Project 2- Hive and Visualization

Sahil Puri

Sethuraman Sundaraman

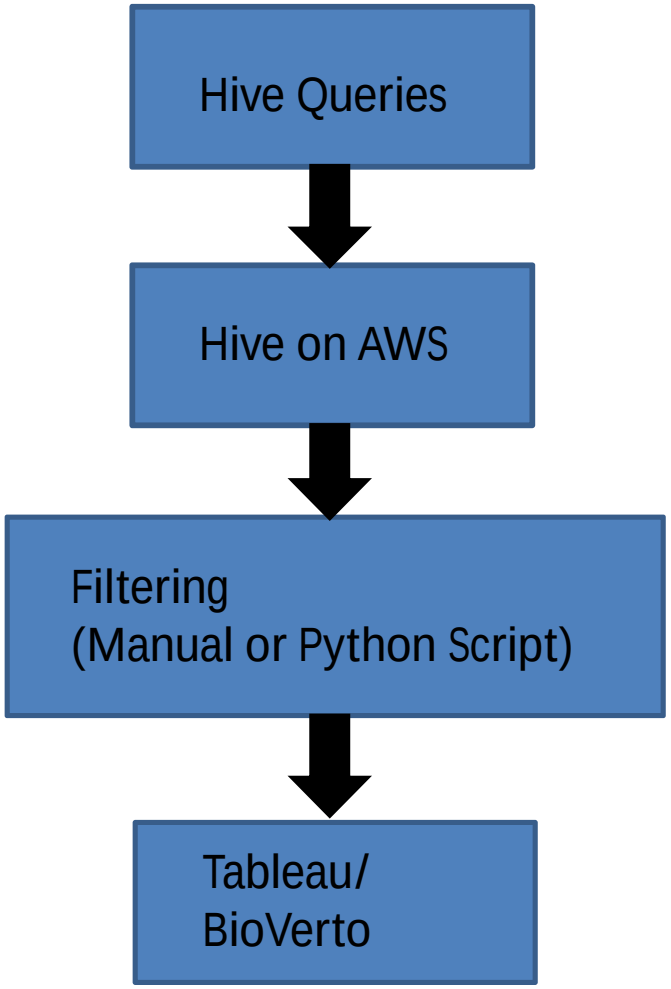
Data Sets

- Netflix Dataset
- Enron Dataset

Work Distribution

- Sahil Puri
  - Netflix Dataset, Analysis and Visualization
  - Attempt on Association Rule Mining
- Sethuraman Sundaraman
  - Enron Dataset, Analysis and Visualization
  - Graph Visualization

Data Processing Pipeline

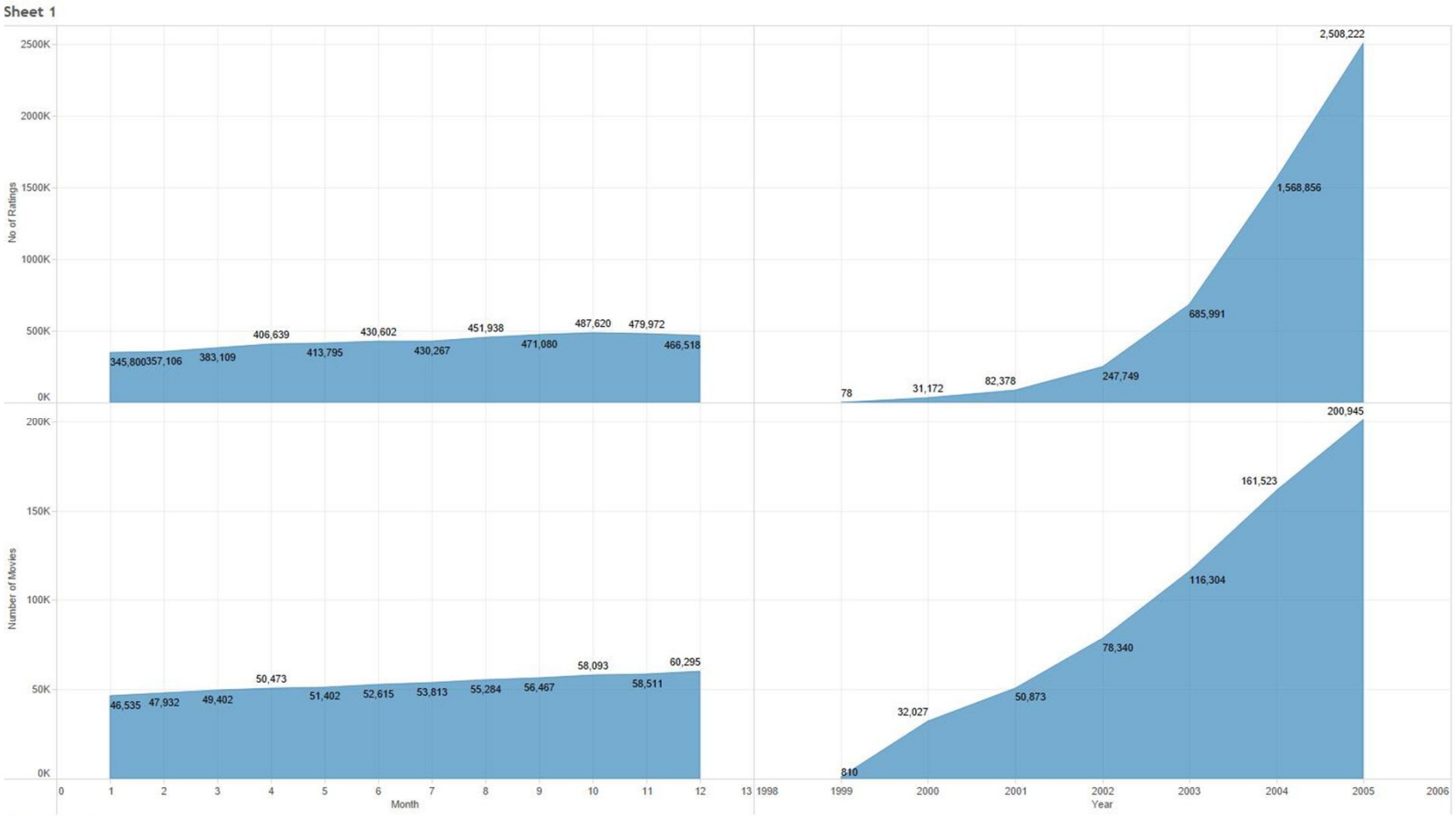


NETFLIX DATASET

Hypothesis 1

- The number of reviews and movies will increase as the internet gets popular, also holiday seasons will have more active users.

Visualization



The plots of sum of No of Ratings and sum of Number of Movies for Month and Year.

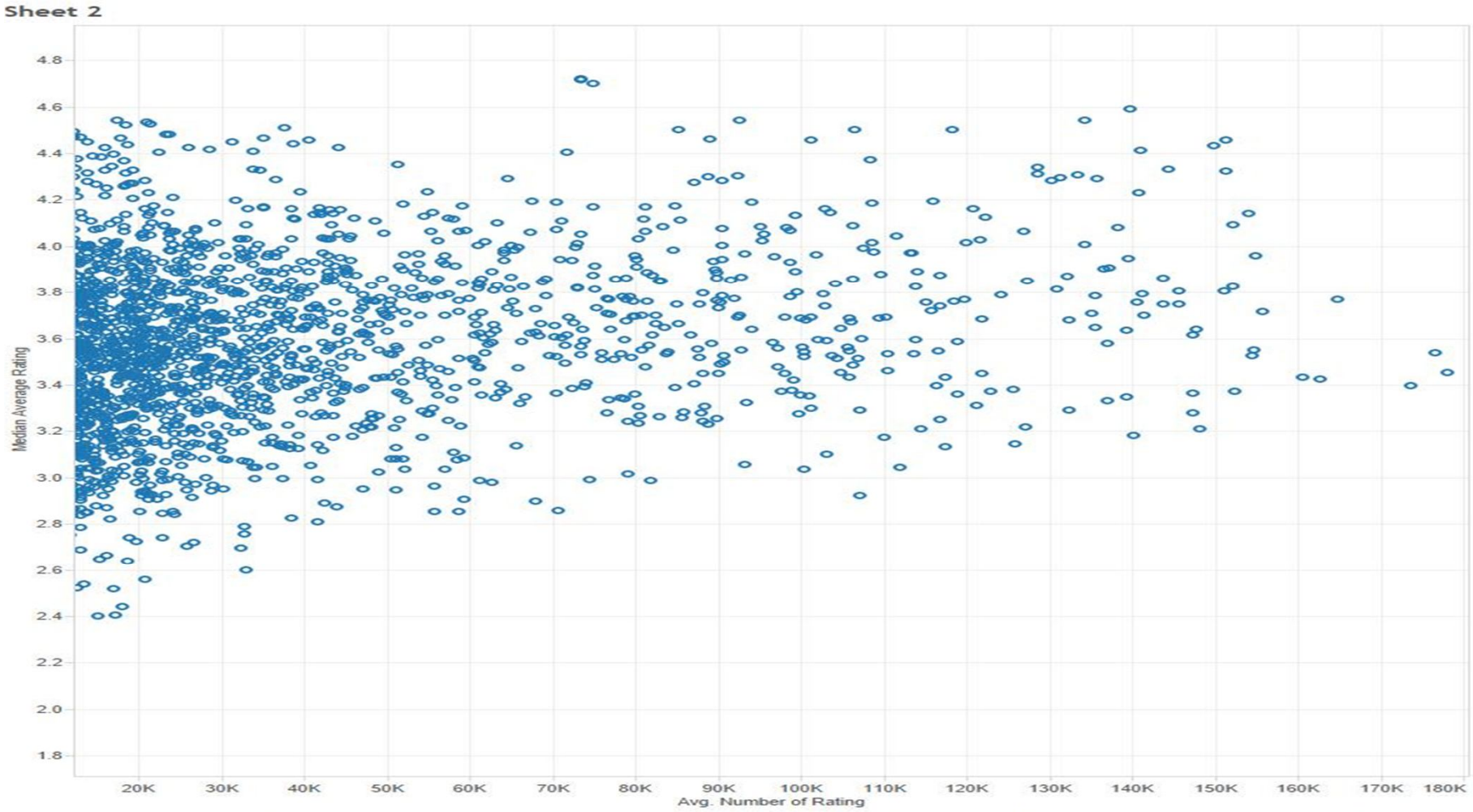
Results

- The number of ratings and movies increase as the years progresses from 2001 to 2005.
- Similarly both the dimensions are high is October to December which is has popular holidays like Thanksgiving and Christmas

Hypothesis 2

- The best movie is a movie with the highest average rating.
- A popular movie is a movie that receives the most reviews.
- Number of users who giving ratings will mostly be comparatively less as it is difficult to receive opinions from customers.

Visualization



Average of Number of Rating vs. median of Average Rating. Details are shown for Name. The data is filtered on sum of Number of Rating, which keeps non-Null values only.

Results

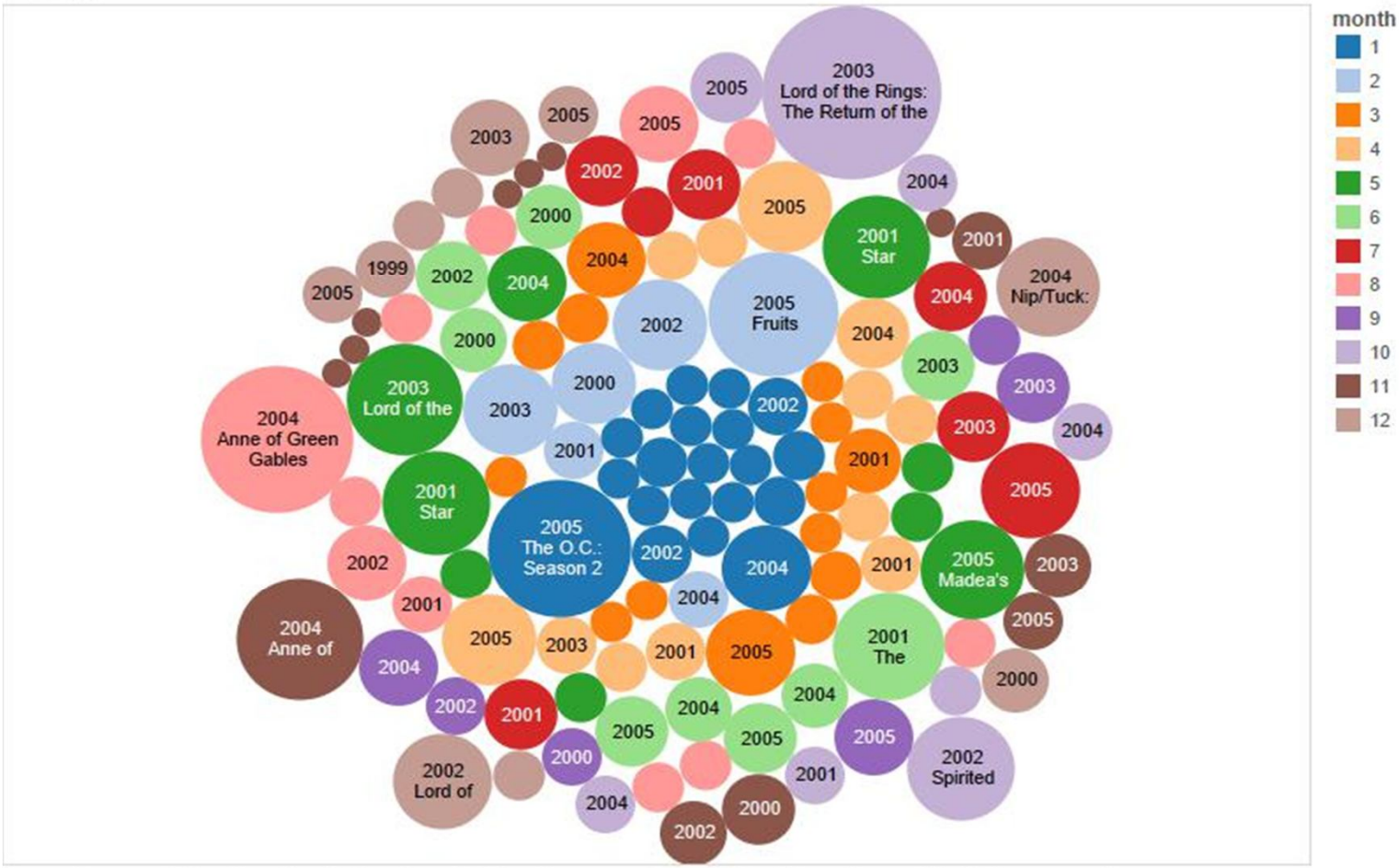
- Bigger cluster of number of ratings are seen towards the left of the graph.
- Thus fewer movies are evaluated by more people.
- The most popular and critically acclaimed movie must be seen on the upper right quadrant.
- The shawshank redemption well known popular and hit movie is in the upper right corner.

Hypothesis 3

- If the highest rated film of every month, of every year is listed, a potential movie which will be highest ranked movie of Netflix of all time and be identified.

Visualization

Sheet 1



Year and name. Color shows details about month. Size shows sum of popularity. The marks are labeled by year and name.

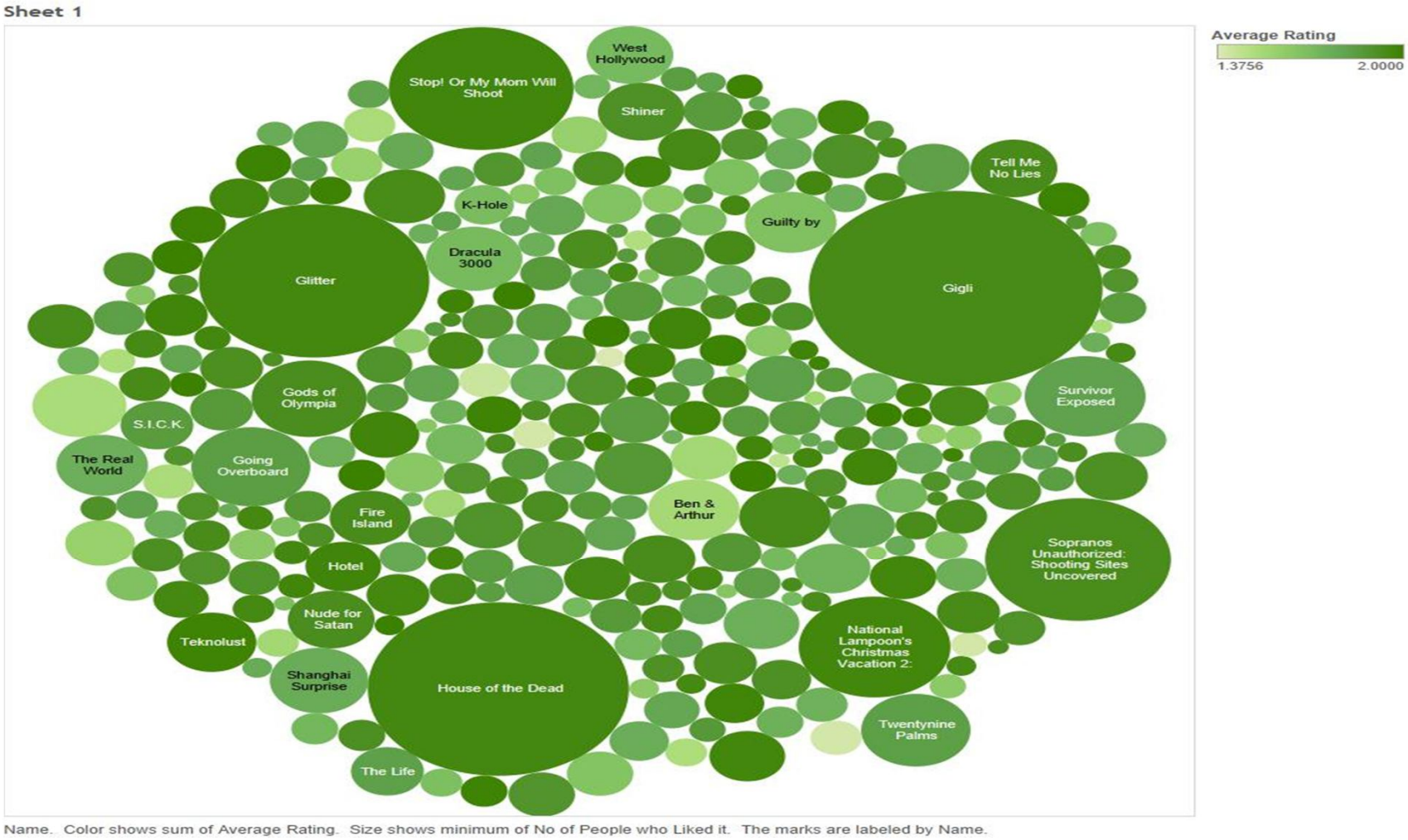
Results

- Colors represent Months and size the popularity of the movie.
- We find that The Lord of the Rings: The Return of the King is seen in multiple months being rated as the best rated movie (ties are broken based on popularity).
- This movie is #9 according to the website http://www.100thingsilearned.com/netflix.php.
- Similarly we see Annie Hall which ranks #142
- Note we have data only till 2005.

Hypothesis 4

- Hated movies are the ones which score an average of less than 2.
- Loved movies will score a 5.
- Most of the hated movies will not have many customers who loved the movie.

Visualization



Results

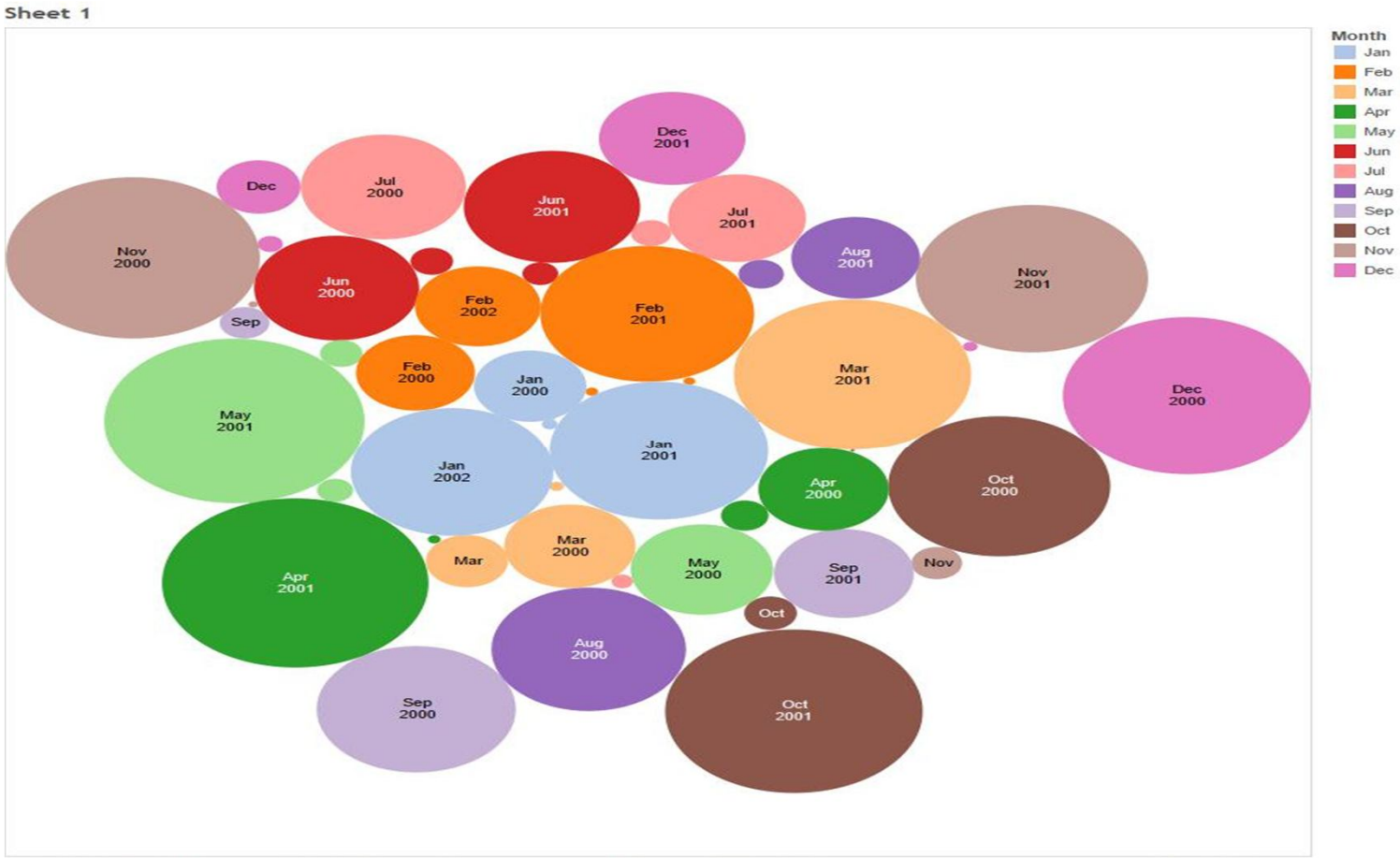
- The brightness of the colors represents relative average scores.
- i.e. the darkest colors score closer to 2 and lighter closer to 0.
- The size of the bubble is the people who love the film.
- We see all larger bubbles are dark green. Lighter color movies are loved by fewer people.

ENRON DATASET

Hypothesis 1

- The number of communications surges when the time approaches the scandal.

Visualization



Month and Year. Color shows details about Month. Size shows sum of Total Number of Communications. The marks are labeled by Month and Year. The view is filtered on Month, which excludes Null.

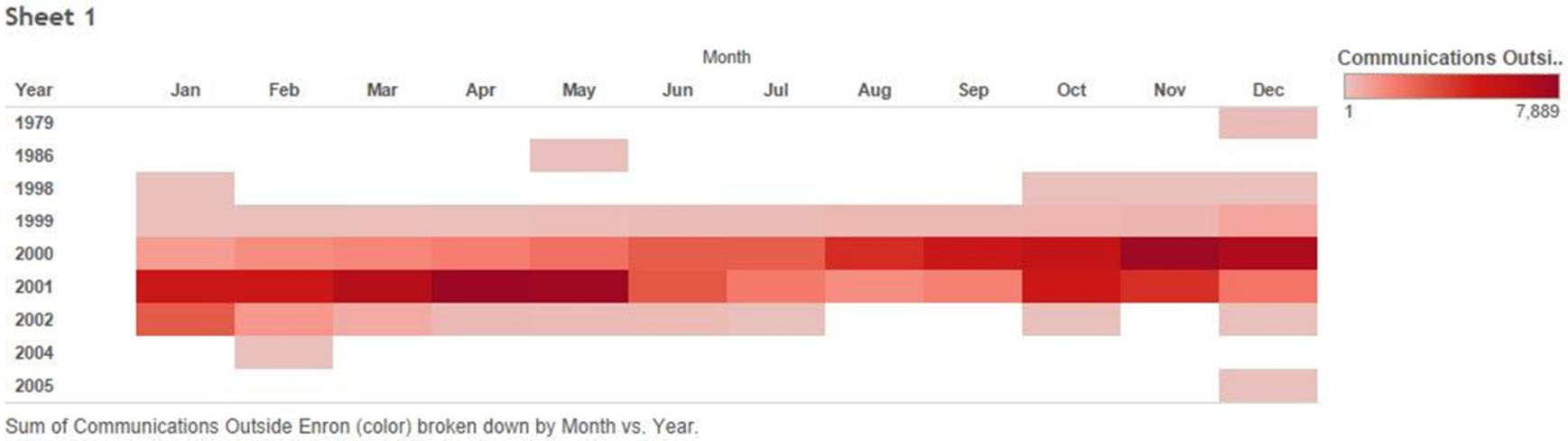
Results

- The color of the bubble represents the month.
- The year is given as text.
- Size of the bubble is the number of emails exchanged in that month.
- We can Conclude there is high activity is April of 2001.

Hypothesis 2

- The number of communications specifically outside enron will increase as the date of scandal approaches

Visualization



Sum of Communications Outside Enron (color) broken down by Month vs. Year.

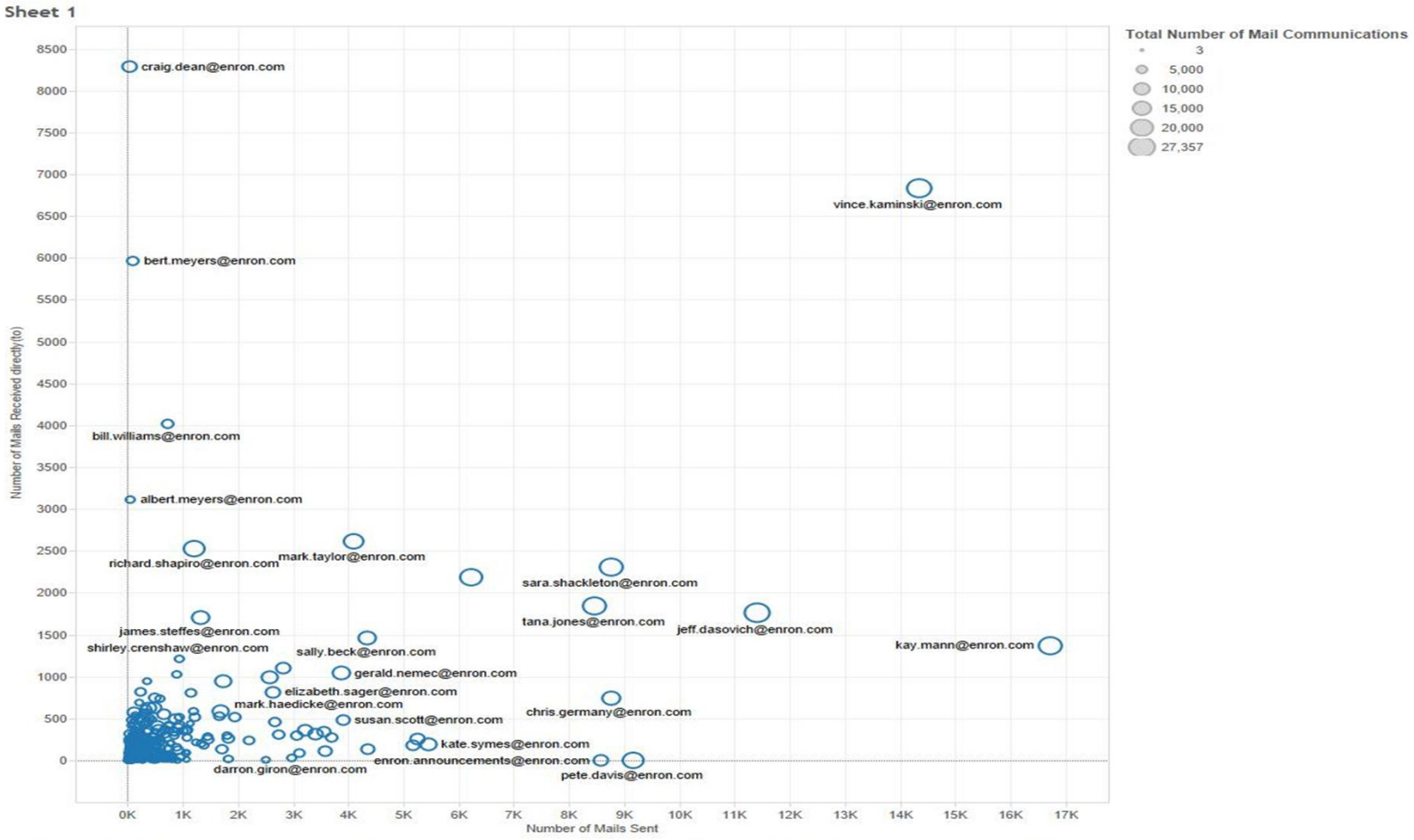
Results

- From the heat graph we can conclude that April 2001 has highest numbers of emails leaving Enron.
- From the above two hypothesis the scandal can be placed in the 2001.

Hypothesis 3

- Degree Centrality: The person will send most number of emails and will receive most mails addressed to them will be an important person in the enron Company.

Visualization



Sum of Number of Mails Sent vs. sum of Number of Mails Received directly(to). Size shows sum of Total Number of Mail Communications. The marks are labeled by Email Id.

Results

- In a graph of number of emails sent versus received people in the upper right quadrant are important.
- Two such important people are Vince Kriminski and Sara Shackeleton.
- Vince Kriminski was the Managing Director of research who flagged the financial Malpractices of the company.
- Sara Shackeleton was the vice president of the company.

Hypothesis 4

- In 2001, if we visualize a graph of every communication between employees, an important person will be centrally connected.

Visualization

Results

- The graph was generated by BioVerto.org (Developed by Karthik Chivukula, Dr. Dobra,).
- The edge represents communication between two people.
- Color signifies degree and size is betweeness (representing the closeness to other nodes).
- Thus Jeff Dasovich represents an important personality in 2001(Big Red Node), indecently he was the Government Reprehensive of Enron.

PROBLEMS FACED AND LESSONS LEARNT

Problem Faced and Possible Solutions

- Absence of sub-queries
  - Use of sub queries was restricted to ‘from’ clause.
  - Restricted support in ‘where’ clause with operators like EXISTS only.
  - Workaround: Use inner Joins
- Absence of set operations like INTERSECTION, UNION and EXCEPT
  - Workarounds possible using inner and outer joins.

Problem Faced and Possible Solutions

- Problems with JOINS on the version in hive installed on AWS
  - Some Joins(self joins) did not work. The query was stuck in reduce job.
  - Others worked on re provisioning the machine
- https://issues.apache.org/jira/browse/HIVE-4222
- https://issues.apache.org/jira/browse/HIVE-3739

Lessons Learnt

- The previously listed points prevented us from implementing Association Rule Mining
- Hive is missing some of the sophistication of SQL.
- Inner joins serve as workarounds for implementing missing functionalities.
- Absence of Procedures, triggers and views being rudimentary, Hive can operate on non real time data.
- Thus hive(true to its definitions), should be used for data ware housing.