NETWORK ANALYSIS: ENRON EMAILS

PROGRAMMED IN R & PYTHON
BY CHUNMEI GAO

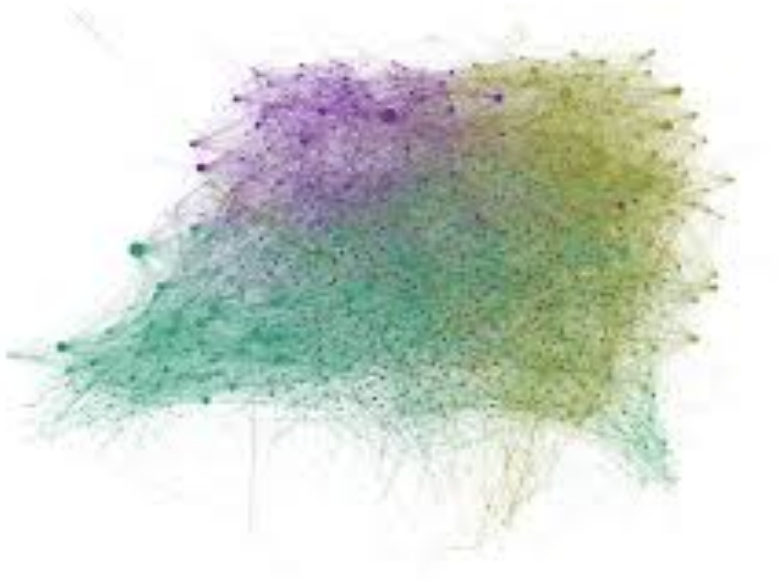
NETWORK ANALYSIS

- Network Theory:
  - The study of graphs as a representation of the relationship between discrete objects

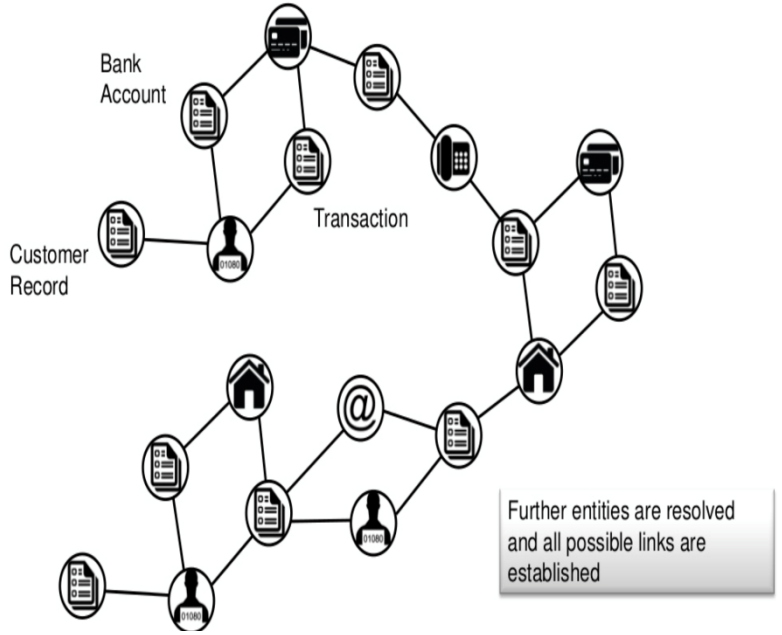
- Those objects could be various items:
  - Facebook ~ Friendship connections
  - Insurance Companies ~ Fraud detection
  - Corporate Firms ~ Email communications between employees

EXAMPLES OF NETWORK ANALYSIS GRAPHS

Facebook Social Network



Fraud Detection



NETWORK ANALYSIS PROJECT IN RELATION TO ENRON EMPLOYEES EMAILS

- Project Instructions & Objective:
  - Plot communications between employees
  - Select an employee folder in order to build a subset of the data
    - Provide employee names and email direction in R & Python
    - Identify the subset's emails subject line titles and graph creatively
  - Identify the name of the central person

ENRON'S PROJECT DATASETS

- Datasets Used: http://snap.stanford.edu/data/email-Enron.html
  - Size of Data: 2.73GB
  - Folders: 156 Employee Subsets
    - 13807 Emails
    - 4014 Unique Enron Employees
    - Direction
    - Email Details:
      - From
      - To
      - CC
      - BCC
      - Date
      - Subject Line

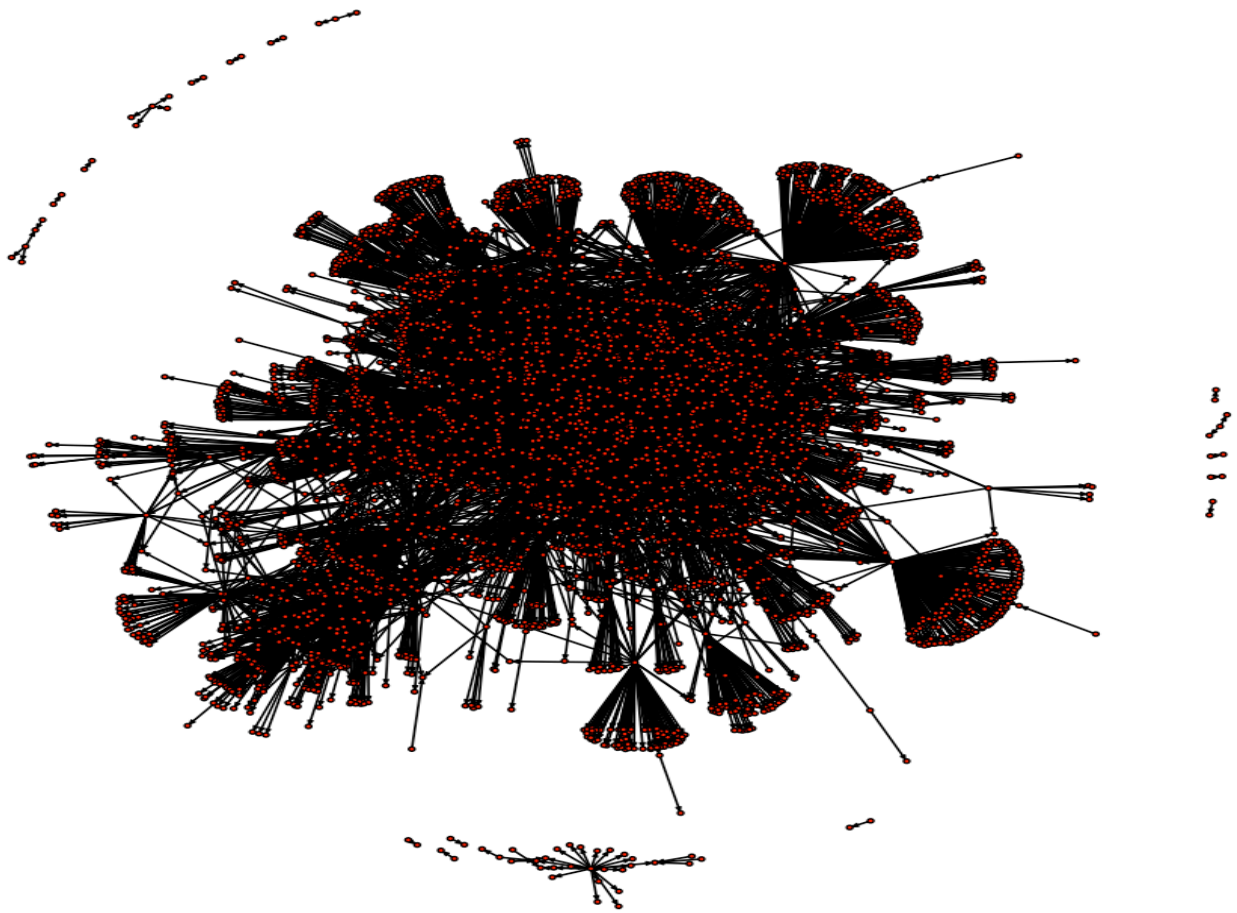
EXAMPLE OF THE RAW DATA

```
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: carrie.blaskowski@enron.com, jennifer.stewart@enron.com
X-From: Tracy Ramsey
X-To: Colleen Koenig
X-cc: Carrie Blaskowski, Jennifer Stewart
X-bcc:
X-Folder: \John_Arnold_Nov2001\Notes Folders\Active international
X-Origin: ARNOLD-J
X-FileName: jarnold.nsf

ACTION ITEMS:
(1)  Include additional EES and Enron Corp conferences where trade credits
may be applicable.  Identify cash spend estimates for existing and new
```

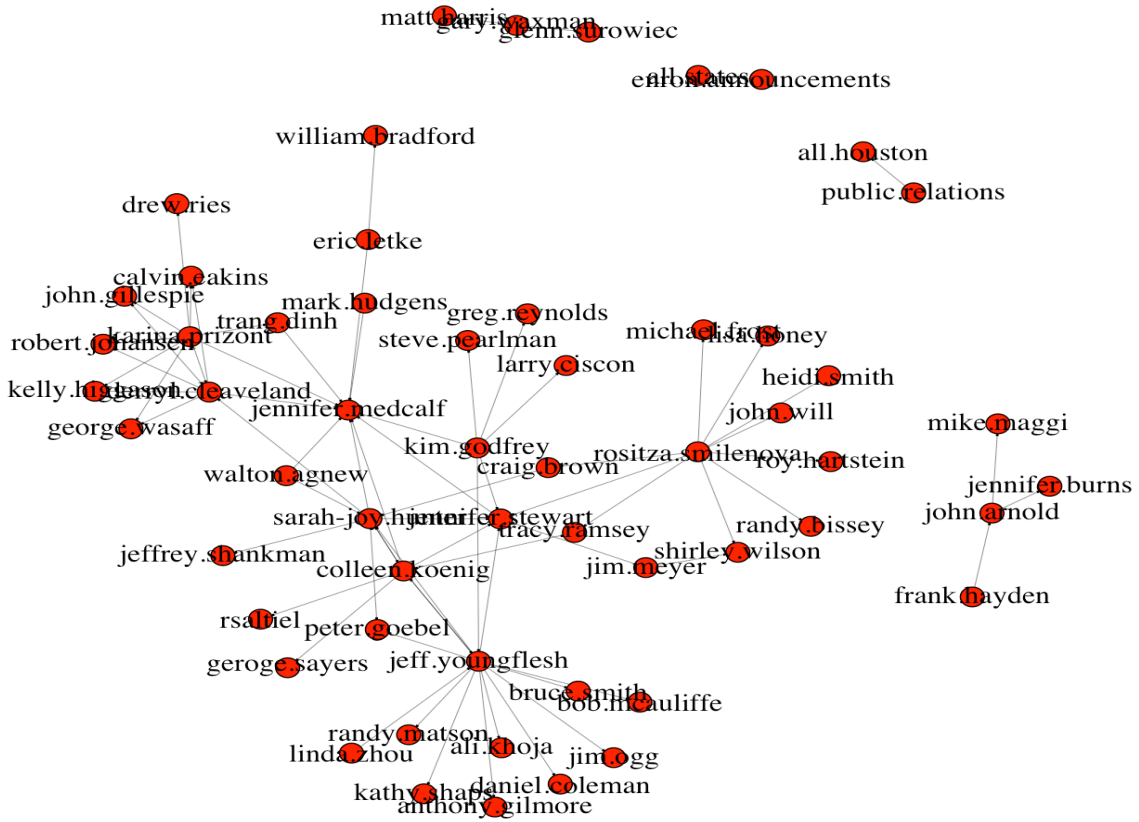
SNAPSHOT OF ALL ENRON'S EMAIL COMMUNICATIONS - AGGREGATED VIEW

- Red Dot, indicates unique employee.
- Line between points indicates email communication.
- Programmed in R.



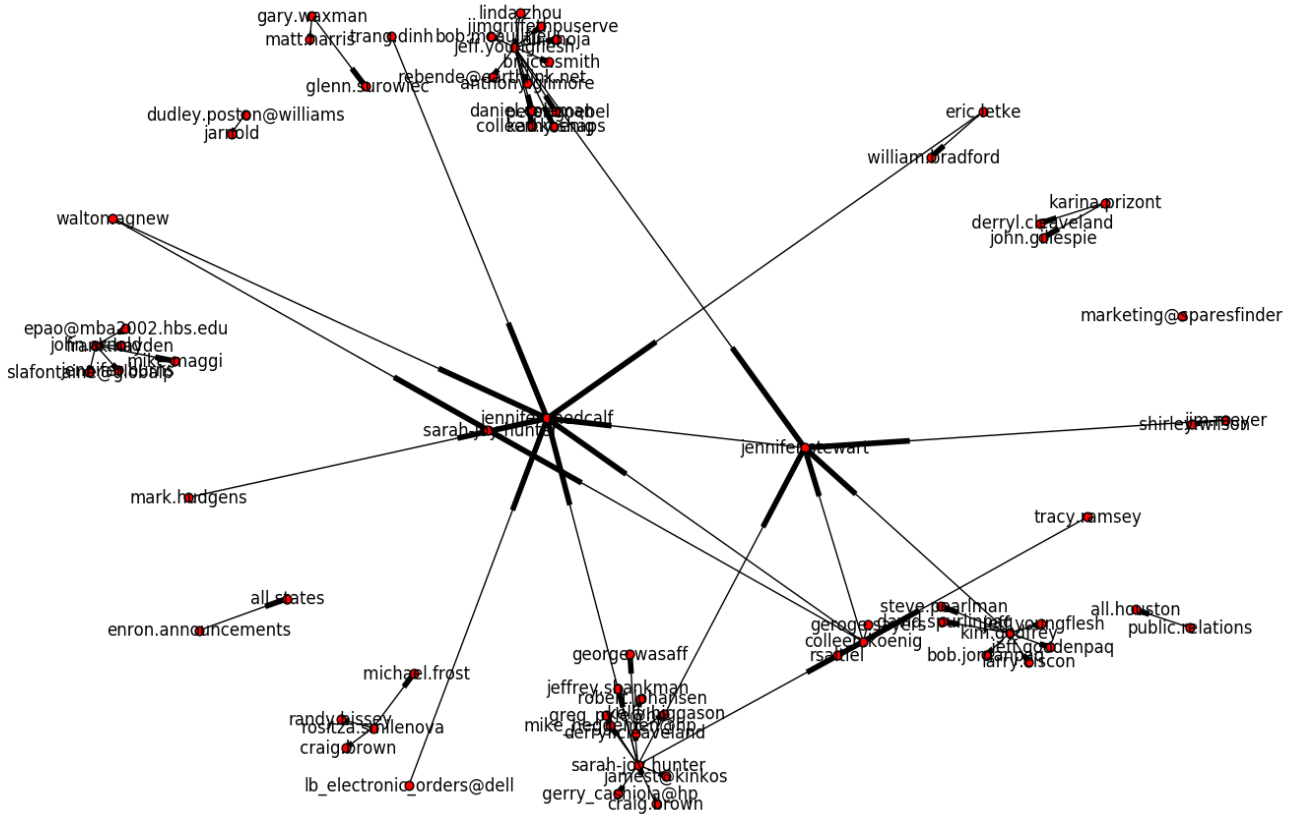
SUBSECTION BY AN EMPLOYEE FOLDER

- Red Dots indicate an unique employee and their name.
- Employee connections by name.
- Structure of the subset's network.
- Centrality
- Programmed in R.



SUBSECTION BY AN EMPLOYEE FOLDER

- Centrality of Employee within the network.
- Email contact direction.
- Programmed in Python.



CONNECTION MATRIX

- Numbers of email contact between each employee

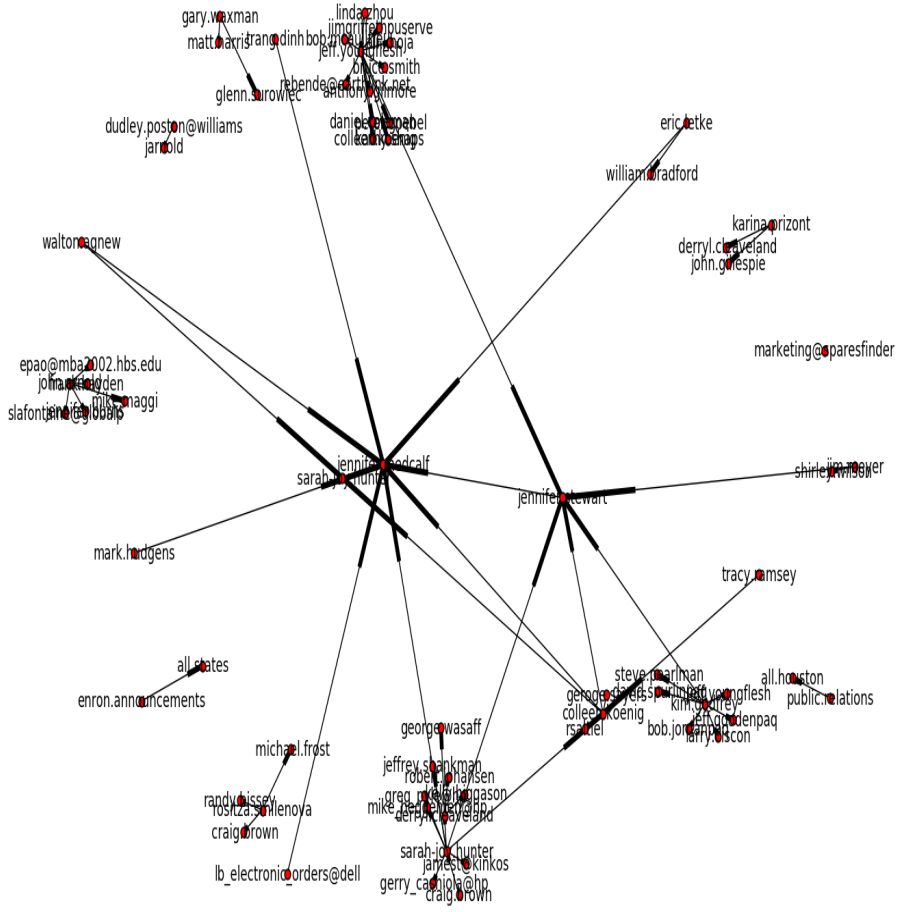
	A	B	C	D
		jennifer.med	colleen.koen	anthony.gil
1	jennifer.medcalf	0	0	0
2	colleen.koenig	2	2	1
3	anthony.gilmore	0	0	0
4	jennifer.stewart	0	0	0
5	sarah-joy.hunter	3	0	0

EMAIL SUBJECT LINE: WORD CLOUD

- Email Subject Line Frequency
- Programmed in Python.

CENTRAL EMPLOYEE BY NAME

- What does Centrality mean?
  - Most messages sent / received
  - positional advantage, fall on the shortest pathway between other pairs of employees.
- Who was central?
  - Maximum Indegree: Ginger Dernehl
  - Maximum betweenness: Steven Kean



CODE

- https://github.com/gaochunmeibb/EnronEmail-Social-Network-Analysis

QUESTIONS