Project Presentations 1

CMSC 498J: Social Media Computing

Department of Computer Science
University of Maryland
Spring 2015

Cat 1: Groups G4, G5, G6, G7, G10
Presentation 1

- **G4**: Evolution of Amazon Product Reviews Over Time
  - Cody, Ben, Chris, Matthew

- **G5**: Predicting the Popularity of a Restaurant
  - Hannah, Rattanawalee, Alishba, Julia

- **G6**: Using Tweets to Determine Postitive/Negative Sentiment towards Movies
  - Jack, Paige, Gedalia, Danny

- **G7**: The Social Effect on Foursquare CheckIn
  - Erik, Dan, Ben, Brandon

- **G10**: Using Social Media to Improve Music Recommendations and to Aid Music Discovery
  - Julia, Anthony, Mario, Delante
Evolution of Product Review Data

Team FORCE
Ben, Cody, Chris, Matt
Motivation

Something, somewhere went terribly wrong

~$300,000,000,000 worth of sales were made online in 2014 in US

~$1,250,000,000,000 worth of sales were made online globally
Why Understand Reviewer Networks?

- Online reviews are highly influential in consumer purchasing decisions
- Business owners can enhance the reception of their reviews
- Customers can be more informed by better understanding reviews
What Did We Research?

1) Do review scores become more homogenous over time? (social influence)
2) Do initial review scores differ significantly from later review scores? (average)
3) What’s the best time to release a product (month, time)?
4) Do connected reviewers tend to review similar products overtime? (triadic closure)
Dataset Acquisition and Preparation
Dataset

- For this iteration, we used Electronics Reviews
- 1.2 Million Amazon Electronics Reviews
  - Range from Nov. 1996 - Mar. 2013
  - 50,000 removed for errors:
    - “unknown” Reviewer_ID
    - Invalid timestamps
    - “Bot” users that spammed reviews
- Future iterations - Movies/Patio Furniture
Fields That Interested Us

- **product/productId**: B00006HAXW
- **product/score**: 5.0
- **review/time**: 1042502400
- **review/summary**: Pittsburgh - Home of the OLDIES
- **review/text**: I have all of the doo wop DVD's and this one is as good or better than the 1st ones. Remember once these performers are gone, we'll never get to see them again.
- **Product/title**: Rock Rhythm & Doo Wop: Greatest Early Rock
- **product/price**: unknown
- **review/userId**: A1RSDE90N6RSZF
- **review/profileName**: Joseph M. Kotow
- **review/helpfulness**: 9/9

The “Foci” of our graph

The “People” of our graph
The Data Conundrum

- Data was in a .txt output file that had no ability to be parsed in real time.
- Total data file was >1GB which was very unruly to edit or open (tons of RAM)
- After writing a conversion script to make .csv files, the files still had too many rows to be easily edited by Excel, Access, etc.
The Data Conundrum Fix

- Break all data into multiple .csv files and push to a personally hosted MySQL Server
- Break primary table into secondary tables for accessing important data like:
  - Overall review score for a product
  - Number of reviews for a reviewer
  - Score breakdowns of products (1, 2, 3, 4, 5 star)
- Use SQL to create custom tables/data sets for individual parts of the project
Environment and Tools

Data Management:
● FTP Server - Personal Windows Server
  o Central location for data so that everyone is using the same data sets, updates can be made easier
● MySQL Server - Raspberry Pi 2
  o Data broken across multiple different tables

Trend Analysis:
● Excel (PowerPivot) - Individual Machines
  o Pulled data from SQL queries
Environment and Tools 2

Network Analysis:

- Anaconda/Jupiter Environment
  - Anaconda is a scientific/analytically geared distribution of Python 2.7 made for speed on large data sets
  - Hosted on a remotely/team accessible server
  - Same local network as Raspberry Pi/MySQL server so that work would be split between the two.
Excel (PowerPivot) Trend Analysis
Consistency of Review Score

• Approach:
  o Group reviews into buckets of 10s, 50s, and 100s
    ▪ Order all reviews by product and timestamp
    ▪ For each product 1st review is 1, 2nd review is 2, 100th review is 100…
    ▪ Reviews labeled 1-10 in bucket 1, 11-20 in bucket 2, ect...
  o Plot the standard deviation of review score for each bucket
Standard Deviation Trend Data

Standard Deviation of Review Scores
(1-100 Broken into 10s)
Standard Deviation Trend Data

Standard Deviation of Review Scores
(1-400 Broken into 50s)
Standard Deviation Trend Data

Standard Deviation of Review Scores
(1-800 Broken into 100s)
Consistency Data Analysis

- **Finding:**
  - Standard deviation of review score trends downward

- **Theory:**
  - Reviews become more consistent over time
  - Reviews possibly influenced by previous reviews
  - Reviews possibly close in on true valuation
Trend of Average Review Score

- Approach:
  - Group reviews into buckets of 10s, 50s, and 100s.
  - Plot the average of review score for each bucket
Average Trend Data

Average of Review Scores
(1-100 Broken into 10s)
Average Trend Data

Average of Review Scores
(1-400 Broken into 50s)
Average Trend Data

Average of Review Scores
(1-800 Broken into 100s)
Review Average Data Analysis

● Finding:
  o Average of review score trends upward

● Theory:
  o Products with lots of positive reviews continue to be positively reviewed
  o Products with negative reviews stop being reviewed
  o Products with many reviews are generally positive
Timing Product Release

- Approach
  - Plot the number of reviews, average of review score, and standard deviation of review score by month
Reviews Per Month Data

Total Reviews Per Month (1997-2013)
Average Per Month Data

Average Review Score Per Month (1997-2013)
Standard Deviation Per Month Data

Standard Deviation of Review Score Per Month (1997-2013)
Timing Data Analysis

● Finding:
  ○ January has the most reviews, highest average review score, and lowest standard deviation.

● Theory:
  ○ Products are more likely to be favorably and consistently reviewed if released during the December/January holiday season.
Evolution of Reviewer Network

Treated as a Bipartite Graph:
  Reviewers are Nodes
  Products are Foci

Researched Network Evolution
  Clustering Coefficient (Triadic Closure)
  Cosine Similarity
Evolution of “Triadic” Closure

❖ Segment network into time slices
  ➢ 5 was a good fit {1997-2000, 2000-2003, ...}

Computed bipartite clustering coefficients for select users for each time slice
  ➢ Similar to average neighborhood overlap
  ➢ Computation included older slices network data

❖ Additional Pruning
  ➢ Removed user <= 5, products <= 10 reviews
Clustering Coefficient Contracts To 0.2

BCC of Chunk 1 Users

BCC of Chunk 3 Users
Conclusions

1) Discounting additional product offerings, reviewers who review similar products are more likely to **diverge** in future reviews
   a) Users: reviewer offer diverse perspectives
   b) Business Owners: Predicting reviewers from only the network link structure will fail
Cosine Similarity of Reviewers

- Each reviewer has a list of products they’ve reviewed
- Similarity is determined through comparing the products that reviewers have/have not reviewed
- Method: Cosine Similarity
Cosine Similarity ctd.

- Two Reviewers: R1, R2, Five Products p1-5
- R1 has Reviewed: p1, p4, p5
- R2 has Reviewed: p2, p4, p5

\[ R1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad R2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \]

Cosine Similarity = \[ \frac{R1 \times R2}{\sqrt{(R1^2)(R2^2)}} = \frac{2}{\sqrt{3 \times 3}} = .6666 \]
Methodology

- Split up review data into 5 buckets by time.
- Bucket -> b[‘a’,’b’] = cos_sim(a,b)
- Bucket 2 includes Bucket 1, 3 includes 2, etc.
- Plot each bucket
- Note: Removed data where Cosine Similarity is zero
Cosine Similarity: Results

^Buckets 1,2

^Buckets 2,3
Cosine Similarity Results ctd.

Buckets 3,4

Buckets 4,5
Cosine Similarity Results

- Over time, cosine similarity increases overall within a set of reviewers. Increased density of plot indicates a growth in non-zero similarities between reviewers (At least in this metric)
- Application: Identify and market to highly similar users over-time
Extensions

- Compare and contrast to more datasets
- Perform link-prediction based on similarity of review and helpfulness score
- Further research differences between BCC and cosine similarity
Any Questions?
Tripp Lite cables are the best!
By George from Charlotte on February 19, 2012
Size Name: 25-feet | Color Name: Green | Verified Purchase
Needed to attach a few more devices thru an 8-port switch, and needed both short and medium length ethernet cables. Have used several brands in the past, including Belkin, Tripp Lite and "no-names". Tripp Lite is heavier duty (both the exterior insulation and the internal wires) and the snagless connectors are great quality. Wouldn't use any other brand at this point.

Comment: Was this review helpful to you? Yes No Report abuse

Tripp Lite cables are the best!
By George from Charlotte on February 19, 2012
Size Name: 20-feet | Color Name: Blue | Verified Purchase

Needed to attach a few more devices thru an 8-port switch, and needed both short and medium length ethernet cables. Have used several brands in the past, including Belkin, Tripp Lite and "no-names". Tripp Lite is heavier duty (both the exterior insulation and the internal wires) and the snagless connectors are great quality. Wouldn't use any other brand at this point.

Comment: Was this review helpful to you? Yes No Report abuse

Tripp Lite cables are the best!
By George from Charlotte on February 19, 2012
Size Name: 10-feet | Color Name: Green | Verified Purchase

Needed to attach a few more devices thru an 8-port switch, and needed both short and medium length ethernet cables. Have used several brands in the past, including Belkin, Tripp Lite and "no-names". Tripp Lite is heavier duty (both the exterior insulation and the internal wires) and the snagless connectors are great quality. Wouldn't use any other brand at this point.

Comment: Was this review helpful to you? Yes No Report abuse
Predicting Restaurant Popularity in College Towns

By: Julia Eng, Rattanawalle Surason, Alishba Khawaja, Hannah Pallack
Problem

60% of restaurants fail within the first year\textsuperscript{[1]}

80% of restaurants fail within the first five\textsuperscript{[1]}

Restaurants in areas with high concentrations of college students are more likely to succeed\textsuperscript{[2]}.

- Colleges students frequently go out to eat, so there is a need for restaurants in the area
- But certain restaurants have difficulty staying open
  - College Park examples: Chidogos, Garbanzo Mediterranean, Roti

Goal

To predict whether a restaurant will be “popular” or “unpopular” if opened near a specific college campus.

popular $\geq$ 4 stars

unpopular $\leq$ 2 stars

(5 stars being the max)
Yelp

A social media site where users can rate and review local businesses, including restaurants.
Yelp Data

- Used the Yelp Search API and Business API along with scraping individual restaurants’ profiles to obtain data.
- 283 restaurants within 5 miles of UMD
- Info for each restaurant contains some of the following:
  - Type of food, distance from college, rating, review count, price, reservations, delivery, take out, good for X (parking, kids, groups, etc), has alcohol, has waiter service
Yelp Data Concerns

- Only specific part of population posts on Yelp
- Not every restaurant has info on all features
- Potential information cascades
  - People who visit a restaurant might be biased by other people’s reviews
Yelp Data Concerns Cont.

- Popularity of a restaurant changes over time
  - We obtained the academic dataset from Yelp created in 2012
  - Compared ratings of restaurants in College Park to their ratings from 3 years ago
  - Of the 122 in both datasets, 45 (37%) had the same rating, 60 (50%) differed by 1.0 star or less, and 17 (13%) differed by > 1 star
1. create graphs
2. summary statistic
3. offer prediction functions and examples are:
   - Linear Regression
   - Decision Tree
   - Logistic Regression
Training & Testing

1. Train Data
   - Decision Tree
   - Linear Regression

1. Test Data
   k-Cross Validation
Best Features for UMCP
5 most important features

1. Review count
2. Catering
3. Price
4. Delivery
5. Distance from college
Linear Regression

Formula is \( Y \approx \beta_0 + (\beta_1)(X_1) + (\beta_2)(X_2) + \ldots + (\beta_n)(X_n) \)

Pick top 5 variables

- \( Y = \) Rating
- \( X_1 = \) Caters
- \( X_2 = \) Price
- \( X_3 = \) Delivery
- \( X_4 = \) distance_from_college
- \( X_5 = \) BikeParking
**Residual values**

Residual values = true y - predicted y

- Residual values can be negative or positive.
- If *residual* value is *zero*, it means the prediction is corrected.
- If *residual* is *far from zero*, it means the prediction is very off.
### Result From Linear Regression

- Split data into 5 section.
- Train \( \frac{4}{5} \), test \( \frac{1}{5} \) and run it for 5 times:

**Residual summary when trying to predict #rating are:**

<table>
<thead>
<tr>
<th>Min.</th>
<th>Median</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.860818</td>
<td>0.0771</td>
<td>1.664812</td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1.664812
Result from Decision Tree

- Split data into 5 section.
- Train $\frac{4}{5}$, test $\frac{1}{5}$ and run it for 5 times:

**Residual Summary**

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>Median</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg.</td>
<td>-1.6502</td>
<td>0.034324</td>
<td>1.3234</td>
</tr>
</tbody>
</table>
Homophily among ratings

Busboys and Poets

Marathon Deli

Krazi Kebob

Average rating: ★★★★★

High review count

Close to campus

Price: Under $10

Does not deliver

Caters

Same rating as

Same rating as

Same rating as
Homophily among ratings

- **Dennys**
  - Same rating as **Moose Creek Steakhouse**
  - Same rating as **Tara Thai**

- **Average rating:** 2 stars
  - High review count
  - Close to campus
  - Price: Under $10
  - Does not deliver
  - Caters

- **Moose Creek Steakhouse**
  - Same rating as **Dennys**
  - Same rating as **Tara Thai**

- **Tara Thai**
  - Same rating as **Dennys**
  - Same rating as **Moose Creek Steakhouse**

**Steakhouse**

**Tara Thai**

**Moose Creek Steakhouse**
Yelp Reviewers’ Comments

- Considered the reviews of the UMD data
- Recurring phrases might correspond with the restaurant’s rating
- Scraped all reviews & collected 20 most common bigrams for each restaurant
Yelp Reviewers’ Comments

- Graphed terms that corresponded to 5+ restaurants

- **Findings**: Most people talked about what they ate
Food Poisoning, 2nd Time Around. I ordered the Shrimp & Black Bean sauce on Sunday, May 6th, 2012 around 2:30pm and they sent me this meal that tasted fine initially but I found myself throwing up for 24 Hours. Poison me once, shame on you. Poison me twice, I'll report you and blast your restaurant. If you order from here, the risk is yours.
mexican food

Chipotle

Taqueria El Mexicano

Alamo Mexican Restaurant

Azteca Restaurant Cantina

La Sirenita Restaurant

Qdoba Mexican Grill

Tacos Cinco de Mayo

La Fondita

Average rating: ⭐⭐⭐⭐️
## Bigram findings

<table>
<thead>
<tr>
<th>Phrase</th>
<th># Restaurants</th>
<th>Restaurant Avg. Rating (stars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“great food”</td>
<td>9</td>
<td>3.8</td>
</tr>
<tr>
<td>“indian food”</td>
<td>8</td>
<td>3.8</td>
</tr>
<tr>
<td>“wait staff”</td>
<td>8</td>
<td>3.3</td>
</tr>
<tr>
<td>“happy hour”</td>
<td>17</td>
<td>3.3</td>
</tr>
<tr>
<td>“fast food”</td>
<td>21</td>
<td>3.3</td>
</tr>
<tr>
<td>“chinese food”</td>
<td>23</td>
<td>2.8</td>
</tr>
<tr>
<td>“delivery guy”</td>
<td>5</td>
<td>2.2</td>
</tr>
</tbody>
</table>
10 college campuses
Some of which were:
- Duke University
- Purdue University
- Johns Hopkin’s University
- University of Michigan
- UC Berkeley
- University of Montana
UMD
BU
Duke
JHU
Northern Iowa
Purdue
Rice
RIT
U of Michigan
Berkeley
U of Montana

Has TV
Full Bar
Delivery
Caters
Good for Groups
Waiter Service
Quiet
Takes Reservations
Bike Parking
Rich Get Richer Phenomenon

- Restaurant’s rating is proportional to the number of reviews
- Ones that are around for a longer period of time will naturally have more reviewers and a higher average of “stars”.
Rating and Reviewers Pattern

Restaurants near University of Maryland College Park
Rating and Reviewers Pattern

University of Berkeley

rating

review_count
Future Applications

● Web application that predicts success of restaurant before it opens
● Enter:
  ○ name of college nearby
  ○ distance from college
  ○ price range
  ○ features important to restaurants in that area
● Return probability of “popularity”
Summary

- Update model for best predictions
- Common phrases indicate users’ attitudes toward restaurants
- Best features for different campuses may vary.
  - 3 features are universal: Price, Distance, Review Count
- Homophily among ratings
- Rich get richer phenomenon
Twitter Movie Sentiment

Gedalia Kott | Jack Lotkowski | Paige Nelson | Danny Shaw
Problem

• Where do we get our reviews about movies?
  • Rotten Tomatoes- only critics
  • IMDB- only active members with accounts
  • Word of Mouth- small pool of people
• There’s a near endless amount of tweets where people give their movie opinions!
Proposed Project

Use the mass amount of data on Twitter to create a review system based on the general opinion of Twitter users.

85% of Twitter users liked Boyhood.
Process

• Collect Tweets about various movies
• Determine if each tweet is a positive, negative, or neutral comment
• Determine overall percentage positive reviews based on tweets
• Compare this percent to review sources like IMDB and Rotten Tomatoes
Data Collection

Rotten Tomatoes

IMDb

Twitter

King of Diamonds @KOD26M · 13m
Nathan Forster @Tolkien_Nerd · 7h

King of Diamonds @KOD26M · 13m
Nathan Forster @Tolkien_Nerd · 7h
I forgot Movie 43 existed. *retch*

5:47 AM - 3 May 2015 · Details
Data Sets and Sources

● 100 Movie Titles and Ratings
  - IMDb and Rotten Tomatoes
    ● Bad
    ● Good
    ● Recent

● 1,000 Tweets per Movie
  - Twitter API

● Total:
  - 300 movies
  - 300,000 Tweets
Getting Movie Titles and Ratings

- Scraped HTML pages from Rotten Tomatoes and IMDb
  - Simple, well-structured data
  - Lots of user ratings per movie
  - Easy to parse data with DOM manipulation libraries (jQuery)
  - Organized pages by Bad, Recent, Good
    - Matches our required data sets exactly
Twitter API

● Using tweepy python module
  – Great client, not the best documentation...

● Options
  – REST API
    ● Collection Rate: High
  – Streaming API
    ● Collection Rate: Depends on the Movie/Situation
      – Bottom line: Not enough data in time
Some Technical Details and Lessons Learned..

- Rate Limits – Collection is time consuming
  - 180 (user auth), 450 (app auth) Requests / 15-min window
  - 10 Requests for 100 tweets per movie
    - Paginated results

- Search Query Parameter
  - Trade-offs: Getting tweets about the movie without the junk

- tweepy Features
  - Wait and block on rate limit, error handling, keeping track of tweets

- Data Storage
  - Data needs change over time — save the tweets to a database!

- Other parameters: count, lang, result_type
  - More trade-offs...
Training Data

1. Curated dataset from Kaggle
2. Movie reviews → tweets
3. Wrangling
4. Rotten Tomatoes and ImDB
Classification

1. MetaMind (… Vowpal Wabbit)
2. 5 v 3 class classification
3. Output as JSON

```
[
    {u'user_value': u"Furious7 was the worst movie I've ever seen. Period.", u'probability': 0.819143650919281, u'label': u'1'},
    {u'user_value': u'I loved Skyfall. Brilliant!', u'probability': 0.9619668949957287, u'label': u'5'}
]
```
Taking it Further

1) http://leonardowater.herokuapp.com/twitter-sentiment

2) SentiWordNet
The Social Effect On Foursquare Check-In

cmsc 498j, spring 2015

Erik Koebke, Ben Summers, Brandon Nguyen, Dan Ochieng
What is Foursquare?

Foursquare is a local search and discovery service mobile app which provides a personalized local search experience for its users. The app takes into account where users go, what places they like and advice from other trusted users to provide personalized recommendations of the best places to go around a user’s current location.
The Foursquare Dataset

- Online Dataset holding 5 charts: checkins, users, venues, social-graph, ratings.
  - Checkins: User ID, location, time (1,000,000+)
  - Users: ID, location (2,000,000+)
  - Venues: ID, location (1,000,000+)
  - Social Graph: 1-1 connections of User IDs
  - Ratings: User ID, Venue Id, Rating (1-5)
Our Goals For This Project

- Identifying the role that social connection plays in users checking in to new venues.
- Identifying the role that social connection plays in the ratings of Foursquare venues.
The Foursquare Network

- **Two Types of Nodes**
  - Users: The people who download and use the Foursquare application.
  - Venues: Different locations of the world that users visit and “check-in” to.

- **Social-Affiliation Network**
  - A network that contains both original nodes and contexts.
    - In the case of Foursquare, original nodes are the users that download the application and contexts are the venues that these users “check-in” to.
    - Venues give context to link formation because of their role as the focal point of interaction within the network.
The Foursquare Network (Cont’d)

User-to-User connection

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>8</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>U2</td>
<td>2</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>U3</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

check in total
rating(1-5)
Mechanisms For Link Formation

- Triadic Closure
  - Two user’s of the network who share a mutual friend are more likely to form a link.

- Focal Closure
  - Two user’s of the network who share a mutual foci are more likely to form a link.

- Membership Closure
  - A user who is linked to another user who is linked to a foci is more likely to link to that specific foci.
Mechanisms For Link Formation (Cont’d)

- Focal Closure and Membership Closure are examples of the Foursquare network exhibiting Homophily.

- Two mechanisms for link formation:
  - Selection (Focal Closure) - selecting friends with similar characteristics
  - Social Influence (Membership Closure) - modifying behaviors to make them closer to behaviors of friends.
Homophily In The Foursquare Network

Selection:
- Two user’s who attend the same venue are more likely to form a connection.

Social Influence:
- A user who has a mutual friend who attended a specific venue is more likely to attend that specific venue.
Mechanisms For Link Formation (Cont’d)

Diane

Bob

Triadic

Joey

Membership

Verizon Center

Focal

Chipotle

Jaclyn
Measuring Homophily In The Foursquare Network

- We measured the degree of selection in the Foursquare network by calculating the % of users that attended the same venues are connected in the social graph.

- We measured the degree of social influence in the Foursquare network by calculating the % of users that check-in to a new venue that has already been attended by a connection.
Homophily Measurement Results

- **Degree of selection**: 34% of users who have attended a venue are connected in the social graph.
- **Degree of social influence**: 38% of the time, users checked in to a venue that had already been attended by a connected user.
Limitations In Measuring Network Homophily

- We do not know when user connections were made.
  - In calculating selection and social influence, we do not know whether a connection was made before or after a user attends a venue.
  - In further experimentation, we would request to know this information in order to more accurately represent the homophily of the network.
On The Other Hand...

Location appears to play a major role in where user’s check-in as well.
Venue Graph

- Another way to view Network
- Each node is a venue; contains: vid(venue id), ckn(checkin number insofar), sts(start checkin timestamp), ets(end checkin timestamp), lat(latitude), lng(longitude), rating
- Edges: Transactions between Venues (Users moving from one venue to another within 2 hour time span)
- Properties for Edge Creation:
  - spatial distance: places that are nearby are more likely to be connected
  - temporal distance: places that can be reached in a short time are more likely to be connected
  - affective/social distance: places that belong to a common lifestyle are more likely to be connected
Venue Graph

- Analysis of Venue graph showed that node degree and check in count distribution follow power-law distribution ($\alpha = 1.33$)
- Skewed Bowtie structure with large out-component and SCC ratio, likely due to testing in a dense city, with small location differences.
- SCC ratio: 53.1%, out-component: 46.65%
- Large amount of locations that become “endpoint” of user traveling (Living communities, work)

-Li, Huimin “Structural and Dynamic Analysis of Foursquare Network”
Correlation Between Social Connections And Venue Ratings

- Foursquare gives users the ability to rate venues on a scale of 1-5.
- In our representation of the network, the rating that a user gives a venue represents the strength of an edge that connects that user to the venue.
- We are interested in the idea that a user’s social connections could influence the ratings that they give to venues.
Measuring Social Influence On Venue Rating

- On the scale of 1-5, we defined the following:
  - Positive rating (+): rating > 3.0
  - Neutral rating: rating == 3.0
  - Negative rating: rating < 3.0
- It may be possible that social influence works differently for a positive rating vs. a negative rating.
- We analyzed venues with a positive average rating (+) separately from venues with a negative average rating (-).
Structural Balance from Positive/Negative Connections

- Structural Balance in Social-Affiliation network, between users and venues
- Balanced Network shows that user ratings align with user connections
Measuring Social Influence on Venue Rating (Cont’d)

We aimed to answer two questions:
- What is the likelihood an individual gives a venue a positive rating, given that their social connections in aggregate have given that venue a positive rating?
- What is the likelihood an individual gives a venue a negative rating, given that their social connections in aggregate have given that venue a negative rating?
Social Influence On Venue Ratings

Results

- 68% of connected users had similar positive ratings at their common venues making the likelihood that an individual will give that venue a positive rating based on their connections rating fairly high.

- 74% of connected users had similar negative ratings at their common venues making the likelihood that an individual will give that venue a negative rating based on their connections very high.
Limitations In Measuring The Role Of Social Influence On Venue Ratings

- We do not know what the average rating of a venue for a user’s connections was at the time the user rated the venue.
  - So, we cannot be certain of the exact role the connections played in the user’s rating.
- In further experimentation, we would request that we be provided with a snapshot of a user’s connections average rating of the venue at the time a user rated the venue.
Social Connection (Local) Impact On Average Venue Rating (Global)

- Social connection is a local measure that has a major impact on the global measure that is average venue rating.
- The larger amount of individuals that choose a rating similar to their social connections, the greater the global impact social connection has on venue rating.
In order to show this, we decided to analyze the top 5 most highly rated venues and the top 5 least highly rated venues.

- **Top 5 most highly rated**
  - Determine what % of positive raters of a specific venue have a connection with another positive rater of that venue.

- **Top 5 least highly rated**
  - Determine what % of negative raters of a specific venue have a connection with another negative rater of that venue.
Social Connection (Local) Impact On Average Venue Rating (Global) Results

- 48% of positive raters of a specific venue have a connection with another positive rater of that venue.
- 52% of negative raters of a specific venue have a connection with another negative rater of that venue.


Using Social Media as a Predictor for Music Recommendations

Julia Narakornpichit, Anthony Mascone, Mario Finelli, Delante Desouza
The Problem...

Music recommendations are often inaccurate

Relevant titles are often omitted from recommendations

Social Media Impact?
The Relevance...

Why should we care?

Multimedia is integrated in daily life

Multimedia follows a flow of information
- understanding this flow is crucial
The Purpose...

...to answer the following questions:

Why are music recommendations often inaccurate?

How can social media influence these recommendations?

Are these influences positive or negative?

How can music recommendations be improved?
The Method…

Using the Last FM dataset

Determine relational effectivity of songs listed in the dataset

Provides the dataset
The Method...

Using the Twitter API

Determine how tweets affect the relationship between different songs

Determine how tweets are related when describing the same song

Determine if a user likes one song, if they like a song that is defined as similar by LastFM

Provides the platform
Together…

Last FM and Twitter help provide the infrastructure necessary to tackle our problem

Use Twitter to verify the relationships made using Last.FM and to see if Twitter can be used to make effective connections between two songs
The Data...
The Scripts...

Preprocesses dataset

Analyzes dataset against Twitter
The Results...

Restrictive queries

Limited twitter dataset

Small subset of lastfm dataset
The Results...

Many “radio” users

sometimes few tweets for particular songs

sometimes many tweets but users rarely tweeted about the same related songs

not always tweet about music (wedding in the office)
The Outcome...

Inclusive results

Natural Language Processing is difficult
-language statements are vague

Are Last FM and Twitter ideal resources?
Where to go?

Expand the dataset (entire last fm dataset -- expensive!)

Potentially look for other datasets (easier to track number of listens and relations between songs)