Exploring Enron's Emails via Graphs

Gary Shetye, Aneesha Sreerama, Dachuan Zhang, Shivangi Sood

Northeastern University

DS4300 - Large-Scale Storage/Retrieval

John Rachlin

June 24, 2021

Introduction

Comprehending large datasets and the stories they tell can be very difficult given the sheer accessibility of large volumes of data nowadays. In this class we've explored big data such as the gene-disease associations dataset. We've also considered how to build databases that can handle large amounts of data such as the amount that comes with Twitter. As business and data science students, we turned to finance datasets for our final project. Unfortunately, gathering data on financial holdings by mutual funds and comparing data holdings through ETFs required capital intensive subscriptions to services such as WRDS and CapitalIQ.

Eventually, through resources at CMU, we were able to download and view over 500,000 emails between Enron executives in the years prior to the infamous Enron financial scandal. Given that our "data" was stored in roughly organized folders instead of .csv files or a relational database, our project came down to how well we could extract and organize the data from the emails. We decided to look specifically at the top two executives implicated in the case, Jeffrey Skilling and Kenneth Lay, both ex-CEOs, who took great lengths to cover up Enron's sketchy dealings by means of cover-ups, fake email addresses, and confidentiality within the company. We asked questions regarding how often the convicted executives messaged each other and did they use their hidden personal emails or work emails? We wanted to understand the story our email data could tell so we chose NEO4J's graph database system to visualize these stories.

Methodology

To map these relationships and understand the inner workings of this scandal, we used Neo4j. In terms of our data model in Neo4j, given that we stored an email as a record with all its properties, we decided it would be best to model our data with a sender and receiver node with the email as the relationship. The sender node has the attribute of the *from* email address and a

receiver node has the attribute of the *to* email address. The email would be the relationship between who's sending and receiving and its attributes include a unique message id as the identifier, the date that the email was sent, an encoding, the length in characters of the email, the subject of the email, and then the body of the email.

Initially, to first get a clearer picture on what was going on we decided to ask what accounts Jeffery Skillings was using to send emails. Interestingly enough we found that he was actually sending emails through five email accounts, four of which were not even his. Through graphing our results as shown in Exhibit A, you can see all of Jeffery Skilling's email accounts through which he was sending emails from. These email accounts include jeff.skilling@enron.com, which was his main email account, and the rest were other employees accounts he was using to hide his real identity. These accounts include sherri.sera@enron.com, katherine.brown@enron.com, joannie.williamson@enron.com, and sherri.reinartz@enron.com. We decided to further investigate these five email accounts to see who these email accounts were sending emails to. In Exhibit B there is a graph that maps all the accounts that Jeffery Skillings was sending emails to. We were surprised to learn that Sherri Sera was the one who was sending the most emails, while Jeffery Skilling's main email account was sending the fewest. We were able to conclude that the fact that he was sending the most emails through Sherri Sera's accounts alludes to the fact that he may have been trying to hide some illegal activity. Additionally, you can see in Exhibit C that the top account he was sending to was Sherri Sera, which implies that he was keeping things between his assistant or he was emailing messages to himself.

In regards to Kenneth Lay, we also wondered what accounts he was using to send emails through. We found that just like Jeffery Skilling, he also was sending emails through other

people's accounts. The email addresses of the other account besides his main account include Rosalee.fleming@enron.com and Tori.Wells@enron.com. Next, we decided to analyze the accounts that Kenneth was sending emails from to see if the majority of the emails he was sending were not coming from his main account, similarly to Jeffery Skilling. Through querying the database we were able to find that this was the case for Kenneth Lay as well as Rosalee Flemming was sending the most emails as shown in Exhibit D. To take a deeper look into who was receiving the most emails from Kenneth Lay as shown in Exhibit E, we can see that it was Sherri Sera. Coincidently, both Jeffery Skilling and Kenneth Lay were sending the most emails to Sherri Sera. This could have meant this account could have been a point of communication for the two to hide any illegal activity.

To further support our hypothesis we analyzed the frequently used words in emails between top executives. We found that "energy" + "california" were among the frequently used words as Enron was confirmed to have illegally manipulated the energy grid in California in the early 2000s. We mapped out the network of these emails and found that the majority of the mentioned emails were between email accounts that were not affiliated with "@enron.com" (Exhibit F). Furthermore, the majority of these emails were between Kenneth Lay and Jeffrey Skilling as seen in Exhibit G.

Insights/Conclusion

In the project, we effectively revealed some suspicious behavior of Kenneth Lay and Jeffrey Skilling, the two leading players in the Enron Email Scandal. Through our preliminary research before the model creation, we made the assumption that both characters used multiple email addresses in the network. It was interesting to discover that neither of the two executives used their actual company email for the majority of their communication and adopted

impersonated accounts instead. We also found a close relationship between Kenneth Lay and Jeffrey Skilling as both of their email accounts, including impersonated ones, were among the top recipients of each other.

We can pursue our project further by investigating the relationship between convicted and unconvicted employees through Neo4j Graphs. Specifically, we wish to identify the existence of some sub-networks that are yet to be discovered. We also would have had the opportunity to establish a shortest-path algorithm to compare the distance of lower employees to executives who were convicted if we had emails from employees that weren't only a part of Senior Management. This would have given a much better image of how much of the company was involved with Enron's illegal activities

Author Contributions

Aneesha Sreerama found and prepped the data and organized it into tables in a .csv file via a Python script. Gary Shetye researched background information on Enron, drew up the presentation deck, and contributed to the report. Shivangi Sood contributed to the creation of the graphs and queries, as well as the report. Dachuan Zhang contributed to the creation of the graphs and queries, as well as the report.

Appendix

Exhibit A

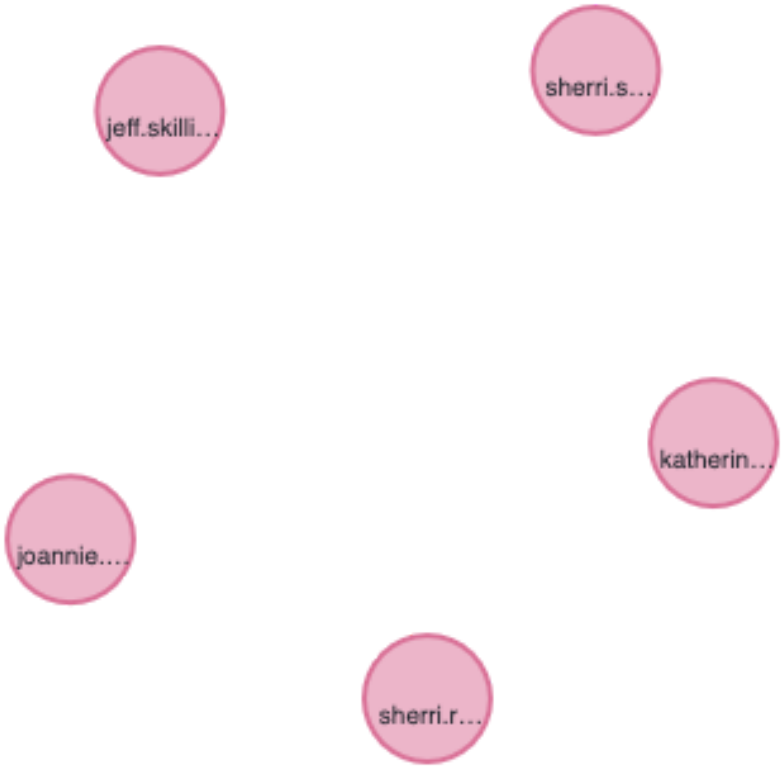


Exhibit B

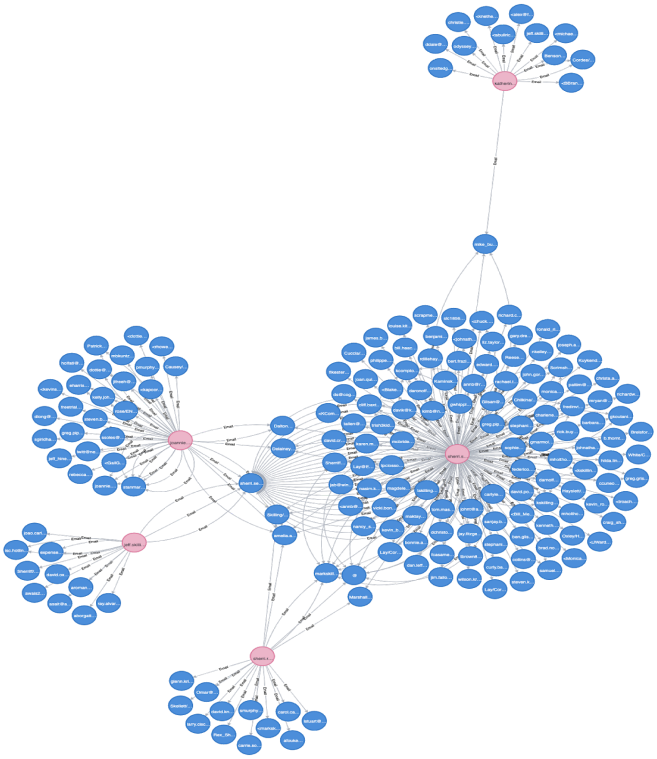


Exhibit C

	to_email	num_emails
1	"sherri.sera@enron.com"	32
2	"markskilling@hotmail.com"	10
3	"kskilling@ehshouston.org"	7
4	"amelia.alder@enron.com"	6
5	"kcompton@kpcb.com"	6

Exhibit D

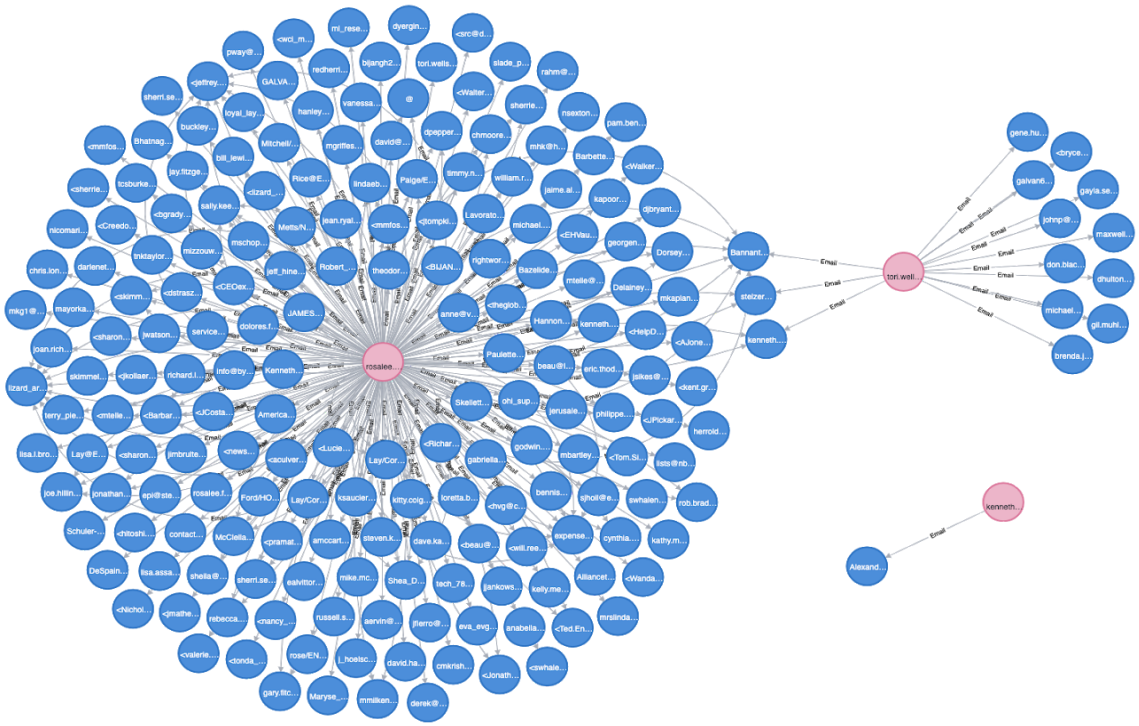


Exhibit E

	to_email	num_emails
1	"sherri.sera@enron.com"	34
2	"markskilling@hotmail.com"	10
3	"Lay/Corp/Enron@ENRON"	9
4	"expense.report@enron.com"	9
5	"lizard_ar@yahoo.com"	9

Exhibit F

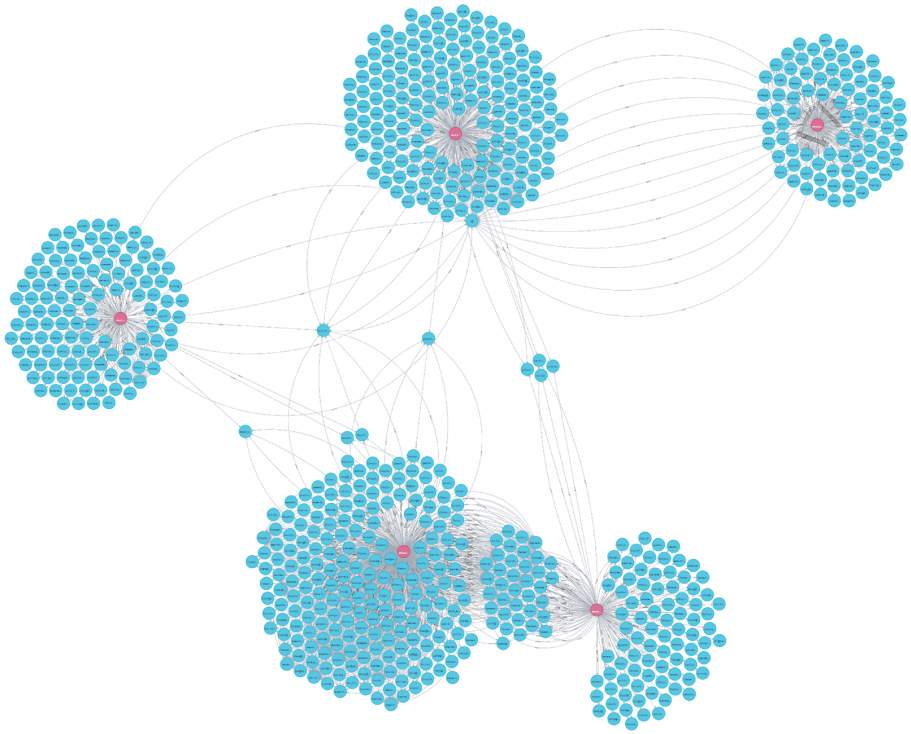
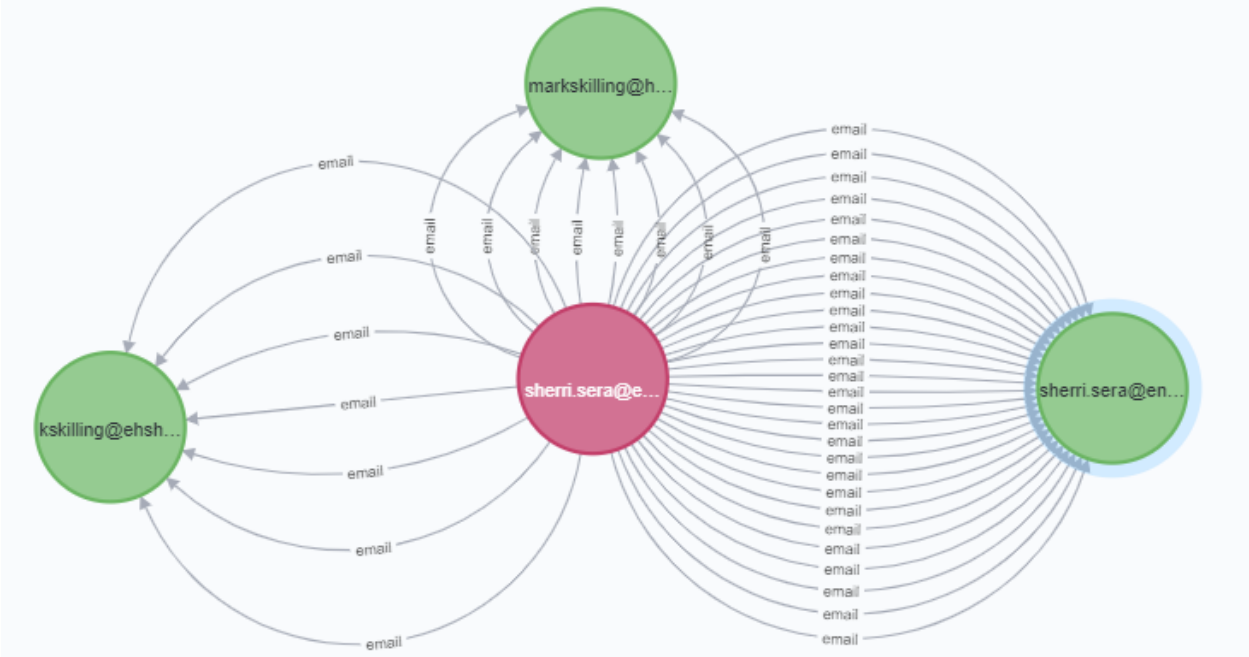


Exhibit G


markskilling@h...
kskilling@ehsh...
sherri.sera@e...
sherri.sera@en...
email