Computer Science with Applications III: Final Report
Mitsue Iwata, Kyle Magida, and Scarlett Swerdlow

Data

We worked with the Yelp Dataset Challenge dataset. It includes about 1.6 million reviews by 366,000 users for 61,000 businesses in ten cities. The dataset is 1.5 gigabytes. We used three JSON files: one for reviews, one for users, and one for businesses. (There are also JSON files for tips and check-ins, which we did not use.)

Hypothesis

We did not enter the project with a hypothesis so much as a series of questions we hoped we could answer. We were interested in how businesses relate to each other are based our analysis on pairs of businesses joined by common reviewers. Specifically:

- Which pairs of businesses have the highest “lift” (a term we will explain shortly)?
- Which businesses are most connected through a shared customer base?
- Can we predict whether a business will have high connectivity from other features?

Algorithms

We relied on three algorithms for our analysis:

Association rules

Association rules are essentially a technique to identify a basket of goods that are commonly purchased together. Association rules are all around us. For instance, on Amazon: people who buy *MapReduce Design Patterns* also buy *Hadoop*. In the context of our project: people who review business A also tend to review business B.

There are three key statistics in association rules. The first is support, which is the probability of an event. For instance, the support for business A being reviewed is:

$$Support\,(A) \;=\; Pr\,(A) \;=\; \frac{\#\ users\ who\ have\ reviewed\ A}{\#\ users}$$

The second key statistic is confidence, which is the probability of an event given some other event. For instance, the confidence for reviewing business B given a review of business A is:

$$Confidence\ (B \mid A) = \frac{Support\ (A\ \&\ B)}{Support\ (A)}$$

This leads to the last statistic: lift, or the increase in the probability of an event given some other event. Mathematically:

$$Lift\ (B \mid A) = \frac{Confidence\ (B \mid A)}{Support\ (B)}$$

We discuss how we calculated each of these statistics in the next section.

Network analysis

In addition to pairwise relationships between businesses on Yelp, we were also interested in the network of Yelp businesses. We learned, though, that network analysis and visualization does not scale well. Therefore, we used the association rules to approximate the network of Yelp businesses. Specifically, we focused on businesses with high lift—in other words, businesses that are closely connected to other businesses through shared customers. Ultimately, we sampled businesses that were over the median number of reviews (more than eight) and business pairs in the top 0.1 percent of lift.

In addition to graphing the network, we calculated two statistics of connectivity. The first, degrees, is a count of the number of edges a node has. In this case, the number of other businesses a business is connected to through a shared Yelp customer. Having a high degree count suggests that a business shares its customer base with many other businesses. The second statistics is betweenness. Betweenness counts the number of shortest paths between every other pair of nodes in a network that go through a given node. Having a high betweenness here suggests that a business has high cross-over appeal. People with different tastes all go to business A, for instance.

Prediction

Finally, we wanted to know whether we could predict whether a business has high connectivity (defined by betweenness) from other features, such as number of reviews, average rating, location, hours, or type. To that end, we modeled a random forest, essentially an average of several decision trees, to predict high-connectivity from hundreds of features in the Yelp dataset as well as supplemental information from the Census Bureau.

To make the model, we labeled the 61,000 businesses in the dataset as having high connectivity or not (based on their betweenness score). We then added in Census data which reduced the size of the dataset to ~51,000 since we only had census data for the US businesses. We then applied a Boosted Random Forest algorithm to the data with as analysis that we had done on the dataset for another class indicated this was the best model.

Big data

MapReduce

Our primary use of big data techniques occurred in calculating the association rules. We used EMR on a cluster of Amazon EC2 instances to calculate the support of each business as well as the support confidence, and lift of each business pair. This required two EMR jobs and a local script implementing a heap:

- The first job yields a business id and the number of unique users who reviewed it.

- The second job has three steps. The first yields each business pair that shares at least one Yelp user as a customer. The second yields each business pair that shares at least one Yelp user and the number of shared customers. The third step only involves a mapper. It uses the count from the second step and the output from the first job to yield a business pair and a list of support, confidence, and lift statistics. Ultimately, this job yields more than 48 million pairs.

- The final script limits the business pairs used in the network analysis to those above a certain lift and also applies an absolute numerical cut-off as well. For our network analysis, we wanted to look at businesses in the top 0.1 percent for lift, so we set a lift threshold of 62,178.41.

We used S3 to store our initial data as well as intermediate steps that needed shared access. All of the raw data that we received from Yelp was uploaded to a shared S3 bucket. We also added the business frequency file to the S3 bucket once it was complete so it could be run in the AWS query.

Multiprocessing

In addition to AWS storage and applications, we also used multiprocessing to reduce the time of computations done locally on our own machines. Specifically, we used the multiprocessing package in Python to parallelize the machine learning pipeline that led to our predictive model, this was done in a pooled k-folds function so that it ran each of the random trials of the model separately, all of the other code was written exclusively for the machine learning class.

We did not use any big data techniques for the network analysis. It was conducted in R using the iGraph package.

Results

Association rules

We calculated the lift of all 48 million pairs of businesses that share at least one Yelp user as a customer. We found that the highest lift business pairs are all in Scotland.

Business A	**Business B**	**Location**	**Lift**
Forbidden Planet	Southside Books	EDH	464903.2
Princes Mall	Jack Wills	EDH	474816.9
House of Fraser	Princes Mall	EDH	474816.9

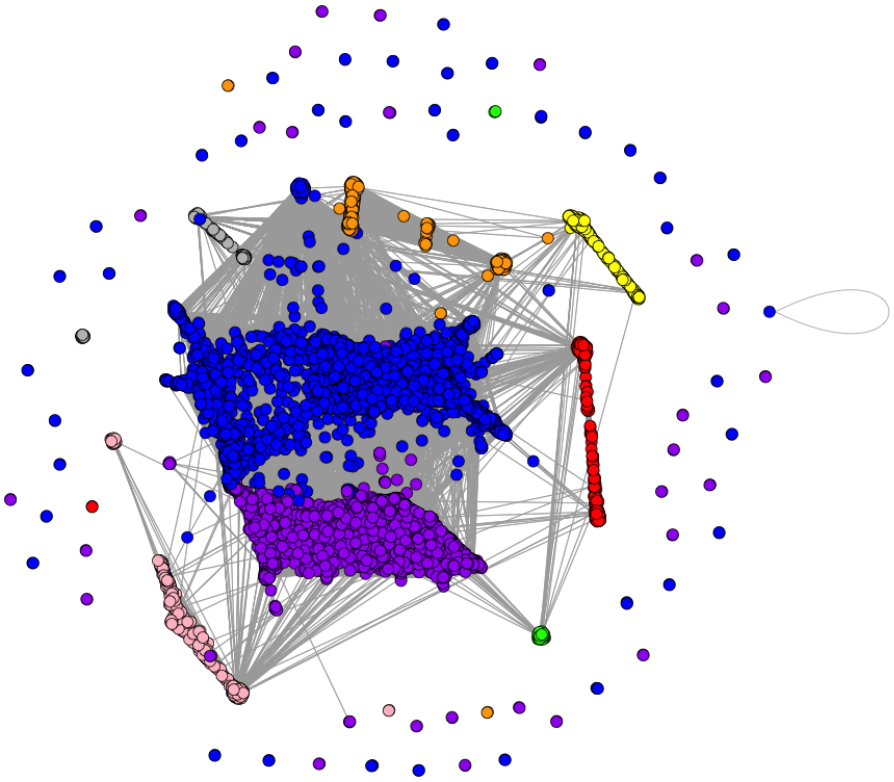
We theorize that there are not many Scottish businesses on Yelp compared to other states in the dataset, but Yelp users who visit one of these Scottish businesses visit all the other ones too.

We also looked at the business pairs with the highest lift in each state as well as the business pairs across state lines with the highest lift. Some of the latter appear to be travel-related (places we expect tourists to go when traveling), but not all.

Business A	**Business B**	**Location**	**Lift**
Carolina Family Restaurant	Air Canada Lounge	NC-QC	82904.5
Walmart	Museum of Edinburgh	SC-EDH	74614.1
Chicks w Spiritual Gifts	Spiritual Connections	AZ-CA	74614.1
BodyN’Sole	Rockstar Snowcones	IL-NV	82904.5

Network analysis

We graphed the network of Yelp businesses in the figure below. Each node is a business with at least eight reviews, and an edge between two businesses means they have a lift in the top 0.1 percent of all business pairs.



We also calculated the degrees and betweenness for each business in this network. We found that the businesses with the highest degrees were similar to the businesses with the highest lift. Again, they are all in Scotland. This seems to support our theory that Scottish businesses have high lift because they share customers.

Business	**State**	**Degrees**
Multrees Walk	EDH	982
Bruntsfield	EDH	962
Bedlam Theatre	EDH	940

Businesses with the highest betweenness all appear to have tourist appeal. Le Mondial De La Biere is an international beer festival in Quebec, and Rascals is a Scottish pub that has been featured on the Food Network.

Business	**State**	**Betweenness**

Cold Stone Creamery	NC	3878903
Le Mondial De La Biere	QC	2851884
Rascals	EDH	2533497

Prediction

Our model identified the following features as meaningful predictors of high-connectivity:

- Certain Census tracts are influential, more so than the attributes of those tracts
- Opening and closing times matter a lot
- Whether the business serves pizza is very predictive

The model has a precision of 0.36, meaning one-third of the businesses we label as highly connected are actually highly connected. It has a recall of 0.46, meaning we correctly label about half of the highly connected businesses in the dataset. While neither of these are especially strong in terms of being certain of the results, it still is much better than random guessing which would get the baseline percentage of around 10%. The overall results are in the machine learning folder, with item by item predictions, feature importance and overall metrics.

Lessons learned and challenges

We learned a lot about how to implement EMR, MRJob, and association rules to help answer our questions. We learned how to set configure options with an MRJob class to pass files to nodes and set parameters from the command line. We needed the former option, in particular, to ensure the nodes in our second EMR job had access to the support statistics calculated in our first EMR job in order to calculate the confidence and lift statistics. Along those lines, we were able to successfully utilize MRStep in our calculations. Applying association rules and interpreting the statistics in the context of our data was also challenging, but a worthwhile exercise.

Probably the biggest challenge we had was with the output of our second EMR job. We wanted to limit the output to the top k business pairs by lift. What we discovered, though, is that we could limit each node's output to the top k business pairs on that node, there was not a way to limit the output across all nodes to the top k business pairs across all nodes. We did not discover a direct way to overcome this challenge. Instead, we created a script to run locally that reduced the output of the second EMR job. Generally, we found debugging and verifying results difficult.