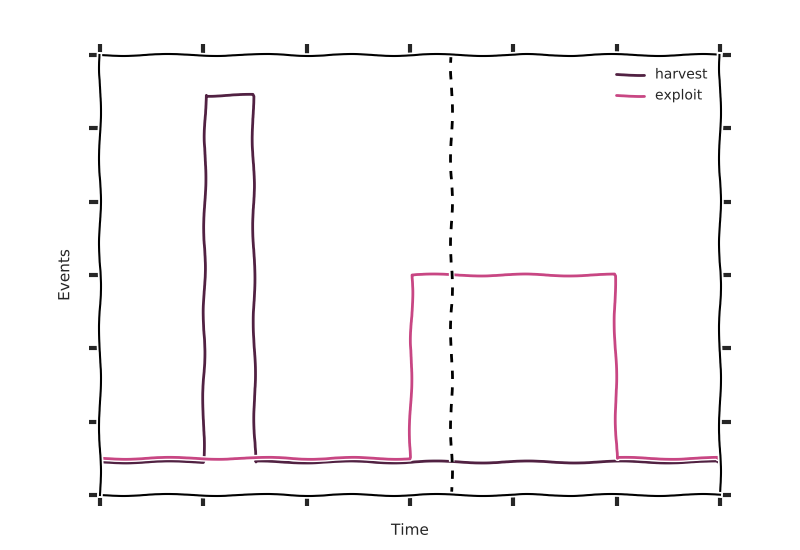
A fraudulent transaction is an unauthorised use of a card. A fraudster may use a stolen card or card details (card number, security code, PIN number) to make a purchase or to withdraw money from an account. The information required to commit fraud may have been acquired opportunistically, for example by finding a lost card. However, in many cases, card data will have been acquired at a common point of compromise (see below).

**Common Points of Compromise**

A common point of compromise is a breach point used by fraudsters to collect card data. This can be a physical location, such as a skimmer fitted on an ATM; alternatively, a virtual breach, such as data stolen from a merchant’s database of transactions, can be the point of compromise. Identifying common points of compromise is important for commercial banks, as it allows them to block or monitor other accounts at risk of fraud.

Exploitation of a common point of compromise typically occurs in a two-stage process. This is illustrated in the schematic plot below.

In the first stage of the compromise, account details are harvested from the compromise point. In the second stage of the compromise, these compromised account details are used to make fraudulent transactions. There is typically a delay of several days between harvest and exploit, as it takes time for a fraudster to process the data that they have harvested.



In a set of transactions, a common point of compromise can be identified when an unusually high number of defrauded accounts transacted with the same merchant some time before fraudulent transactions were seen for that account.

**Task Description**

You will receive a simulated dataset of transactions from a 60 day period. These transactions are supplied in a single .csv file. The format of this file is given below.

One of the merchants in this dataset is briefly compromised and accounts that transacted with that merchant during that period are at risk of fraud.

Your task is to identify which merchant is the common point of compromise, on which days accounts are compromised and on which days the compromise is exploited. You can use any programming language you choose for the task.

Once you have identified the compromised merchant, prepare a brief presentation (5 minutes) that describes the method you used to find the compromise.

Please submit:

* The identity of the compromised merchant
* The dates of the compromise
* The date of the exploitation
* The code that you used to identify the common point of compromise
* Your presentation describing your methodology

**Points to Note:**

* The dataset contains exactly one compromised merchant
* There may be some opportunistic fraud in the data, which is unrelated to the compromise

**Data Format**

**Transactions\_simple.csv**

|  |  |
| --- | --- |
| **Field** | **Description** |
| Account | Unique id of the account |
| Date | Date of the transaction |
| Fraud | Boolean flag for fraud |
| Merchant | Unique id of the merchant |

**Things you may want to include in the presentation**

* What is the question you are trying to answer?
* How did you approach the problem?
* How well did your solution work?
* How do you know?
* If you had more time, how would you refine the solution?