

EPSRC

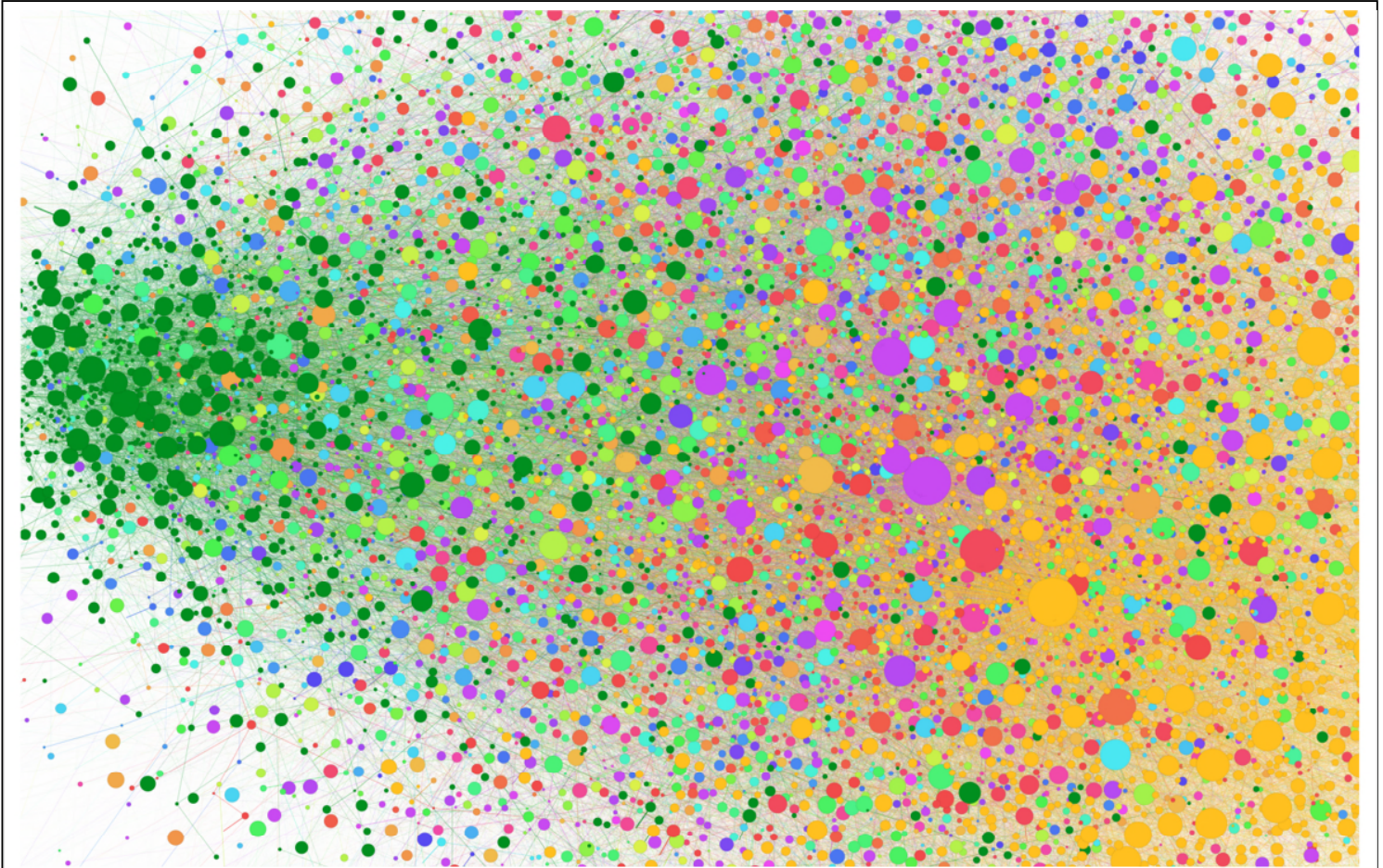
Engineering and Physical Sciences
Research Council



InFoMM

Industrially Focused
Mathematical Modelling

EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling



Visualization of a customer-product network, done with Gephi (<https://gephi.org/>)

Community Detection in Product-Purchase Networks



UNIVERSITY OF
OXFORD

dunnhumby
Roxana Feier



Table of Contents

1. Introduction	2
Background.....	2
2. Methodology.....	2
Transaction Data	2
Network.....	3
Communities.....	3
Analysis.....	3
3. Data Insights.....	3
Robust Community Assignments	4
Community Profiles	4
Comparing Network Partitions.....	5
4. Discussion and Conclusions.....	6
Network.....	6
Communities.....	7
Analysis.....	7
5. Potential Impact	7
References	7

1. Introduction

A personalised recommendation and discount programme is most successful when it identifies products a customer is likely to buy, regardless of previous purchase history. Focusing on grocery shopping data, we aim to identify groups of shoppers with similar behaviours and the corresponding products that they purchase. The store can then recommend products to customers that people similar to them have bought.

Background

Networks are structures used to model pairwise relations between objects. We represent customers and products as a network, where edges indicate whether a product has been purchased by a customer in a given time window. The edges can be weighted to reflect how often a purchase has occurred and thus keep more information from the transaction data.

In this project we use a technique known as *community detection* to identify groups of nodes – both customers and products – that are more densely connected with each other than with other groups. These communities should reflect not just past purchasing practices but also indirect connections between customers and products, “missing links” that could be incorporated in a recommendation system.

Transactions in a store can be represented as a network, with edges indicating which customers have bought which products.

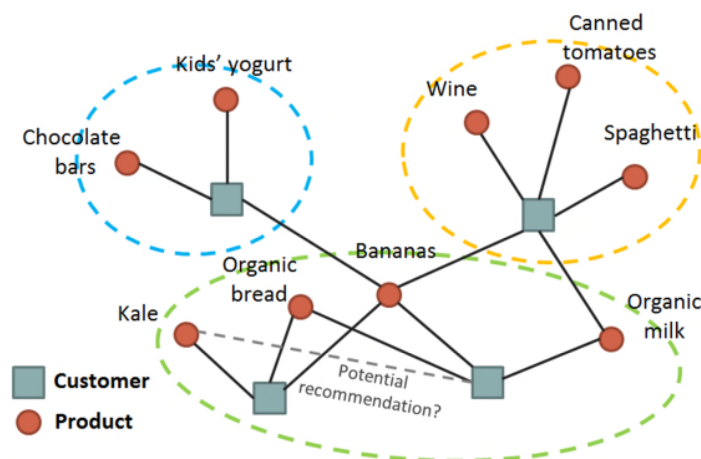



Figure 1: Example of a product-purchase network with three communities. Customers and products that are not directly connected but are in the same community suggest potential personalized recommendations.

2. Methodology

Our process consists of two stages: first, transactions are aggregated and represented as a network; then, an algorithm is used to identify densely connected communities. We will briefly mention some of the modelling choices affecting both stages of this process.

Transaction Data

We were provided with anonymised transaction data from stores of varying sizes in the Oxford area covering the period from March 2013 to March 2015. Results reported here are from the



smallest store in this data set and cover the 3-month period between July and September 2014, to avoid unusual shopping patterns observed around major holidays. This subset of the data consists of 8032 unique customers and 3737 unique products, covering 115238 transactions.

Network

Edges in the network connect customers to products they have bought in the time window considered, with a higher weight for products purchased more often. Three different weighting schemes were implemented. We focus on one which scales the number of products of a certain kind bought by a customer by the total item count for that customer. Thus, if one person bought 3 apples, 5 bananas, and 2 oranges across all shopping trips in the 3-month period, the edge connecting this person to the node for apples will get a weight of $3 / 10$.

Communities

Having set up the data as a network, we use an existing algorithm known as the Louvain method, implemented in MATLAB [1], to find its community structure. The number of communities is determined automatically by the algorithm. However, the user needs to specify a null model – a mathematical description of what a network without community structure would look like – as well as a resolution parameter which can be increased to reveal finer communities in the data. For now, we use the standard null model from the literature and the default resolution value.

In general, community detection is known to be computationally difficult, meaning that the “best” solution is almost never found by the algorithm. Instead, the code returns one of several locally optimal community assignments, which are solutions that could not be improved without moving many nodes around to different communities.

Analysis

The output of the code is an assignment of each node to a community. Understanding and assessing the quality of the output is as much (if not more so) part of the modelling as getting the community assignments in the first place. Specifically, we ask two questions:

1. Which community assignments are the most meaningful? What can we say about the customers and products in these communities?
2. When do two partitions of a network into communities (obtained, for instance, by choosing different parameter values) convey the same information?

We explore these questions in the next section.

3. Data Insights

For the store analysed in this report, the median customer shops just once in the 3-month period considered. Realistically, we cannot expect to get the proper community assignment for such a customer, or for a niche product which is only bought a handful of times. It is then important to understand which community assignments are meaningful and which are not so we can focus the analysis efforts on the former.

Qualitatively, *communities* are sets of nodes that are more connected with each other than with other nodes in the network.

A *network partition* is a division of all the nodes in the network into distinct communities.

Robust Community Assignments

The algorithm used for community detection has some stochastic steps, which means that node assignments can differ between runs. We can use this to our advantage. By considering pairs of nodes that are (almost) always classified together across several trials, we can focus on the more robust structures in the network.

We are interested in community assignments that are persistent across a large number of runs of the community-detection algorithm.

An *association* or *co-classification matrix* is used to record, for each pair of nodes, what fraction of the time they are assigned to the same community. For example, if entry (100, 200) in this matrix is equal to 0.8, this indicates that nodes 100 and 200 are placed in the same community in 80% of the trials. After proper rearranging of rows and columns, the association matrix reveals two sets of nodes – both customers and products – that are classified together in nearly all 100 trials ran. These blocks of nodes are subsets of the two largest communities initially found by the algorithm. We study them in more detail in the next section.

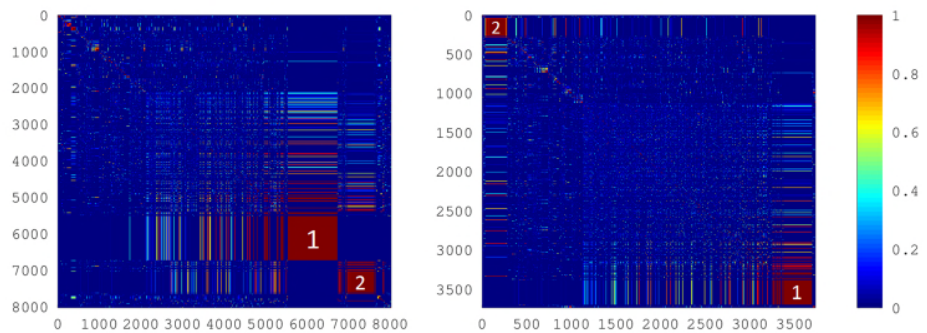


Figure 2: Association matrix for customers versus other customer nodes (left) and products versus products (right). A third association matrix for customers versus products is calculated but is not shown here. We notice two blocks of nodes – both customers and products – that are consistently assigned to the same communities across all trials.

Community Profiles

We describe the composition of these two blocks of nodes in terms of average statistics (such as money spent per shop or number of items bought) relative to the overall network. Additionally, we show in Fig. 3 the distribution of customers across an existing segmentation, looking for categories which are over-represented in either of the two blocks relative to the distribution observed for the network as a whole. Similarly, Fig. 4 shows the distribution of products across broad categories, which bundle similar products together.

The largest identified community is made up of customers focused on healthy and adventurous eating.

We find that block 1 contains customers that shop more often, buy more items and spend more money each time compared to the average shopper. These customers are overrepresented in categories defined by healthy eating and adventurous tastes (segments 1, 2, 3). Products assigned to this community are overwhelmingly produce, meat, and “hard groceries” (a category which includes all items relating to foreign cuisine) and span the whole popularity range based on the number of unique items sold and the number of customers that buy them. They are slightly cheaper than the average product in the network.

Customers in block 2 also shop more often than the average customer but they buy fewer items and spend less money each time, and are overrepresented in categories defined by tried and tested products as well as products aimed at children (segments 4, 5, 7). Compared to other



A second set of customers buys a disproportionate amount of crisps, soft drinks, chocolates, ready meals, and aisle end products which are usually on promotion.

customers, they buy more “soft groceries” (which include soft drinks, crisps, chocolates), “cabinet provisions” (ready meals and sandwiches, cheese, yogurt), and “end panels units” (products displayed at aisle ends, usually on offer). These are more popular products based on items sold and the number of customers that buy them. They are cheaper than the average product in the network.

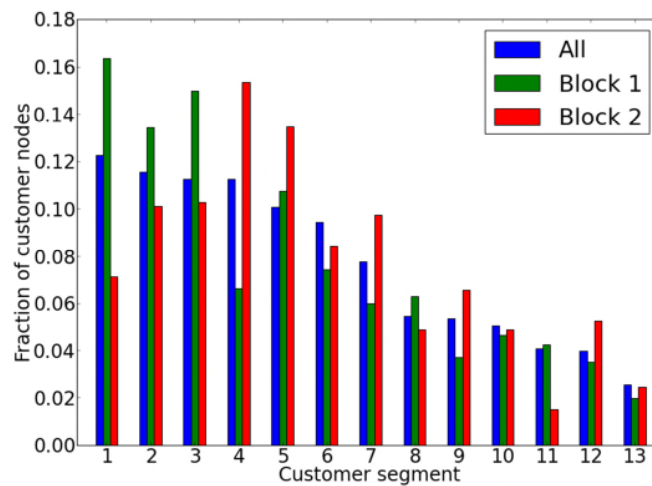


Figure 3: Distribution of customers across pre-defined segments

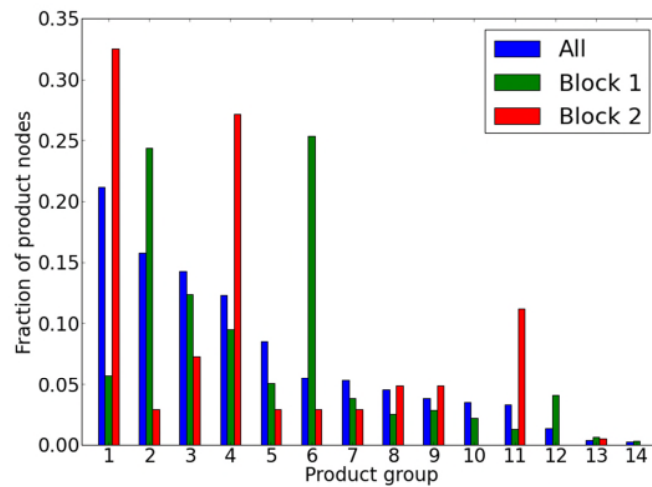


Figure 4: Distribution of products across pre-defined groups

Comparing Network Partitions

It seems that the second block of customers profiled above may be more influenced by promotions than the first group. We can test this by building a second network where all purchases of discounted products are excluded from the transaction data. We ask how the identified communities of customers are different in the two networks.

The *interlock matrix* is one way to quantify the similarity of two network partitions by looking at the number of nodes shared by pairs of communities in the two partitions.

When all transactions are included, the two largest identified communities have 2318 and 1505 nodes, respectively. When discounted purchases are omitted, the two largest communities have 2005 and 492 nodes, suggesting that the second group of customers described earlier comes apart. To explore this quantitatively we look at the pairwise overlap of communities of customers from the two partitions of the network, described by the *interlock matrix*. High values in this matrix correspond to communities which share a large number of common nodes. In our case, these high values are encountered for groups of customers that are minimally affected by discounts, emerging in either of the two networks considered.

It is generally incorrect to simply pick the highest values in the interlock matrix, as larger groups have more nodes in common simply due to chance. Instead, for each pair of communities we can calculate a *p-value* with the assumption that the two partitions are independent of each other (i.e. that the null hypothesis is true). Low p-values correspond to overlaps that are much higher than what would be expected at random.

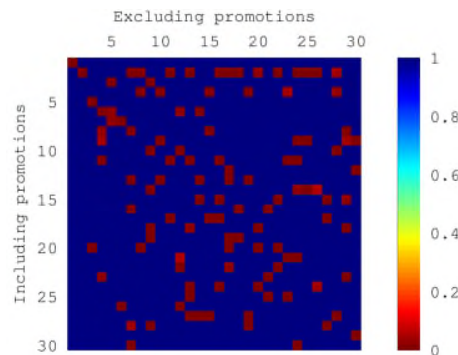


Figure 5: P-values of community overlaps between the full network and the network built after excluding all items bought on promotion. For clarity, we show just the 30 largest communities in each partition. We consider p-values higher than 0.05 insignificant and set them to 1.

The matrix of p-values is shown in Fig. 5. We see that the two largest communities in the two partitions show a significant overlap with each other and not with any other community, suggesting that these customers are not very much driven by discounts. In contrast, the second largest community in the full network breaks down when promotional items are excluded, showing significant overlap with several different communities from the second partition. This agrees with the intuition developed in the preceding section, although it also raises the possibility that the first block contains more niche products than the second block, products that are more likely to be on promotion. Nevertheless, this discussion illustrates one way to compare network partitions in a systematic way, using in this case a statistical test.


Future work will impact all steps of the modelling process: building the network, identifying communities algorithmically, and analyzing the output to derive useful insights for our application.

4. Discussion and Conclusions

Before translating community assignments into product recommendations, additional work is needed to assess and improve the quality of these assignments. Below we describe some possible directions for future work.

Network

It is possible to construct the customer-product network so that it better represents the underlying transaction data. For example, we can use a different weighting scheme, incorporate



known information into the network (such as customer metadata or product hierarchies), or set up the network to be time-dependent rather than static.

Communities

The analysis done so far suggests that there might be additional structure in the two largest identified communities. These structures can be explored at a finer scale by changing the resolution parameter in the community-detection algorithm. Another idea is to use a technique known as consensus clustering to extract robust community assignments from a set of partitions obtained through repeated runs of a community detection algorithm.

Analysis

The final step in the process is to understand what community assignments say about the different types of customers that come into a store and the products that they buy. There are many questions that can be asked on this, some of which we have already explored in earlier sections: Which community assignments are meaningful and which are not? When do two partitions convey the same information? How should identified communities translate into product recommendations? How can we test whether recommendations are effective?

5. Potential Impact

The short-term impact of this work, already partially achieved, is to start a discussion about networks within the company and demonstrate that community detection techniques provide practical insights into grocery shopping data. Going forward, this work could suggest ways to test existing customer or product segmentations and identify some of their potential shortcomings. We might also identify other projects in the business that could benefit from a network theory approach. In the long term, this approach could suggest new customer and products segmentations and form the basis of a personalised recommendation system (which could be tested on a subset of customers to compare its effectiveness against the status quo).

Outside the company, this work will likely require new contributions to network theory. We will have to improve some existing tools to handle the fairly large networks we work with. Finally, we are applying community detection techniques to a type of social network – product purchasing – that has not been studied extensively before.

Giles Pavey, Chief Scientist, said *“Roxana’s research has really caught the imagination of our data science team and beyond.”*

Lorna Barclay, Data Scientist, reported *“Initial results show strong promise that understanding the bipartite networks made up of products and customers could power significant improvements in the design of online and offline services for shoppers to find relevant products.”*

Rosie Prior, Research Manager, noted *“Her research suggests that the use of network theory could provide a superior method for creating customer segments within our business.”*

References

1. IS Jutla, LGS Jeub, and PJ Mucha (2011-2014). *A generalized Louvain method for community detection implemented in MATLAB*. <http://netwiki.amath.unc.edu/GenLouvain>