Data Analysis Project

Abstract— This document discusses our findings during the course of investigating two datasets. During our exploration, we made use of a variety of programming languages and packages, including Matlab, R, Gephi, Bash, Python, Java, and occasionally Excel (pivot tables are great!). We used our own laptop and desktop computers, as well as the high-performance computational resources provided by the Advanced Research Computing (ARC) group. We used a variety of analytical models to explore the data files, as well as writing our own custom searching, filtering, and analysis code.

This document is broken down into the following sections: Section 1 describes the storylines and scenarios that we found in both datasets. Section 2 begins to describe how we made these findings, beginning with a discussion of data handling and variable creation. Section 3 discusses the models that we created, including methods of aggregation, explanatory models that we used, and justifications for using those models. Section 4 discusses computational issues that we encountered during our exploration. Finally, Section 5 discusses the importance of our results and conclusions.

1 Storylines

In this section, we summarize the scenarios that we discovered within each dataset. The methods used to uncover these scenarios are described in more detail in subsequent sections.

1.1 Dataset 1 – ACME

As we began to explore this dataset, we were biased towards looking for an evil plot of corporate subterfuge or a scenario along those lines. Under the rationale that most nefarious behavior would take place at night when fewer employees were present in the office, we focused on searching the after hours behavior of the employees. We considered "after hours" to be after 10:00 PM and before 5:00 AM, while "late night" was considered to be between 12:00–4:00 AM.

Among some other bizarre behavior (such as senior manager Fulton K. Rojas, who worked a 22 hour shift but primarily surfed the web all night and throughout the day), we identified some employees who connect devices very late at night. Listed in Table 1 are individuals with such activity. The first four employees listed leave in the table the company, while the other two do not.

ID	Name	Role	Late Connects
GML0105	Germaine M. Lyons	Technician	5/9
CTR0537	Candice T. Ramos	Admin. Staff	3/4
LKY0181	Leroy K. York	Engineer	4/5
AFF0760	Aubrey F. Foster	Foreman	3/3
JSE0020	Jane S. Eaton	Tradesman	3/3
MCH0530	Matthew C. Hayes	Janitor	6/6

Table 1: Employees with late night device connection activity.

We found such activity suspicious since these employees rarely ever connect devices, but some leave the company while others do not. A potential explanation is that some employees may not have been caught doing suspicious late night activity, while other had been caught and fired. Behavior related to after hours connects and disconnects vary from user to user — some log in for a short time and connect the device for a short time, while others may leave the device connected for hours. Neither their web or login activity appeared suspicious, and they only use their personal PC to connect the device. Figure 1 shows a Principal Component Analysis (PCA) plot of all 1000 employees, using aggregated of counts for logons, connects, late night connects, website visits, and months worked as inputs. We see the individuals who left the company (in blue) have somewhat different behavior than the rest of the employees and those who stayed (in red).

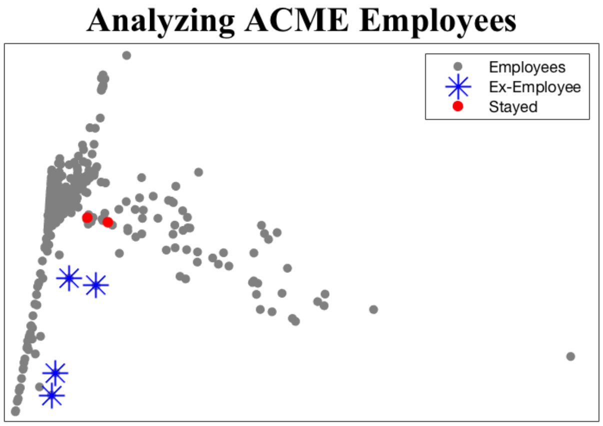

Fig. 1: A PCA plot of ACME employees, highlighting the six employees identified in Table 1.

1.2 Dataset 2 – DTAA

We noticed early on in both datasets that a number of employees left each company during the course of the provided 17 months of information. Interestingly, the pattern of employees leaving was different between the two companies. At ACME, 6.5% of employees left the company and the departures appeared to be randomly distributed. In contrast, 15.5% of employees left DTAA, with an evident increase in the departure rate between March and November 2010 (Figure 3).

Why could this be important? We noted that employees at DTAA worked much more often on weekends and holidays than the employees at ACME. In fact, no one at ACME worked on the weekends outside of Friday late night shifts. The roles are more varied at DTAA than at ACME, but for similar roles, there are some differences. When comparing the activity of IT Admins, we found some stark differences. All of the IT Admins at ACME stay with the company the entire time, and all exhibit similar behaviors: they all log on to many different PCs, have late night activity, and only one connects devices. However, at DTAA, 42% of IT Admins leave the company, and their behaviors are not as consistent as in the first dataset.

1.2.1 Part 1 – The Spread of "Prince"

Throughout the 17 months, we noted that emails containing repeated instances of the word "prince" spread throughout the company. This appears to be a virus or some sort of malware spreading throughout the company. Figure 2 shows several views of the spread of this word over time. Initially, the infection was contained within the Sales and Marketing department, which is understandable since the majority of emails are internal to departments. However, the "prince" virus eventually began to spread to other departments, with all but 18 employees sending an infected email at some point. Indeed, these 18 individuals never receive emails infected with "prince" suggesting the infection

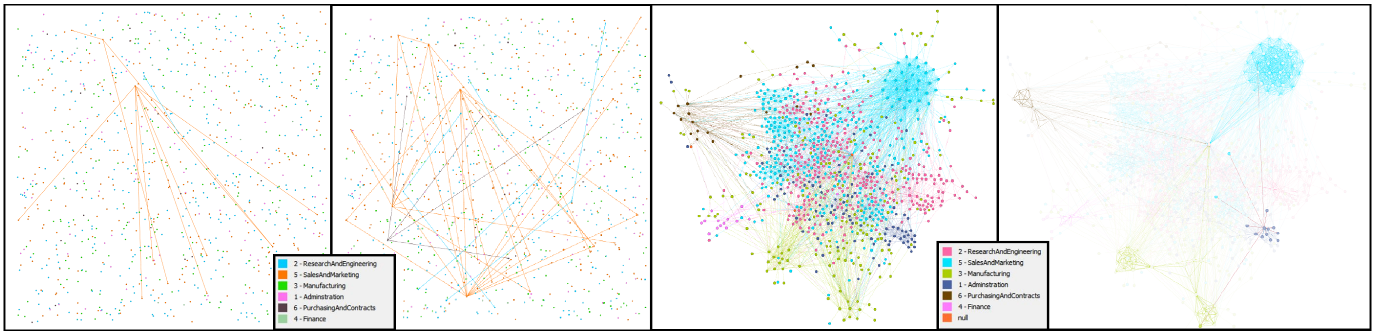

Fig. 2: The spread of “Prince” throughout the company by email. From left to right, this figure shows the spread of “Prince” on Day 1, at approximately Day 100, on the final day, and highlighting the most responsible departments on the last day. Networks generated with Gephi.

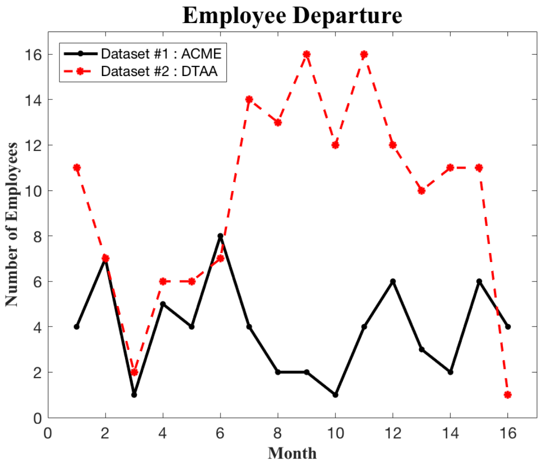

Fig. 3: Employee departures by month at each company.

Functional Unit	FD	Send	Rec.	Send %	Rec. %
Manufacturing	31	11725/ 26979	11087/ 40754	0.43	0.27
Purchasing and Contracts	61	9543/ 24495	8967/ 33672	0.39	0.27
Manufacturing	32	12353/ 33682	11476/ 50511	0.37	0.23
Sales	52	69124/ 744229	65615/ 1038843	0.09	0.06

Table 2: Counts and proportions of *emails* infected with “prince.”

Functional Unit	FD	Infected	Total	Prince %
Purchasing	61	2897	11627	0.25
Manufacturing	32	2080	8949	0.23
Manufacturing	31	480	2553	0.19
Finance	43	96	1212	0.08
Sales	52	3526	66543	0.05

Table 3: Counts and proportions of *files* infected with “prince.”

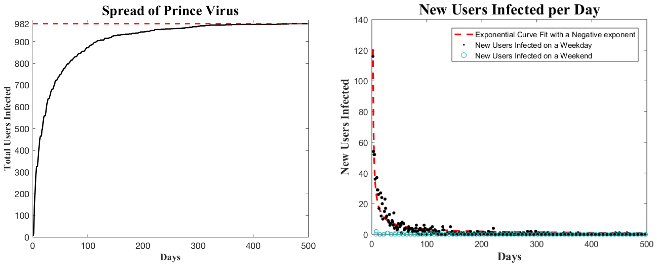

Fig. 4: (**Left**) Cumulative infected users per day. The horizontal line shows the cap at 982 infected employees. (**Right**) New users infected per day. The plot shows the number of unique new users infected per day with a curve fitted.

lies within the organization.

Table 2 shows the number of emails sent and received with “prince” by functional unit and department (FD). We look at the top 4 such units by raw count of infections and then by proportions. These counts and proportions confirm that the functional units found in the final day network graphic in Figure 2 are contributing the most to the spread of this virus. We note that the Sales group have a high raw number of email communications and count of “prince” infected emails, but proportionally is not as high as the Manufacturing groups. In contrast, we look at the top-5 units who access infected files with prince and notice that proportionally the Sales group is not the highest, but is in raw counts of accessing infected files.

We also examined the spread of the virus using the number of employees infected per day (Figure 4). We see this count grow quickly in the first month or so and before midway, almost all the company is infected by the virus except for those 18-employees.

When we break this down further and examine only newly infected users, we see something somewhat surprising. The infection spreads quickly in the beginning weeks, but the number of newly infected users decrease for the remaining months. An exponential curve fitted on this data has a negative exponent (-0.78), which suggests that an attempt was made to contain the infection. Who better to tackle such a problem than the IT Admins? These employees also display some interesting behavior, and we next discuss the IT Admins working for DTAA.

1.2.2 Part 2 – Nine Angry IT Administrators

Within the `file_info` data, we discovered nine users (listed in Table 4) who accessed an executable file that contained suspicious content. The possibility of a malicious program naturally alarms us. Our investigation was furthered by the fact that these programs appear to contain keyloggers (each of these employees visited websites with “keylogger” as a keyword in `http_info`). After combining information from across the collection of data files, we discovered that these nine users were all IT administrators, with each having significantly less device connection activity compared to their colleagues. When examining their email communication as well as file contents, we found a strange pattern of behavior (shown in Figure 5) that repeatedly occurred only a few days before each left the company.

Our analysis of this behavior is that the IT Admins are probably overworked due to the spread of “prince,” and their workload continues to increase as a result. One of the major complaints seen in their emails is that they are made to work on Weekends and holidays too. This has caused them to retaliate against the company. Additionally further investigation of the `email_info` file shows that these IT Admins were

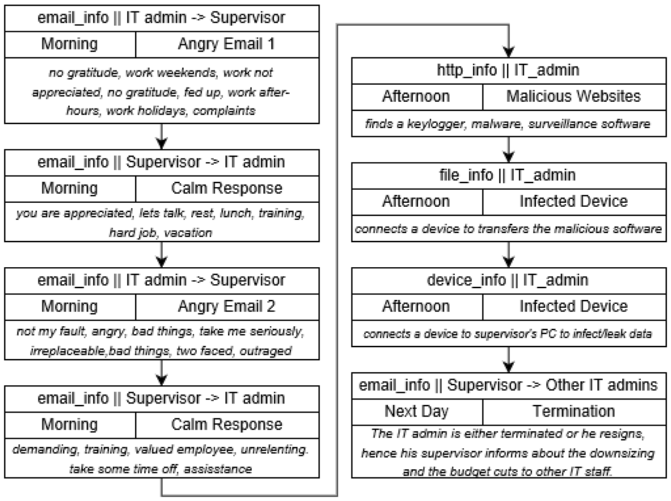
email_info || IT admin -> Supervisor
Morning
Angry Email 1
no gratitude, work weekends, work not appreciated, no gratitude, fed up, work after-hours, work holidays, complaints
email_info || Supervisor -> IT admin
Morning
Calm Response
you are appreciated, lets talk, rest, lunch, training, hard job, vacation
email_info || IT admin -> Supervisor
Morning
Angry Email 2
not my fault, angry, bad things, take me seriously, irreplaceable,bad things, two faced, outraged
email_info || Supervisor -> IT admin
Morning
Calm Response
demanding, training, valued employee, unrelenting. take some time off, assistance
http_info || IT_admin
Afternoon
Malicious Websites
finds a keylogger, malware, surveillance software
file_info || IT_admin
Afternoon
Infected Device
connects a device to transfers the malicious software
device_info || IT_admin
Afternoon
Infected Device
connects a device to supervisor's PC to infect/leak data
email_info || Supervisor -> Other IT admins
Next Day
Termination
The IT admin is either terminated or he resigns, hence his supervisor informs about the downsizing and the budget cuts to other IT staff

Fig. 5: The behavior pattern of our suspicious IT Admins before they left the company, including the applicable files in the DTAA dataset.

ID	Name	File Accessed	PC Used
CSC0217	Cathleen S. Craig	06/10/10 15:20	PC-6377
GTD0219	Guy T. Daniel	06/17/10 15:14	PC-6425
JGT0221	Jonathan G. Terry	07/15/10 15:20	PC-2948
JTM0223	Jerry T. McCall	07/22/10 15:11	PC-9681
BBS0039	Bevis B. Sheppard	08/12/10 14:54	PC-9436
BSS0369	Brenden S. Shaffer	09/30/10 16:10	PC-3672
MPM0220	Meghan P. Macias	11/04/10 15:19	PC-2344
MSO0222	Medge S. O'Brien	12/09/10 15:23	PC-2524
JLM0364	Jacqueline L. Miles	04/28/11 16:06	PC-3791

Table 4: IT Admins who put keyloggers on their supervisor's PC

looking for new jobs. We see that multiple emails have been sent from their personal email to various companies containing keywords such as "resume," "experience," and "passion," which led us to conclude that they are job searching. This is particularly threatening to the company as they appear to have installed a keylogger on their supervisor's PC before they left the company, enabling them to receive and leak out company secrets and data.

We used PCA again to project the psychometric scores of all IT Admins (Figure 6), highlighting those who used a keylogger on their supervisor as well as those who left the company. We do not see much of a difference seen between groups in the projection, suggesting that psychometric scores alone are not a good measure of detecting which individuals would likely become disgruntled in the future.

1.2.3 Part 3 – WikiLeaks: The Real Story about DTAA

Continuing the theme of information leaking from DTAA, we found that a number of employees visited a "The Real Story about DTAA" webpage hosted on WikiLeaks. None of the 30 employees who accessed this page stayed with the company for the entire 17 months. Indeed, these employees departed the company between the months of July 2010 and March 2011, roughly the same timespan as their visits to the webpage, and on average they leave the company 20 days after visiting WikiLeaks.

Some of the keywords listed as associated with the WikiLeaks URL are concerning, including "subterfuge," "clandestine," "forgery," and "lie" among others. Though we have no evidence to support either of these hypotheses directly, we consider the possibility that either (1) these employees have discovered some awful truth about DTAA via this WikiLeaks page and resigned, or (2) perhaps they themselves contributed their own knowledge of company activities to this WikiLeak and then resigned.

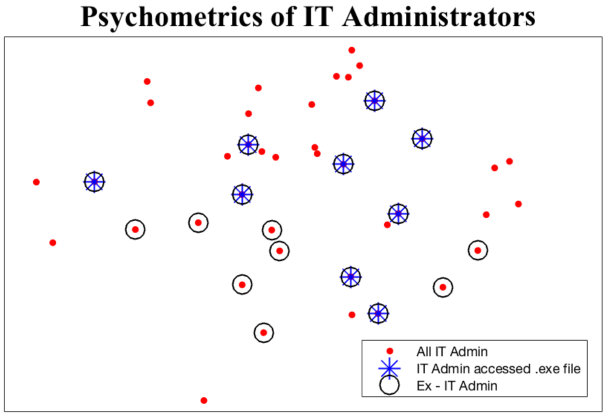

Fig. 6: PCA plot of psychometric scores for all IT Admins. Nothing stands out psychometrically about the IT Admins who used keyloggers relative to the full IT Admin population.

2 Data Summarization

In this section, we describe methods that we followed for loading and manipulating the data, as well as justify new variables that we created as we explored and aggregated the data.

2.1 Data Handling

In initial explorations of the data, we loaded the provided data files into a variety of programs, including Notepad++, R, and Matlab. These programs enabled us to view individual records and to begin to locate interesting observations and attributes in the data. When we detected something that appeared to be worth exploring, we began to filter and sort the data to investigate further. For smaller files, this was computationally feasible using these tools. For larger files, command-line programs such as `grep` were useful for filtering based on keywords that we wanted to investigate. For our DTAA storylines, we were able to use `grep` to locate website visits that contained the "keylogger" keyword or the "wikileaks.org" URL from the `http_info` file quickly and efficiency.

After these initial searches, we began to explore the datasets more deeply. Rather than trying to find storylines with the provided collection of individual files, we worked to create a single master file for each dataset that aggregated and stored all useful information (both variables that were provided to us and variables that we created). This process was considerably easier for the ACME data than for DTAA data, as the resulting size of the aggregate ACME file was not much larger than the initially-provided `http_info` file. These master files gave us more power in locating relationships within the data, such as finding ownership links between employees and their PCs (and identifying shared PCs).

In the case of DTAA, we aggregated some information within the individual files before combining them into a single master file. For example, we aggregated the keywords listed in each individual record in `http_info`, computing a frequency for each keyword by date and employee. Seeing that this file was still quite large, we filtered to only the top 10 keywords aggregated for each date and employee. This more manageable information was then included in the master file. In both datasets, we combined the provided monthly employee files into a single aggregated employee record, tracking the number of months that each user was employed by the company and the month that they left the company (if applicable).

2.2 Variable Creation

The master files that we created for each dataset incorporated a number of variables that were not present in the provided data files. Some of these variables were quite straightforward, such as separating the timestamp field into the individual hour, minute, etc. components. Other variables took a bit more computational effort to create, including computing logon duration and finding logon events that were not followed by an accompanying logoff. These variables were initially created in the ACME master file, but some were duplicated for DTAA as well.

Specific to the DTAA dataset, we created a number of aggregation variables in addition to these straightforward ones. For example, some variables included denoting whether an email or website is "prince" related, as well as the network of users sending and receiving emails (if the email was sent to someone in the company).

Adding these aggregation variables to our master files provided more information to our analysis than using only the individual records, because they enable us to better see patterns in the observations. Understanding areas of more frequent or outlying activity provided hints for where we should focus our current and future investigations.

One possible weakness to our approach of duplicating our ACME variables into the DTAA data was that it biased our initial DTAA exploration. As a result, there was also some bias in the new variables that we created for DTAA. This bias is partly why we did not find most of our DTAA scenarios until shortly before the deadline. We could have potentially improved our DTAA exploration and variable creation by treating it as a new dataset from the beginning, rather than trying to mimic our approach to the ACME data.

Another potential improvement to our variable creation would have been to standardize our variable notation across our group, especially because we started our exploration by each taking ownership of a single data file. The variety of personal notation preferences yielded variables called "emp," "emp_id," and "EmpId" for the same attributes. Minor conflicts then arose when joining tables later in the exploration process when sharing among group members.

3 Data Modeling

In this section, we describe our modeling strategies that we implemented and applied on each dataset. These strategies include aggregating data, applying explanatory and predictive modeling techniques, and justifications regarding the appropriateness of our strategies.

3.1 Data Aggregation

The large data files in both datasets necessitated some aggregation before we were able to merge that content into our master files. Additionally, some of the raw data such as website keywords and monthly employment records were not as useful in raw form as they were in aggregated form.

Starting with the employees, we worked in both datasets to combine the individual monthly records into a global employment picture across the 17 months of data. We implemented some custom aggregation code to create records for each employee, capturing the months that they worked, the month that they departed (if applicable), and the total number of months that they appeared in the datasets. This enabled us to focus our investigation in both datasets on employees who left the company during the time period under investigation.

In the DTAA data, we aggregated the keywords in the `http_info` file, summarizing content about the websites that each employee visited by day via the keywords that accompanied each website visit. After aggregating, we were able to sort the keywords by frequency and detect that the "prince" issue was common in the websites as well as the emails, as well as detecting some accompanying "anhk" and "ahmose" keywords. We were then able to aggregate by items with these words and without. We aggregated the ACME company data on a minute level, but we used a daily level for DTAA to make the file sizes more manageable.

Using Gephi, we created the network graphs as seen in Figure 2 for the emails sent and received by a user. This was useful in displaying the connections between various functional units and departments, as it showed us the spread of "prince" across the DTAA company. The dynamic timeline capability in Gephi enabled us to observe the spread of "prince" over time. We also used Word Clouds to aggregate word frequencies in a subsets of emails, discovering patterns in email groups that lead to some of our Angry IT Admin findings.

3.2 Explanatory/Predictive Modeling

Initially, we tried to visualize the data across many dimensions (both with supplied and created variables for our datasets). We decided to present PCA plots in this paper for ease of explanation, but we also explored the use of *k*-means and other clustering algorithms on the DTAA employee psychometric scores. For psychometric scores alone, we found that there is no inherent structure across all DTAA employees, making these clustering algorithms ineffective for that avenue of exploration. We instead chose smaller subsets of the variables on which to perform *k*-means. One such subset involved looking at infected files and emails. The result of the clustering algorithm found the groups of employees that send emails and never access files and further classified those who do, which failed to capture information previously unknown. In addition to the PCA plots presented here, we generated dozens of others during our explorations using either smaller groups of observations or on a smaller collection of dimensions.

After visualizing the data and mining for relevant information from emails and URLs, we wanted to better characterize groups of employees. For example, consider the group of employees who left each company — were these individuals all fired, did they all resign, or was there some combination of both cases? As we developed our storylines, we created a variety of indicator variables in an attempt to better inform our predictive models. In trying to predict which other employees were at risk of leaving, we decided for the binary responses (left or did not leave the company) to use a logistic LASSO with 10-fold cross validation. We selected this approach based upon discussions from the lecture content, since we knew that not all of our selected variables were going to be relevant to employee departure.

Another important group that we focused on were the DTAA IT Admins. Since they were overworked, some were disgruntled enough to react and leave the company. When looking at a projection of the psychometric scores for groups of IT Admins, we found nothing peculiar (as seen in Figure 6). However, when used in conjunction with other variables, some of these psychometric scores aided in prediction for our models.

3.3 Model Justification

With our binary responses created from our indicator variables, a logistic regression using either a logit or probit link function is suitable. The variables we have considered in our models are related to login activity, device activity, file activity, email activity, some web activity (mainly visiting the WikiLeaks page), psychometric scores, and those activities related to the "prince" virus. We use an aggregated dataset by employee and noted the total number of days active as well as an indicator variable of whether or not this employee stayed. Our choice of comparison model was CART, since this model is appropriate for both continuous and categorical variables. To validate each model, we used a 10-fold cross validation process. In particular for CART, we used bagging within our cross validation to find our best tree.

The initial goal with a LASSO and CART was to predict the number of days employed, but we quickly realized that the activities were in the incorrect scale. We then used both models to predict whether an employee would leave the company or not. To do this properly, we divided the variables associated with activity during that period by the number of days active. We again validated our models using a 10-fold cross validation scheme. A measure of how incorrectly we predict outcomes is reported in Table 5.

Model	Count of Days Active	Left the Company
LASSO	0.200	0.033
CART	0.290	0.007

Table 5: 10-Fold Cross Validation Error Rates

Note the high prediction error rates for the incorrectly scaled data, while the prediction error rates for the latter are much more reasonable. Comparing the prediction errors, CART performs better. However, we are not sure if certain nodes only contain one observation, and so we still find our logistic LASSO to be much more reliable. We show a confusion matrix for the logistic LASSO in Table 6.

	Stayed	Left
Stayed	845	0
Left	25	130

Table 6: Confusion Matrix from Logistic LASSO

Note that the employees who leave around months 14 through 16 will may have similar behaviors as those who stay the 17 months. Hence, some of those who stay the entire duration are much more difficult to classify correctly. Important variables for this classification model include most of the psychometric score coefficients (extrovert being the exception) along with many of the activity variables for log ons and and receiving emails. Many of the "prince" associated variables that we created were not deemed important in determining whether or not an employee would remain with the company.

4 Computational Issues

In this section, we discuss the computational issues related to the size and complexity of the data files used in each dataset. These considerations include how we solved challenges related to data scale and the variety of files, as well as an assessment of the strengths and weaknesses of our modeling choices.

4.1 Computational Considerations and Demands

We have already discussed aggregation strategies for dealing with large files (primarily the `http` files in both datasets). These strategies certainly helped to address computational challenges for dealing with these large files. In addition, we were able to make use of the ARC resources in order to more quickly perform computations on the large data files and our aggregate global datasets. These resources provided a substantial boost in speed over our personal computers. For example, our aggregation code for DTAA's `http_info` was processing one day of website visits in roughly 5 minutes on John's laptop, but accelerated to one day of website visits in roughly 15 seconds on ARC. One difficulty related to using ARC resources was the challenge of remote access to these clusters. Connecting via a VPN from the other side of the planet presented substantial latency issues, nearly to the point of unusability.

Designing multi-threaded code further provided a speed increase, especially when running code in the ARC environment. Despite the additional complications and debugging involved in ensuring that the code was correct, the performance increase was worthwhile when processing large files. For example, we first worked to aggregate the content of the DTAA `http_info` file by date and user in a single thread. After getting a sense for how long that aggregation would take to execute, we preprocessed that `http_info` file, separating it into one file for each of the 1,000 employees. Then, we updated the code to aggregate keywords for each employee in a separate thread, allowing us to make use of multi-core desktops and ARC clusters.

Using Gephi for network graphs was also quite useful, as this software package contains utilities such as filtering, timeline preview, and categorization by label. It also performed efficiently, despite the 1,000 node and 1,300,000 edged graph of email communication that we supplied.

4.2 Computational Modeling Choices

Our choices of models were impacted by how we aggregated and reduced all of the information provided in both the ACME and especially the DTAA datasets. Since our ideas about the stories themselves were finalized shortly before the deadline, the amount of time available to us to run and refine models was greatly reduced. Rather than fitting a more complicated model before downsizing the data or fitting an over-fitted model that may not predict well, we used approaches that are reliable such as a logistic LASSO and CART. The datasets that we ran models on were aggregated by employees in the company, and thus contained 1,000 observations. We could easily run models on datasets of this size and still learn that those employees who stay the whole 17 months are difficult to classify correctly when their behavior is similar to those left the company.

We note that at first we improperly fit models on the raw counts, since we did not account for the number of days active. We also tried to classify those who would be infected with the "prince" virus, but this also produced slight high prediction error rates. A weakness here is not fitting models to the datasets aggregated by the day, but it was difficult with the remaining time to both wrap up the newly discovered stories as well as find appropriate variables to model. Ideally, we would love to understand the data in its raw form, but the sheer size of the DTAA dataset made this nearly impossible.

5 Discussion and Conclusions

In this section, we evaluate the importance of our results to each of the companies, as well as discuss some lessons that we learned while exploring the datasets and completing this project.

5.1 Importance of Results

Our results from exploring the ACME data provide hints towards evaluating the behavior of employees to uncover odd or unusual events, as well as employees who are logging hours without working (as in the case of the 22-hour stint of web surfing). Having the ability to locate and eventually correct odd employee behavior will result in a stronger company overall.

DTAA can use our results to improve their company, especially in planning future enforcement of data management to prevent leaks, as well as better information security policies. We showed that the company has suffered a cyber attack wherein its computers were infected by some malware which corrupted files and email contents. This "prince" malware rapidly spreads throughout the company, primarily via email. The spread likely could have been avoided by using a secure mail client such as Outlook or Proton mail which checks for malicious content.

In reference to the outbreak of the "prince" malware, it appeared that the IT Admins were trying to contain it (Figure 2). However, they were not able to completely contain the infection. We suspect this is the reason IT Admins were overworked, leading to their frustration and retaliation. We recommend DTAA check their staff's working hours and take monthly feedback to estimate their employee satisfaction. Also, IT Admins were able to access suspicious websites and download malicious software. This could have been prevented by using a trusted Anti-Virus software and logging such incidents for review by upper management. They were further able to upload these keyloggers to their supervisor's computer. This is a serious threat to the company, as supervisor's data is being leaked and/or infected. All this could have been avoided with stronger security and data encryption tools.

5.2 Project Lessons Learned

Despite working on these datasets for more than two months, nearly all of our best ideas and findings came in the last two days while writing this report, some even in the final 10 hours. This resulted in a massive rewrite of this document in the final hours before the deadline. In addition to demonstrating the benefits of last-minute panic, this shows that moments of inspiration can occur at any time when exploring the data, even at the last moment.

As noted previously, we felt that our exploration of the DTAA dataset was initially biased towards the approach we followed on the ACME dataset. Because the storylines within the companies and datasets were quite different, this caused us issues with detecting the scenarios that we report in this paper.

Lastly, it is important to have items setup even when not all the pieces are finished. When group members have varying schedules and other deadlines to meet, it can be hard to have all the pieces needed in order to analyze something. Having the code ready to go when those pieces are in place would have saved some time.

Acknowledgments

The authors wish to thank our guiding lighthouses who assisted us in these findings.

APPENDIX

The overseeing of this project was managed very carefully by the lovely daughter of Mai, Ms. Jana ElFishawy. Without this little girl's patience, we would have never been able to finalize our stories or analyze the emails dataset in those very tiring and long weekend and weeknight meetings.

Fig. 7: Jana and Spongebob.

Thanks Jana!