Introduction To Data Science
Project - II

Analysis of

- Enron E-mail Data-set
- Netflix Data-set

Atul Garg (94432505)
Ishadutta Yadav (54931916)

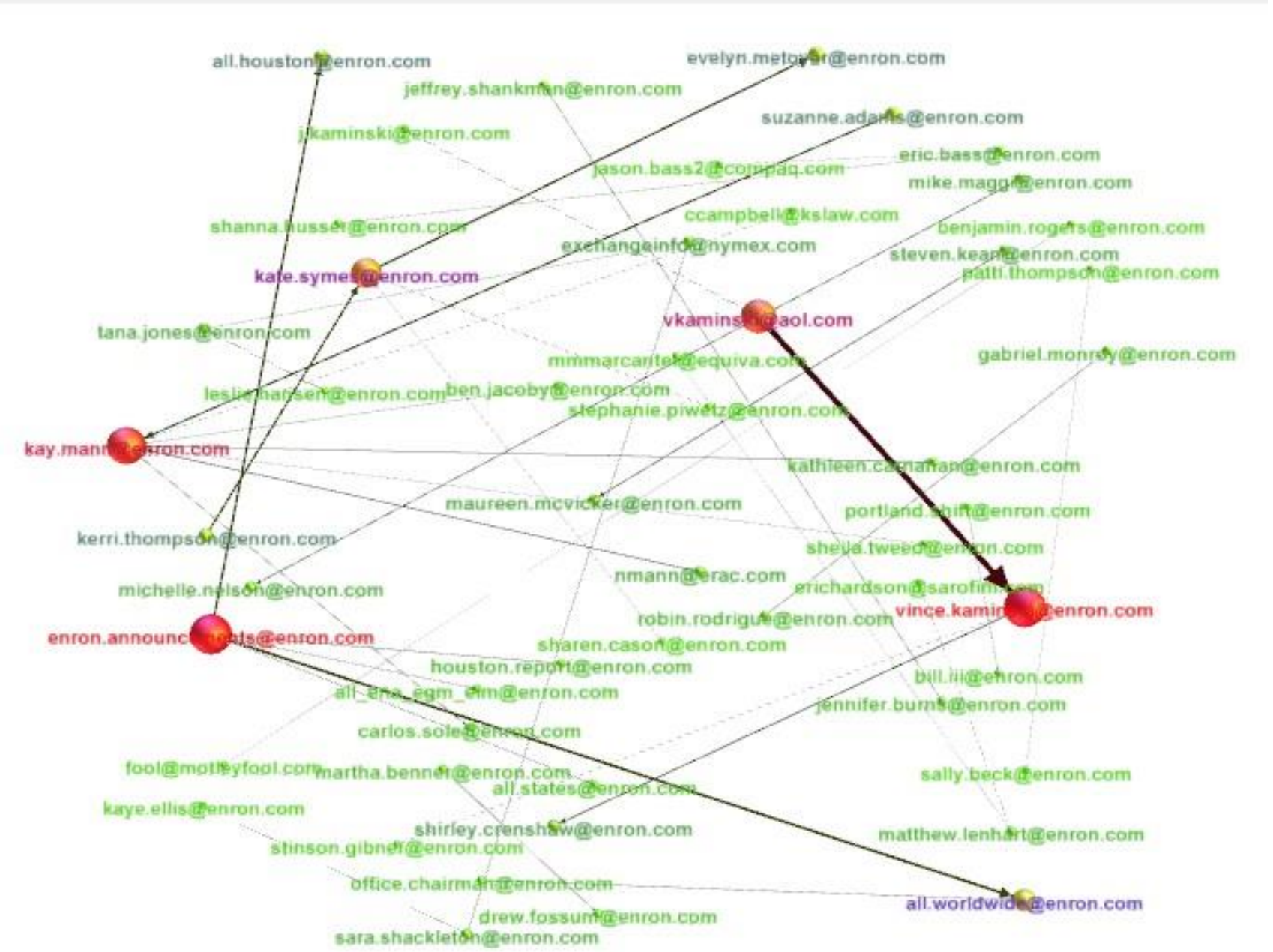
Hive

- What is Hive? A dataware house to store structured data on hadoop file system.
- Provides efficient queries by executing hadoop map-reduce plans.
- Helps for Analysis of data both for engineering and non-engineering people.
- Supports SQL based queries.

Enron E-mail Data-set: Background

- The Enron Corpus is a large database of emails of the Enron Corporation and acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse.[1] A copy of the database was subsequently purchased for $10,000 by Andrew McCallum, a computer scientist at the University of Massachusetts Amherst.[2]
- The Enron scandal, revealed in October 2001, eventually led to the bankruptcy of the Enron Corporation, an American energy company which was found in 1985 by Kenneth Lay. Several years later, when Jeffrey Skilling was hired, he developed a staff of executives that, by the use of accounting loopholes, special purpose entities, and poor financial reporting, were able to hide billions of dollars in debt from failed deals and projects.

Correspondence Network at Enron



* Graph Generated using Gephi[3]

Top People at Enron As per correspondence

- Vincent Kaminski

  Managing Director for research

- C Kay Mann

  Assistant General Councel

- Suzanne Adams

  Legal Assistant

- Evelyn Metoyer

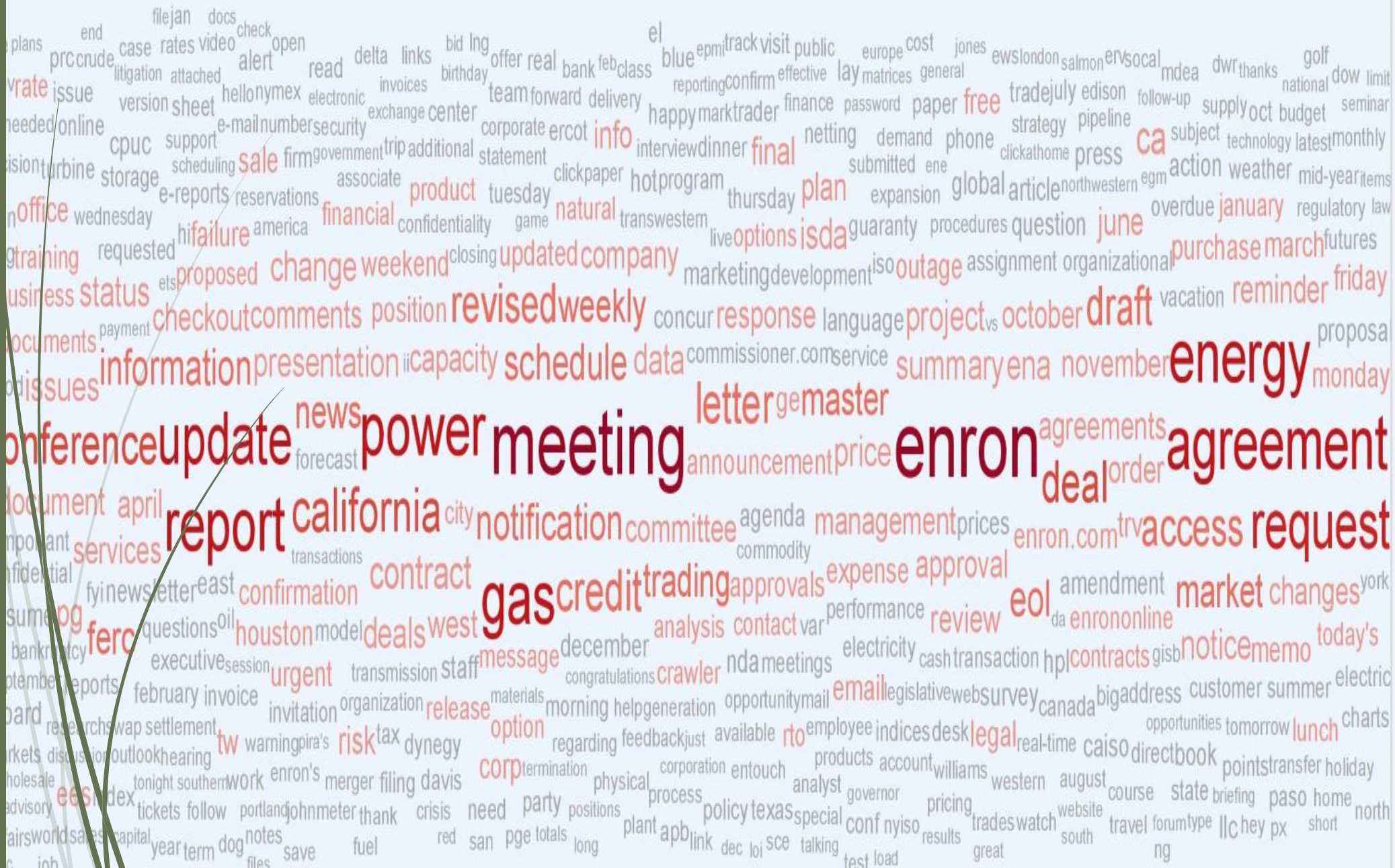
- Steven Kean

  Vice President

- Kate Symes

  Employee

Trending words in Year 2001

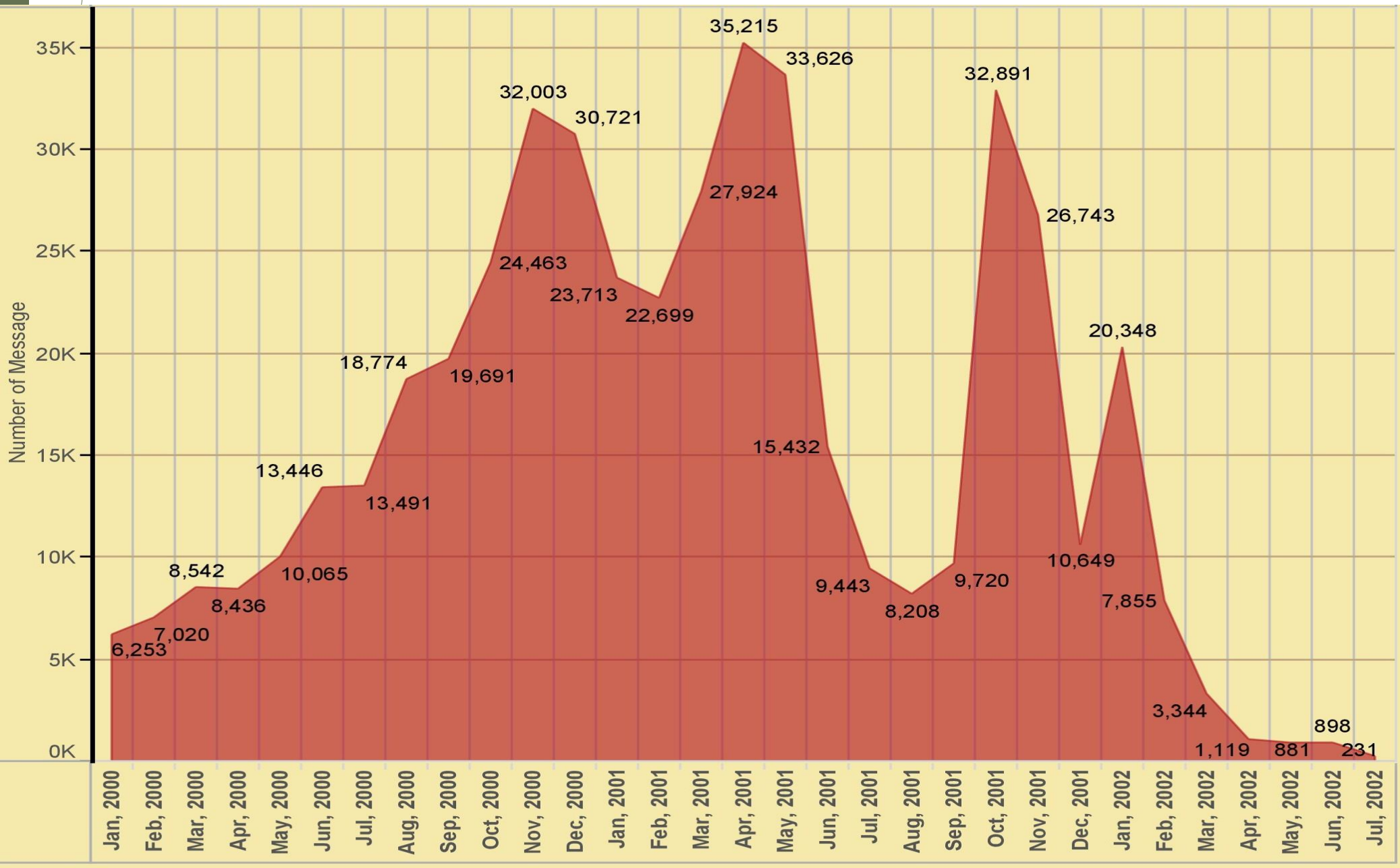


Mails Exchanged In Year (2000-2001)

- **Assumption:**

  Each mail comprising more than one addressee is considered as single mail and not accounted for individual mail to each of the person involved in conversation.

Mails Exchanged In Year (2000-2001)



Inference From Mail Exchange Data

- Based on Number of Mails Exchanged in Enron following spikes are observed:
  - **December, 2000 :**

Enron shareholders filed a $40 billion lawsuit after the company's stock price, which achieved a high of US$90.75 per share in mid-2000, plummeted to less than $1 by the end of November 2001.[4]

  - **April, 2001-October, 2001**:

Company revealed Bankruptcy and was declared public.

Key Players at Enron during Year(2000-01)

- **Hypothesis:**

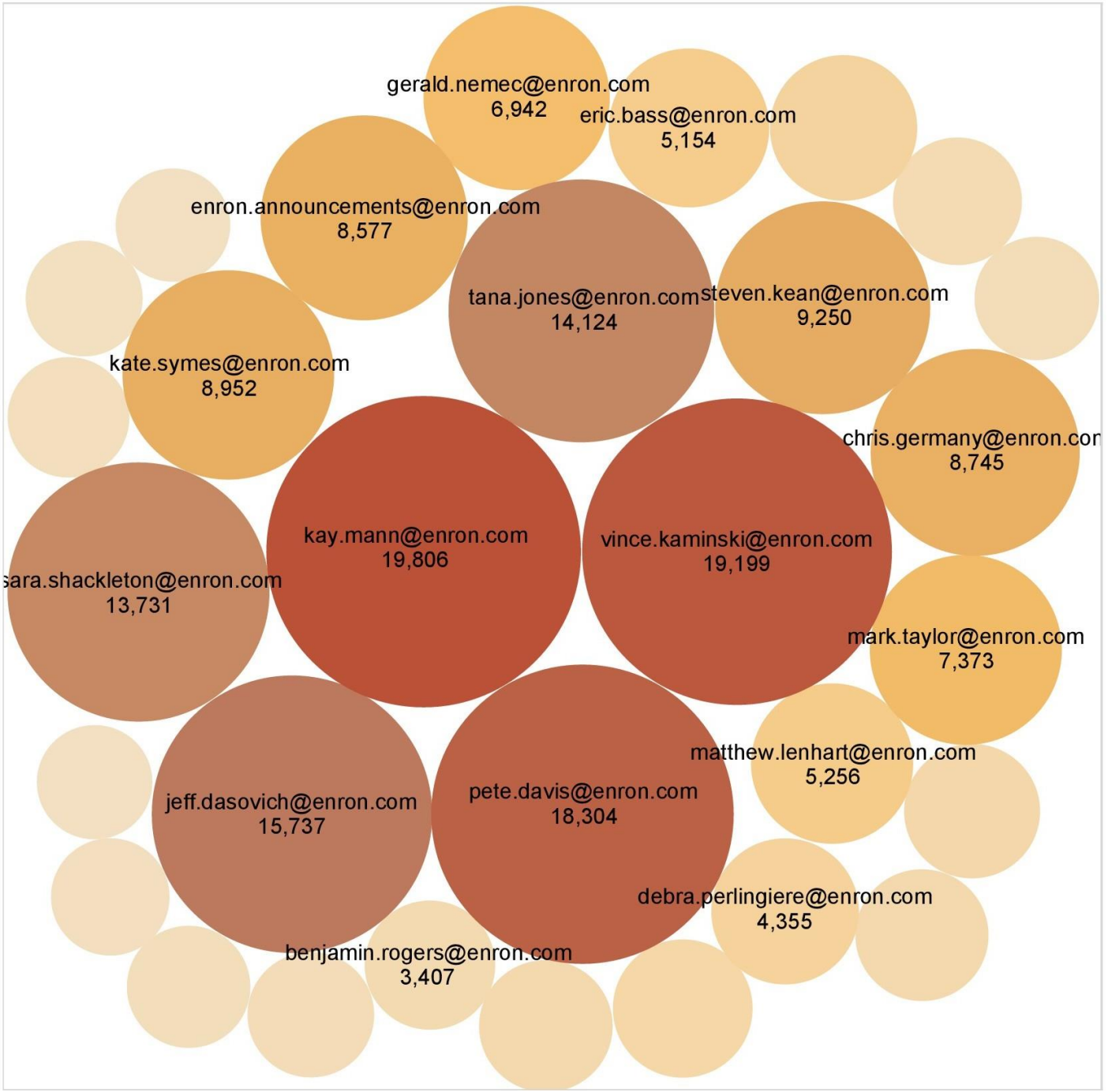
  Key players at enron during year 2000-01 are evaluated based on number of mails exchanged by each person.

- **Assumption:**

  Higher the number of mails sent and received, more important is the person.

  Mails involving more than one correspondent are not considered important enough to contribute.

Key Players at Enron during Year (2000-01)



NetFlix Dataset : Background

- The movie rating files contain over 100 million ratings from 480 thousand randomly-chosen, anonymous Netflix customers over 17 thousand movie titles. The data were collected between October, 1998 and December, 2005 and reflect the distribution of all ratings received during this period. The ratings are on a scale from 1 to 5 (integral) stars. To protect customer privacy, each customer id has been replaced with a randomly-assigned id. The date of each rating and the title and year of release for each movie id are also provided.[5]

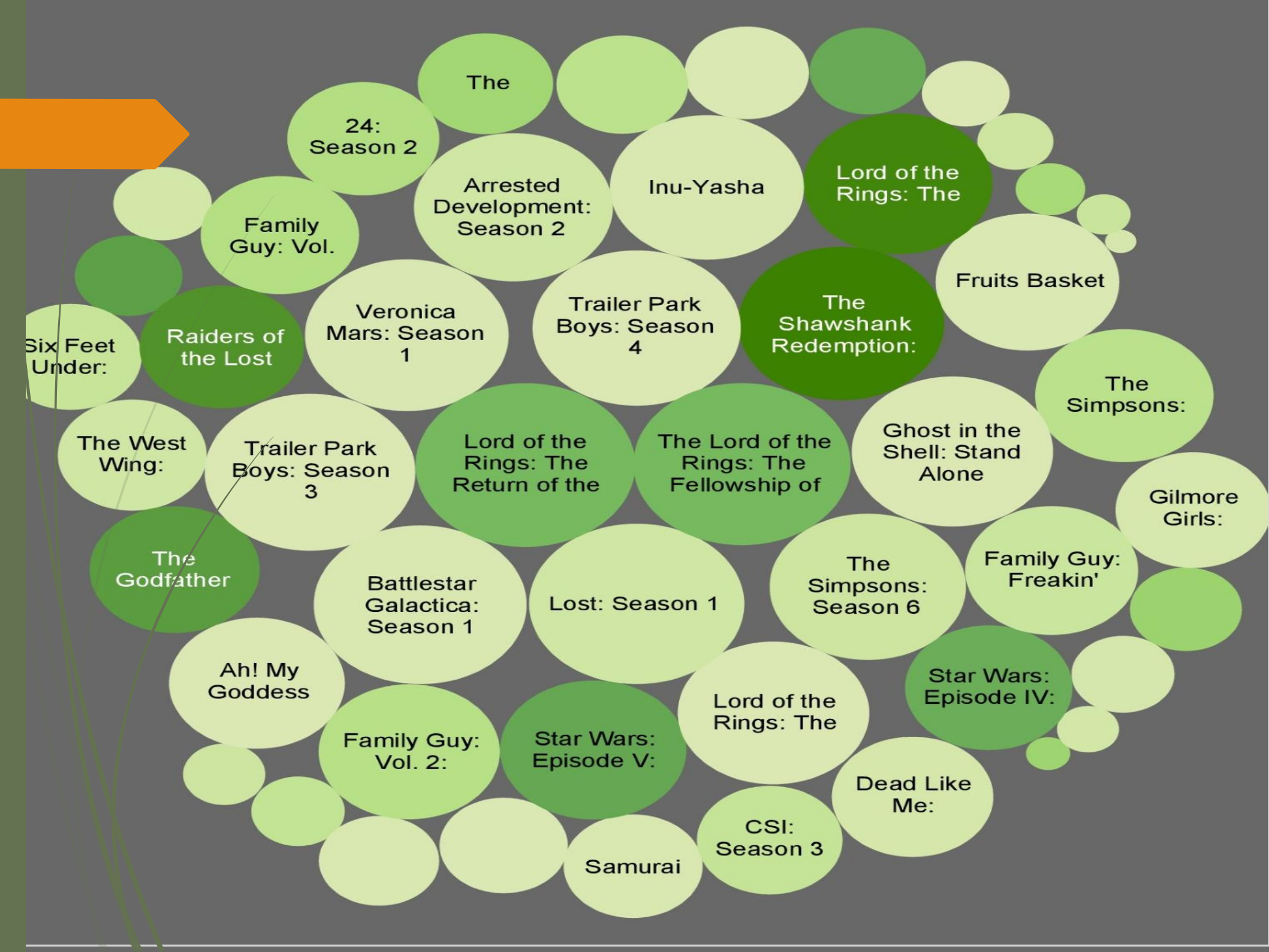
Top Movies

- **Hypothesis:**

  Top movies are evaluated using average rating by each user normalizing rating by number of reviewers of each movie.

- **Attributes:**

  Colors used to represent number of viewers for each movie while size contributes to normalized rating of movie.


The
24: Season 2
Arrested Development: Season 2
Inu-Yasha
Lord of the Rings: The
Family Guy: Vol.
Fruits Basket
Veronica Mars: Season 1
Trailer Park Boys: Season 4
The Shawshank Redemption:
Six Feet Under:
Raiders of the Lost
The Simpsons:
The West Wing:
Trailer Park Boys: Season 3
Lord of the Rings: The Return of the
The Lord of the Rings: The Fellowship of
Ghost in the Shell: Stand Alone
Gilmore Girls:
The Godfather
Battlestar Galactica: Season 1
Lost: Season 1
The Simpsons: Season 6
Family Guy: Freakin'
Ah! My Goddess
Star Wars: Episode IV:
Lord of the Rings: The
Family Guy: Vol. 2:
Star Wars: Episode V:
Dead Like Me:
CSI: Season 3
Samurai

Conclusions

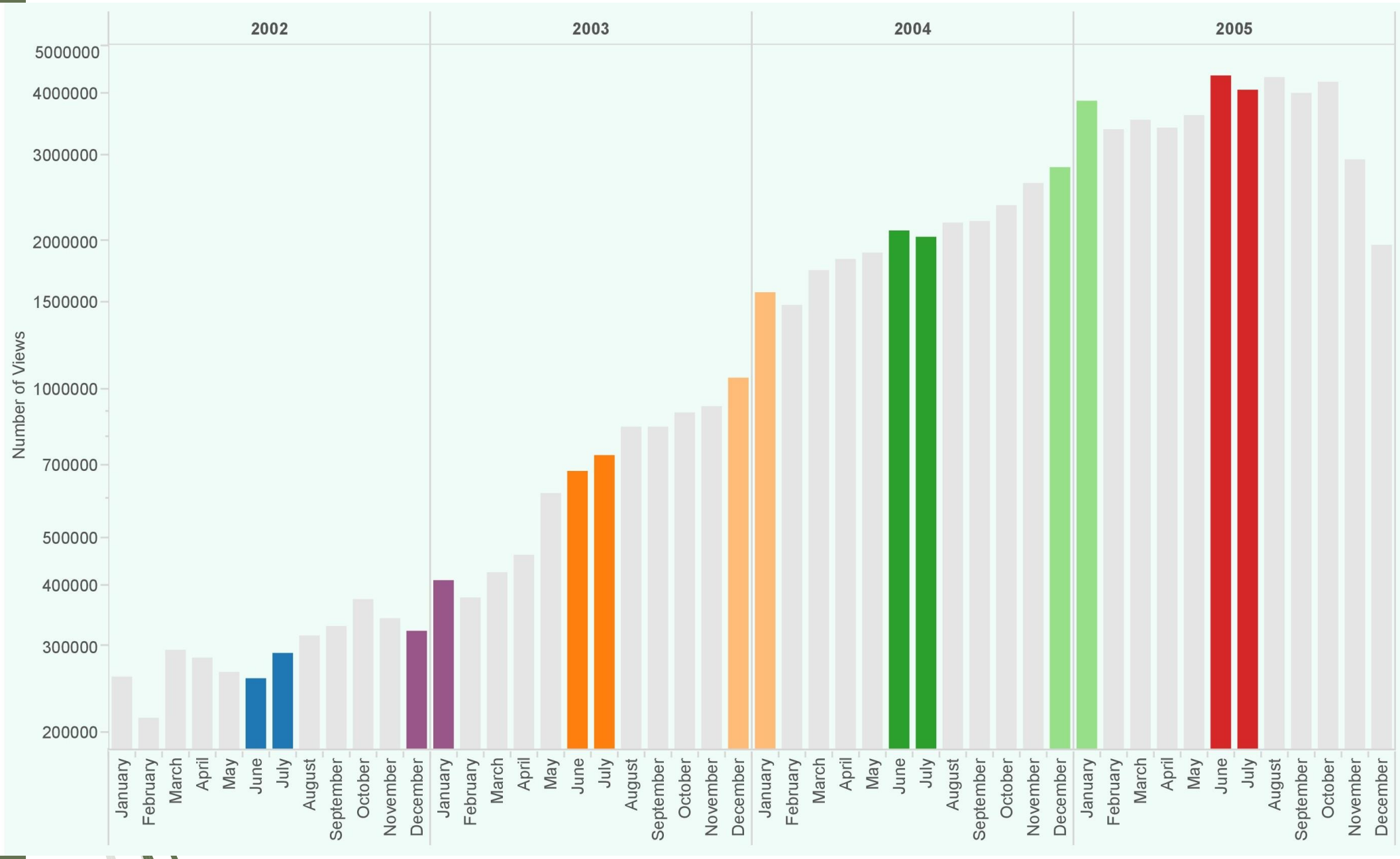
Movie	Overall Rating	Number of Views
Lord of the Rings: The Return of the King: Extended Edition	4.7233	73335
The Lord of the Rings: The Fellowship of the Ring: Extended Edition	4.7166	73422
Lost: Season 1	4.6710	7249
Battlestar Galactica: Season 1	4.6388	1747
Trailer Park Boys: Season 3	4.6000	75
Trailer Park Boys: Season 4	4.6000	25
The Shawshank Redemption: Special Edition	4.5934	139660
Veronica Mars: Season 1	4.5921	1238
Ghost in the Shell: Stand Alone Complex: 2nd Gig	4.5864	220
Arrested Development: Season 2	4.5824	6621

Number of Views

- **Assumption**:

  Every time a user watches movie on netflix, user updates rating for the movie watched, hence every rating for the movie is used to account number of views.

Number of Views (per month)



Inference

- Total number of views experience a spike every January.
- Average number of views during holiday season is moderately high during period of year (2002-2005).

Trends in Movie Titles

References

[1]. Klimt, Bryan; Yiming Yang. The Enron Corpus: A New Dataset for Email Classification Research. CiteSeerX: 10.1.1.61.1645

[2]. "The Enron Email Corpus" Retrieved March 5, 2011.

[3]. Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[4]. "Enron shareholders look to SEC for support in court" (WEB). New York Times (New York Times). May 2007. Retrieved 2013-05-96.

[5]. http://www.netflixprize.com/community/viewtopic.php?id=68