# Yelp final report

Mitsue Iwata, Kyle Magida, Scarlett Swerdlow

# Data set

- Yelp released a sample of its data set as part of the Yelp Dataset Challenge
- 1.6M reviews by 366K users for 61K businesses in ten cities

# Questions

- If you've gone to business A, what is the increased chance you've gone to business B?
- Which businesses are the most connected?
- Can we predict whether a business will have high connectivity from other features?

# Algorithms

- Association rules
- Network analysis
- Boosted random forest

# Association rules

- Used to understand the relationship within data
- People who go to A also go to B
- Key statistics: support, confidence, and lift

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

$$Support\ (A\ \&\ B) = Pr\ (A\ \&\ B) = \frac{\#\ users\ who\ have\ reviewed\ A\ and\ B}{\#\ users}$$

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

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

# Calculation of a-rules / big data

- AWS EMR
- MapReduce to calculate support of every business. Yielded 60,785 pairs.
- Much more complicated MapReduce to calculate confidence and lift of every business pair. Yielded ~48M pairs.

# Findings: Top lift business-pairs

- All in Scotland

| biz_a | biz_b | state | lift |
|---|---|---|---|
| Forbidden Planet | Southside Books | EDH, Scotland | 464903.2 |
| Princes Mall | Jack Wills | EDH, Scotland | 474816.9 |
| House of Fraser | Princes Mall | EDH, Scotland | 474816.9 |

# Findings: Top lift by state

- Seem to make sense

| biz_a | biz_b | state | lift |
|---|---|---|---|
| Vino Lounge | Lyte Lounge & Bistro | AZ | 331618.2 |
| Harris Teeter | Target Stores | NC | 339154.9 |
| Garcia's Pizza In a Pan | CUMTD | IL | 268610.7 |
| Grill Burger Kitchen | Pita Factory | ON, Canada | 186535.2 |
| Cielo | Basislager | BW, Germany | 248713.6 |

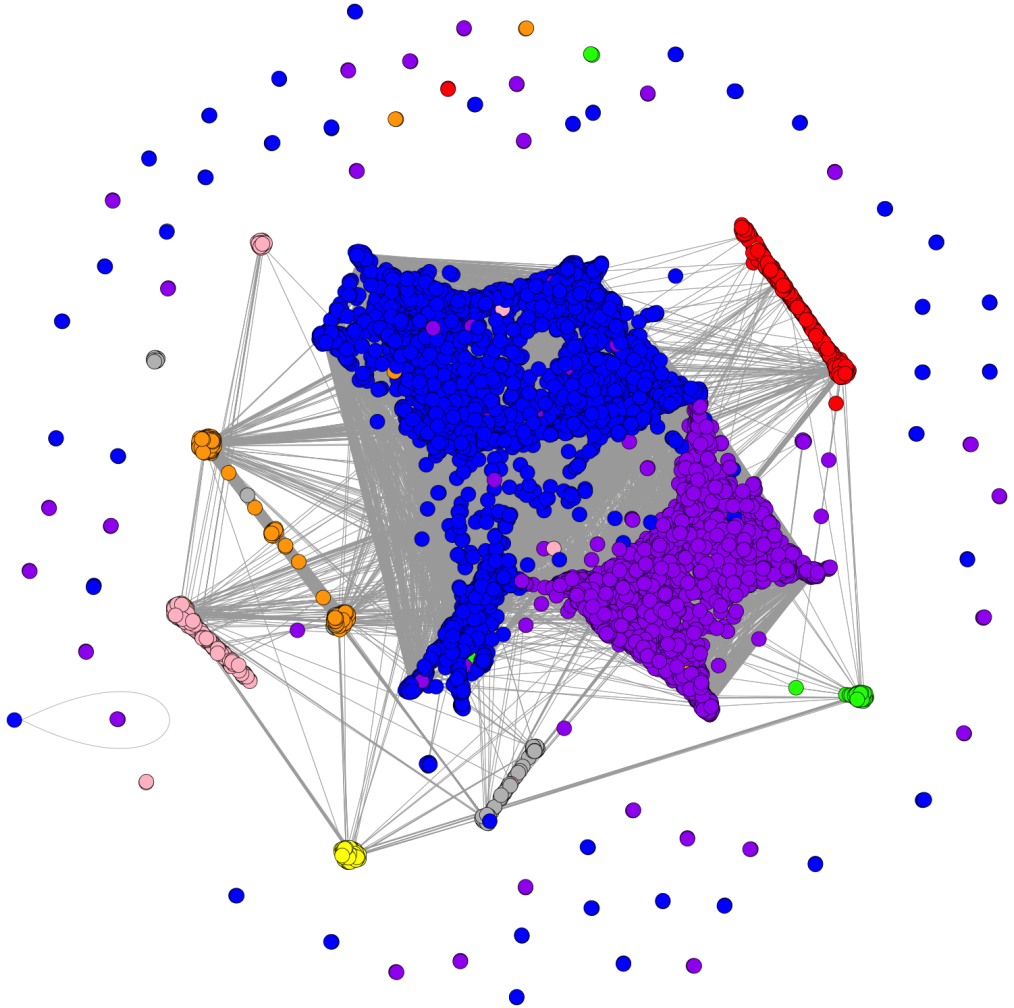
# Findings: Top lift across state lines

- A lot (but not all) are travel related

| biz_a | biz_b | state | lift |
|---|---|---|---|
| Carolina Family Restaurant | Maple Leaf 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

- Network is too big to evaluate in its entirety, but we can use businesses with high lift to approximate the network



# Calculation of connectivity

- R iGraph
- Key statistics are degrees and betweenness.
- Degrees: Biz A shares its customers with many other businesses.
- Betweenness is a measure of crossover appeal. People with different tastes all go to Biz A.

# Findings

| Biz | State | Degrees |
|---|---|---|
| Multrees Walk | EDH, Scotland | 982 |
| Bruntsfield | EDH, Scotland | 962 |
| Bedlam Theatre | EDH, Scotland | 940 |

| Biz | State | Betweenness |
|---|---|---|
| Cold Stone Creamery | NC | 3878903 |
| Le Mondial De La Biere | QC, Canada | 2851884 |
| Rascals | EDH, Scotland | 2533497 |

# Machine learning

- Can we predict whether a business will have high crossover appeal based on its other attributes?
- Census data was included as well so analysis limited to US
- Attempting to parallelize manually

# Findings

- PIZZA
- Best model had .36 precision and .46 recall
- 10% of all business defined as between
- Certain census tracts are influential, more so than attributes of those tracts
- Opening and closing times matter

# Interesting things we learned

- AWS EMR configuration options
- MRJob steps
- Parallelization in Python

# Challenges

## mrjob v0.4.4 documentation



### Table Of Contents



## Troubleshooting

Many things can go wrong in an EMR job, and the system's distributed nature can make it difficult to find the source of a problem. `mrjob` attempts to simplify the debugging process by automatically scanning logs for probable causes of failure. Specifically, it looks at logs relevant to your job for these errors:

# Stuff to come

- Parallelize network analysis and machine learning components
- Winning $5000—Prof. Wach’s share will depend on our grade

# Appendix

Stuff only Prof. Wachs cares about

# MapReduce: Support

- mapper yields business id, 1
- combiner yields business id, sum(counts)
- reducer yields business id, sum(counts)

# MapReduce: Confidence and lift

- First mapper yields user, biz
- First combiner yields user, biz list
- First reducer yields biz pair, 1
- Second mapper yields biz pair, count
- Second combiner yields biz pair, count sum
- Second reducer yields biz pair, count sum
- Final mapper init loads MR Support results
- Final mapper yields biz pair, list with support, confidence, and list stats