Case Study: Community Pantry

Olivier Binette

November 10, 2020

Introduction

A Durham pantry provides food and basic necessities to students in need of assistance. Since the beginning of the COVID-19 pandemic, pantry usage has mainly been restricted to the weekly bag program: students fill bag orders through a Qualtrics survey, and the bags are then delivered to them or ready for pick-up on the next Saturday.

As a volunteer statistician at the Pantry, you are tasked to:

1. describe the composition of pantry users through bag order records; and
2. provide recommendations regarding how the pantry could improve its data collection efforts to better evaluate the need of students.

Your analysis will assist the Pantry in its negotiation with University stakeholders and will help them prepare for the upcoming Campus Food Insecurity Symposium.

Data

The raw Pantry records consists of weekly bag orders filled between June 24 and November 4 2020 using a Qualtrics survey. This is an online survey made available to all of Duke's graduate and professional students. The survey is mainly advertised over email in periodical newsletters and it records all bag orders for the given time period. Its content is summarized in Table 1.

In order to protect the privacy of pantry users, the raw records have been anonymized and only a subset of this data is made available to you. The personal identifiers `name`, `phone` and `email` have been encrypted using an MD5 hash function. Furthermore, only the non free-form survey answers (questions 30-34) have been provided to you for your analysis. This data is contained in the `order_data.rds` file and the anonymization pre-processing script can be found in `encrypt_responses.R`.

Methodology

Many pantry users have filed more than one order in the considered time period. Your first task is therefore to resolve individual pantry users using this data. Next, you will consolidate survey answers for each individual. Finally, you will provide summary statistics and provide recommendations regarding the Pantry's data collection efforts.

Task 1

The fields `name`, `phone` and `email` are all expected to contain variations. For example, I could have entered my name as "Olivier", "Olivier B." or "Olivier Binette" in different days, or pantry volunteers may have recorded

	Question no.	Question	Answer form
Identifiers	2	First name and last initial	Free form
	3	Duke email	Free form
	4	Phone number	Free form
Order	5	Delivery or Pickup?	Multiple choices
	$6-7$	Adress and delivery instructions	Free form
	8	Food allergies	Free form
	9	Number of members in household	1-2 or 3+
	10	Want baby bag?	Yes or no
	$11-29$	Order items	Multiple choices
Survey	30	Degree	Multiple choices or Other
	31	School	Multiple choices or Other
	32	Year in graduate school	Multiple choices
	33	Number of adults in household	Multiple choices
	34	Number of children in household	Multiple choices
	35	Main challenges accessing food	Multiple choices or Other
	36	Feedback	Free form

Table 1: Summary of the Qualtrics "Weekly Grocery Bag Request Survey" questions. Note that number of sub-questions which are not relevant to this analysis have been omitted.

my name as "Oliver". I have two duke email addresses: olivier.binette@duke.edu as well as ob37@duke.edu, and on some days I could have used my gmail address. My phone number changed a few months after I moved to Durham.

In order to identify unique individuals, perform deterministic record linkage using the three fields `name`, `phone` and `email` in combination: define two records to be a match if they agree on at least one of these fields. Given this record linkage, assign a unique entity identifier to each pantry user.

Hint: construct an adjacency list and use `igraph::components()` together with `igraph::graph_from_adj_list()` in order to define unique entity identifiers.

Optional challenge: Can you perform this deterministic record linkage in much less than $\mathcal{O}(n^2)$ time, where n is the number of records? Hint: use sorting to efficiently construct the adjacency list. Do you think that using a disjoint-set data structure to find and managed connected components would be more efficient than using `igraph` as above? Discuss.

Task 2

Construct a data frame where each row represents a unique pantry user and each column is a representative answer to the survey question. This does not have to be complicated. For example, you can choose the representative answer to be the most common available (non NA) answer for this individual.

Task 3

Given the above, answer the following questions:

1. How many individuals have only used the pantry only one time in this time period?
2. What is the composition of degree type, order type and year among pantry users?
3. Does the above differ between one-time versus recurrent users?

Furthermore, discuss the following: How could the pantry improve its data collection efforts in order to gain more meaningful insights into its users and their needs? Keep in mind the need to protect the privacy of the pantry users.

Learn more

I recommend the book *Algorithms, Fourth Edition* by Robert Sedgewick and Kevin Wayne to learn more about data structures and algorithms.

> "Good algorithms can make the difference between being able to solve a practical problem and not being able to address it at all."

There is a free MOOC by the authors on coursera: https://www.coursera.org/learn/algorithms-part1

- The "Union-Find" module of Week 1 covers the disjoint-set data structure and the union-find problem.
- Week 3 covers sorting algorithms.
- Part II of the MOOC, Week 1, covers graph data structures and algorithms to find connected components.