Analyzing and visualizing the data of the VAST challenge 2012

Liveris Avgerinidis 10318828

Jaysinh Sagar 10318771

Nikolaas Steenbergen 10333681

Ioannis Zervos 10333630

University of Amsterdam
Faculty of Science
MSc Artificial Intelligence

ABSTRACT

In the fictional setting of the VAST challenge 2012 a virus is attacking the Bank of Money situated in the World of Money. The data supplied for the challenge contains about 158 million entries of the status of every single computer of the bank of money in the range of 2 days.

In this work we provide a description of the challenge setting, the data supplied, our approach to make the data accessible, visualization steps to get insight into the data, and a summary of our findings and observations.

1. INTRODUCTION

“The Vast Challenge is a participation category of the IEEE VAST Symposium, with the purpose of pushing the forefront of visual analytics tools using benchmark data sets and establishing a forum to advance visual analytics evaluation methods” [1]. It is a yearly event.

The Vast challenge 2012 setting takes place in the fictional world of money. This World of Money consists of several nation-states that stretch over different timezones. In the World of money the only bank is the Bank of Money. The Bank of Money has different branches in all of those national states. In the scenario this enterprise falls victim to a virus attack. The participants are supplied with a large dataset containing detailed information about the status of every single machine the Bank of Money operates in the range of two days of the virus attack. The task is to develop a suitable visualization and through this investigate this virus attack, to find anomalies. For this we follow the Keim's Visual Analytics Mantra: “Analyze first, show the important, zoom, filter and analyze further, details on demand".[2]

In this report, we first look at the structure and nature of the data in Section 2. Then we introduce the problem statement and the description of the VAST challenge and the Bank of Money in Section 3. We then look at means and tools used to process the data so that we may work with them with our visualizations in Section 4. In Section 5 we look at the various tools and methods we use to draw up visualizations to help generate insight from the data. Section 6 talks about the observations and findings from the visualized data and finally in section 7 we see the future scope for the project.

2. DESCRIPTION OF THE DATA

For the challenge we receive a large dataset, containing data from the 2.2.2012 8:15 to 4.2.2012 8:00. It consists of roughly 158 million entries. The time supplied by the data is measured in BMT (Bank World Mean Time), so we have to keep track of the local time in which a single machine is operating, as it differs from the time supplied by the data. The data consists of two table [3]:

Meta Table:
IPAddress (ipAddr): This value will be somewhere in the range of 172.1.1.2 to 172.56.39.254, which is the Bank of Money network.
Machine Class (machineClass): This value will be one of "server", "workstation" or "atm"

Function (machineFunction): For Equipment Type "workstation", this will take the value of either "teller", "loan" or "office". For Equipment Type "server", this will take the value of either "web", "email"," _le server", "compute", or "multiple". For Equipment Type "atm", this will take the value "atm".

Business Unit (businessUnit): This value will be one of "headquarters","region-1". . . "region-50". Regions 1-10 are large regions (see table above) and Regions 11-50 are small regions. Facility (facility): This value will be one of headquarters", "datacenter-1 to "datacenter-5", "branch1". . . "branch200". The Business Unit "headquarters" will have the Facility "headquarters and all data centers, datacenter-1 through datacenter-5". The Business Unit for Regions 1-10 will have a facility either the "headquarters" for the regional HQ, or "branch1". . . "branch200" for the branches in large regions, or "branch1". . . "branch50" for branches in the small regions.

Latitude and Longitude: Coordinates of the equipment location in decimal units. Coordinates map to BankWorld.

Status Table:
IP Address (ipAddr): This value will be somewhere in the range of 172.1.1.2 to 172.56.39.254 Data/Time (healthtime): This value will be the date and time (BMT). Please see the BankWorld description for more information about the BankWorld time zones) [3].

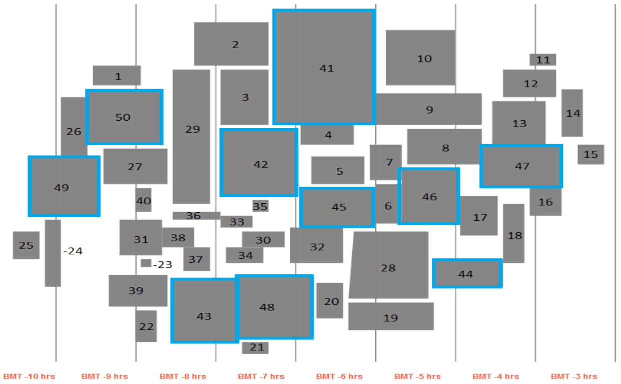
Connections (numConnections): This will be an integer stating the total number of incoming and outgoing connections from a piece of equipment.

Policy Status (policyStatus): This integer's value range between 1 and 5, representing the following:
1. Machine is functioning normally and is "healthy"
2. Machine is suffering from a moderate policy deviation
3. Machine exhibits serious policy deviations and
non-critical patches are failing
4. Machine has critical policy deviations and many patches are failing

5. Machine has a possible virus infection and/or questionable files have been found.

Activity Flag (activityFlag): This integer's value range between 1 and 5, representing the following:
1. Normal. Only normal activity
is detected on the equipment.
2. Going down for maintenance. Machine will be offline.
3. More than 5 invalid login attempts.
4. CPU fully consumed. Machine has been detected as functioning at 100% capacity during this time period.
5. Device has been added. An external device such as a thumb drive or a DVD has been detected on the machine.

The total Bank of Money owns 1 Headquarter, 5 data centers, 1 large regional headquarter in 10 regions, 1 small regional headquarter in 40 regions, 200 branches in 10 large regions, 50 branches in 40 small regions [3]. All offices headquarters and data centers are assigned to geographic regions, which cross Nation-states.



(fig. 2.1)Map of Bank World with numbered regions

3. DESCRIPTION OF BANK OF MONEY

There are certain business rules applicable to the dataset [3]:

Business hours are considered to be Monday-Friday 7am-6pm in each time zone. Since the data set's time entries contain only information about the BMT (Bank World Mean Time), but the computers are situated in different time zones we have to take this into account in our observations and the visualization.

All staff are encouraged to turn off workstations at night. The employees are advised to turn off their workstations during non-working hours. Although as we later will describe (section 6), not all the workstations are turned off, we have to take into account that at least some do not appear in the data any more (if a computer is switched off, it will simply have no data entry in this time slot).

Although Bank of Money engages in planned maintenance, it does not occur on a regular schedule. Normally a computer will flag as going to maintenance, before it goes offline for maintenance.

4. DATA HANDLING

The data files are far too big to be edited or analyzed by conventional spread sheet programs like Microsoft excel or Libre office calc. So we found ourselves in need of a database. A convenient method for this, is to use D3 for visualization and querying the database directly would take too much time, we concluded to the following analysis procedure:
1. make queries in the database
2. preprocess data (e.g. aggregation, min or max of certain values)
3. export the result of those queries to a D3 readable file (csv)
4. visualize data with the help of JavaScript and D3

4.1 Tools

The first and most convenient method for us to create a database was to use python, with the library sqlite3. Sqlite3 creates a local data base as a file in the file system, and can interact with it through sql queries.

4.2 Preparing the database

We loaded both data tables (status and meta) from the csv files supplied from VAST into the local database file. In order to get all information with a convenient query, we then merged the two tables with the status and the meta data (as a sql union over ip addresses). Accessing the data in this stage is still slow. In order to increase the querying speed, we split the data base file into 48 smaller data base files, containing the complete data for 1 hour each. To further increase the querying speed, we created several sql indices (i.e. on location and time). This ensured that the queries could be done in a reasonable amount of time and the size of the database was kept to a reasonable level.

4.3 Database preprocessing and querying

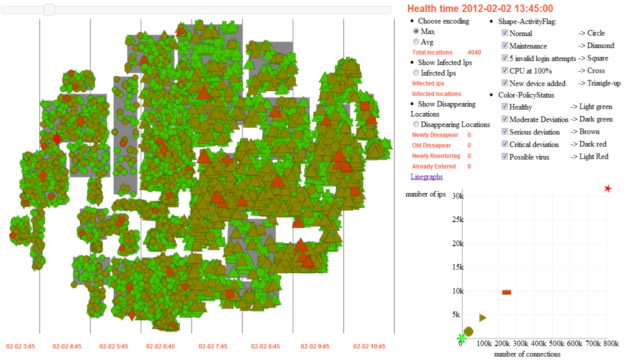
Querying and preprocessing calculations were done in python and the resulting data was written in csv files. Those csv files had to be limited to a reasonable file size as well, as java script and D3 can only handle a certain amount of data (in case of detailed information of each location for example we created 4056 seperate csv files).

5. VISUALIZATION

5.1 Tools

In order to visualize our generated data we used the Data-Driven Documents (D3) JavaScript library [4]. For the interaction with the user we used J Query and our total visualization is build upon HTML. We also used spreadsheets and the WEKA [5] to gain a quick insight of the data attributes.

5.2 Approach


Health time 2012-02-02 13:45:00

- Choose encoding
  - ● Max
  - ○ Avg
- Total locations 4040
- Show Infected Ips
  - ○ Infected Ips
- Infected ips
- Infected locations
- Show Disappearing Locations
  - ○ Disappearing Locations
- Newly Dissapear 0
- Old Dissapear 0
- Newly Reentering 0
- Already Entered 0
- Linegraphs

- Shape-ActivityFlag:
  - ☑ Normal -> Circle
  - ☑ Maintenance -> Diamond
  - ☑ 5 invalid login attempts -> Square
  - ☑ CPU at 100% -> Cross
  - ☑ New device added -> Triangle-up
- Color-PolicyStatus
  - ☑ Healthy -> Light green
  - ☑ Moderate Deviation -> Dark green
  - ☑ Serious deviation -> Brown
  - ☑ Critical deviation -> Dark red
  - ☑ Possible virus -> Light Red

(fig. 5.1)Visualization at a random timestep

Our main visualization consists of a temporal geo-map of the Bank of Money locations with the vertical lines indicating the different timezones. In our visualization we aggregated the almost 890.000 different machines into 4.056 different locations. In this way we can obtain a first impression of the different locations and quickly spot any possible anomalies in the bank network. The size of each shape represents the average number of connections in this location. In the middle, we added some options that will be described later. Because we had to visualize a lot of locations in a limited screen we used 3 levels of sorting. First, the dots are sorted by their policy status and then by their activity flag. In order to make dots with small number of connections visible we sorted them in a way that the smaller ones overlap the larger ones. Above the map there is a time slider, which the user may use to navigate through the different time steps and see how the events evolve through time. Also, the corresponding date and time of each time step is displayed next to the time slider.

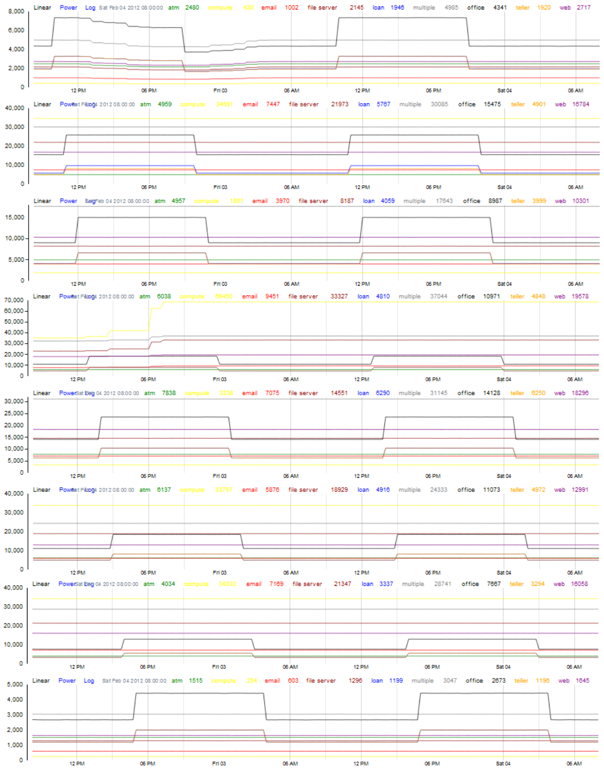
5.3 Interaction

Our interaction techniques were based on the description of the categories of interaction techniques from the paper "Role of Interaction in Information Visualization" by Ji Soo Yi, Youn ah Kang, John T. Stasko, Member and Julie A. Jacko [6]. At a random time step the user may select a location and then a scatterplot of that location will appear where the shape denotes a machine type and the x-axis and y-axis are the number of its connections and the number of its IP addresses respectively. By hovering the mouse over a location additional information of that location will appear.

The user may choose different *encoding methods* for the color-shapes of the locations in the geo visualization. Our first approach for aggregation of the locations was to calculate for each location the average of the activityFlag (shape) and policyStatus (color). Although we know that the policy status is a qualitative variable we did this to represent its most dominant value for each location. With this approach we can obtain a general overview of the situation. However, due to this highly aggregated representation, a lot of information is lost. To overcome the problem of losing specific information, we switch to maximum encoding where at each location, the shape denotes the maximum value of the activity flag and the color denotes the maximum value of the policyStatus of all machines at each specific location. In addition, we added the option "Infected Ips", where the infected machines of each location are displayed as red dots and their size changes according to the number of the infected machines. With that implementation we can see how the virus is spreading throughout the map over time. Finally, with the "Disappearing locations" option we can see where/when entire locations of the region 15 disappear and reappear through time. With that encoding, the newly disappeared locations have light red color, the already disappeared dark red, the newly reappeared light green and the already reappeared dark green. This approach deemed more insightful as we now have information on individual locations and it is now possible to isolate specific locations to pin-point the problem area.

As mentioned above the data in our geo-map visualization are aggregated by location. In order to *elaborate*, a scatterplot appears on clicking on a location. In this scatterplot we can see a detailed representation of the number of IP addresses of each machine type and their number of connections at the specific time step at a specific location. There is no information in the size of the shapes, as a result all the shapes have the same size.

Furthermore, the *filter* interaction is implemented by the filtering menu that is available on the top right of our visualization. With the filtering the user can choose to *filter* the data in the geo-map by either the policy status and/or by the activity flag.



(fig. 5.2) Line graphs of the number of machine types of all the eight timezones through time(top-bottom timezones1-8)

With the line graphs of the eight different timezones at fig. 5.2, the user can explore and observe how the number of each machine type changes through time. The x-axis represents the time, the y-axis the number of machines of each type and the color represents the nine different types of machines, as specified above each line graph.

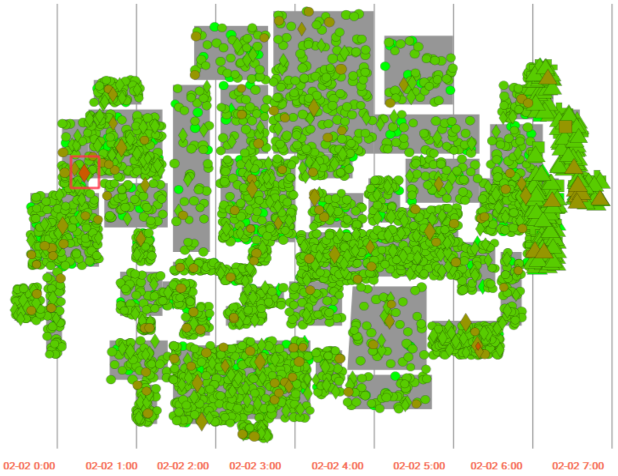
6. OBSERVATIONS AND RESULTS

Through the created visualizations mentioned above, we take a look at what the data has to show us in terms of what is happening with the locations, specific IP addresses and the trends that are observed as we iterate through the visualizations on a time scale. To illustrate this we use several approaches to get a precise interpretation of what anomalies and irregularities exist in this problem set. Although the problem description helps highlight some key concepts to help us understand the nature and rules of the operations of the Bank of Money, yet there is a lot of room for speculation and possible work methods that we see in common everyday banks and how they work.

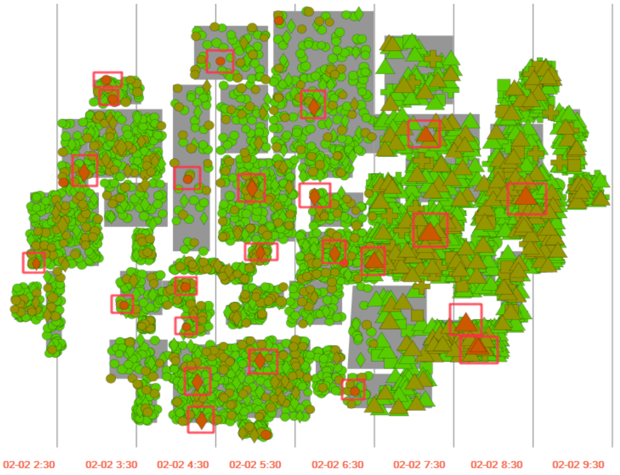
Following Keim's mantra, we use a layered approach where we create an abstract visualization, highlight key areas of interest, expand on those areas, analyse and pinpoint problem spots, zoom into the problematic IP addresses and finally represent those IP addresses individually to pinpoint exactly what is happening. Using this approach, we observed a number of interesting points that shed some light into the problem set:

6.1 Start of Virus infection and Spread:

Navigating through the geo-map visualization, we see that a number of IP addresses show a policy status of 5. Initially this is seen only on 1 IP address (172.2.194.20) but then we see a sharp increase in various locations demonstrating a steep rise in their respective policy status. Over time, we see that the majority of the locations have at least one computer affected with the virus.

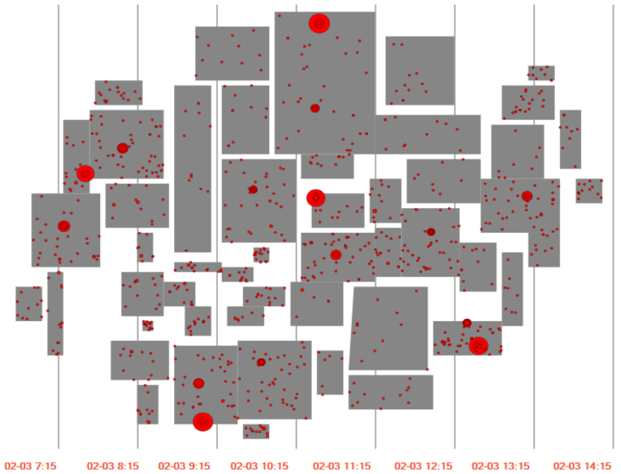


(fig 6.1) Locations of BoM with first infected location highlighted

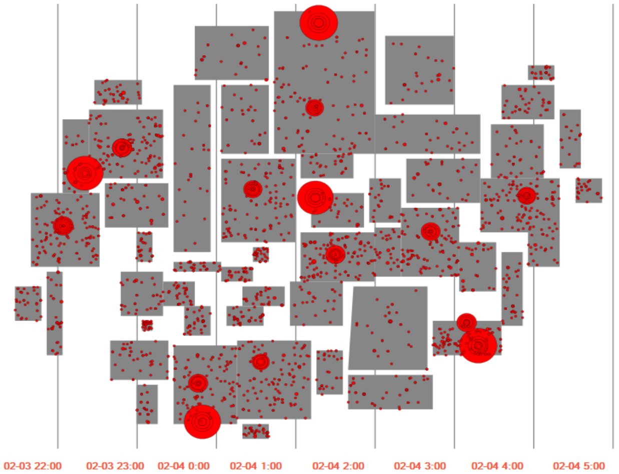


(fig 6.2) Locations of Data centers affected in predominant regions

In fig. 6.1, we see the first location of the infection at IP 172.2.194.20 at 12:45 (BMT) on the first day. What was interesting to see with this IP is that the policy status of the IP starts from 2 from the first entry itself and over time we see a consistent rise in its policy status all the way to 5 which never goes down. After this IP signals policy status 4, we see a lot more IP addresses also reporting policy status 4. Majority of the first infected IP addresses belong to the data centers of each region as highlighted in fig. 6.2. By this we can say that the virus first spreads through the data centres onto other IP addresses over time as seen in the case of fig. 6.3 and 6.4 at a later stage in the visualization.

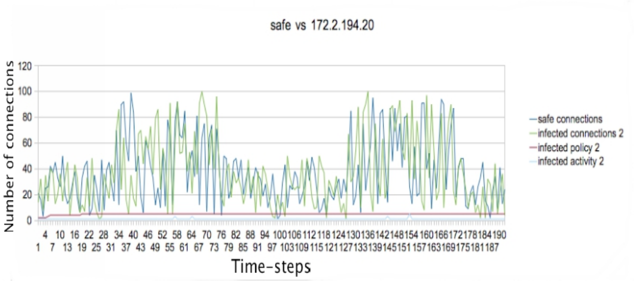


(fig. 6.3) Spread of virus at the midpoint of the time line



(fig. 6.4) Spread of virus towards the end of the time line

Looking at these time-steps of the visualization we can clearly see that there is a clear start to the infection of the virus but we don't see an end of the situation, however it still shows that most computers with the infection are still active and infected.

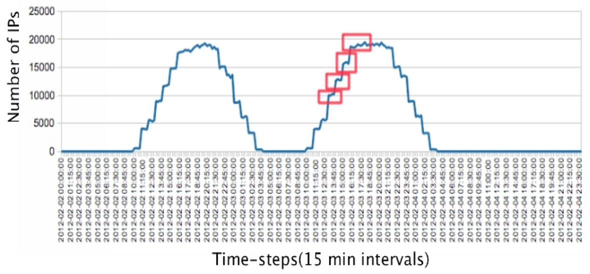


(fig. 6.5) Comparison of the number of connections between the first infected IP and a 'safe' IP at the same location

In fig. 6.5, we compare the number of connections between the first infected IP with another IP that we term as “safe” as this IP never has a policy status above 2 and shows no irregular activity flags. Both these computers are of type compute and at the same location. In the graph we see that there is no abnormal activity taking place in terms of number of connections during working and non-working hours. To illustrate this, we use only 1 'safe' IP although we compared the infected IP with a number of 'safe' IP addresses in the same location of the same type but saw no irregularities. What is interesting to see in this graph is that the activity flag of the infected IP has a series of peaks of invalid login attempts in less than 2 hours and a 100% CPU utilization entry at around time-step 143, however we don't see a significant increase in number of connections at that point as we had hypothesized that possibly a virus infection may lead to a computer experiencing a steep rise in its number of connections to spread the virus especially during non-working hours.

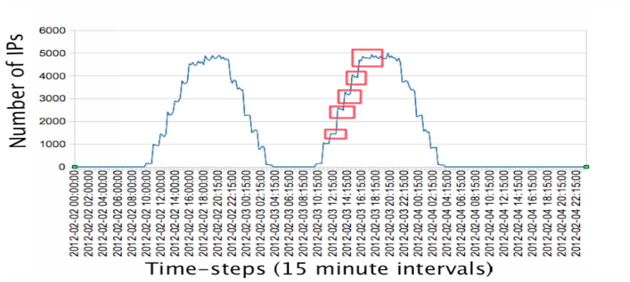
6.2 Global trends of Activity Flags (Invalid login attempts and CPU at 100%):

At this stage in the visualization, we ask ourselves, 'Does the rise in policy status affect the number of invalid login attempts among all the computers ?'. To illustrate this point, we again make use of a simple line graph (fig. 6.6).



(fig. 6.6) Line graph showing the rise and fall of number of IP addresses reporting Activity flag 3 (invalid login attempts) on both days

In fig. 6.6, we see a that on both days, the number of invalid login attempts remain more or less similar during both days of the data set where by the second day, there is a very large rise in number of infected computers as compared to the first day. Although we do see a slight rise as higher peaks on the second day going up as each time zone enters their working hours. These peaks are highlighted in red. Another aspect of this graph that intrigues us is that once all locations are in their work time, it would be natural to see the number of invalid login attempts going down but we don't see this happening. Instead we observe that the graph stays at a rough plateau when all locations are in their working hours. This is especially strange as even the number of IP addresses at the plateau is alarmingly high at about 19000 to 20000 IP addresses.

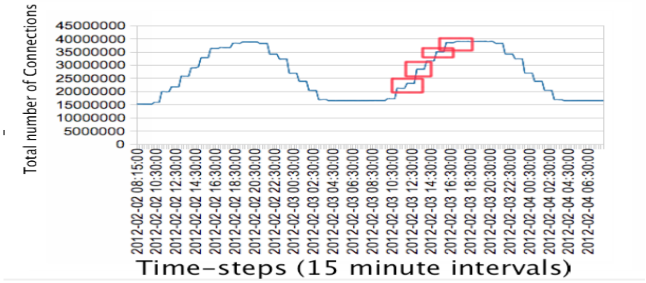


(fig. 6.7) Line graph showing the rise and fall of number of IP addresses reporting Activity flag 4 (100% CPU) on both days

In fig. 6.7, we see a large number of IP addresses with activity flag 4 indicating 100% CPU running. Again in this case, we don't see a significant rise or drop in the total number of IP addresses even after the infection has spread significantly thoughout the Bank World. However we do see a small rise in this count when each time zone enters its working hours indicated in red on the second day. In this case, having a plateau in its peak when all time zones are in working hours as we consider it normal for computers to run at max during working hours.

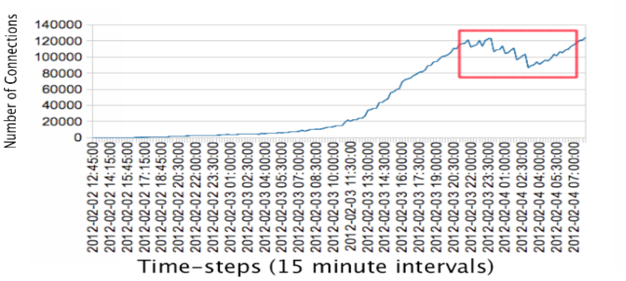
6.3 Virus affecting number of connections:

One question that we hoped would indicate an anomaly is 'Is there a significant rise or fall in the number of connections after the spread of the virus?'. We demonstrate this with the following graph; fig. 6.8.



(fig. 6.8) Line graph showing total number of connections of all IP addresses for both days

In fig. 6.8, we see that there is a slight increase in the total number of connections on the second day for each time zone, however, we had hoped to see a larger difference after the virus has spread. This indicates that there might be a connection between the virus spread and the number of connections but its very loosely coupled and we confirm the earlier point that there isn't a strong connection between the two. We then take only the infected IP addresses and count their total number of connections at each time step to see if there is any indication to fortify our claim in fig. 6.9.

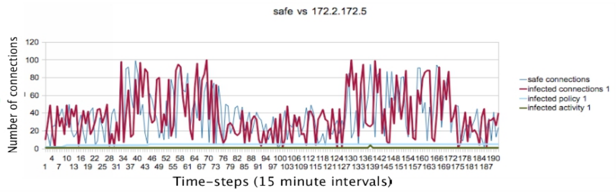


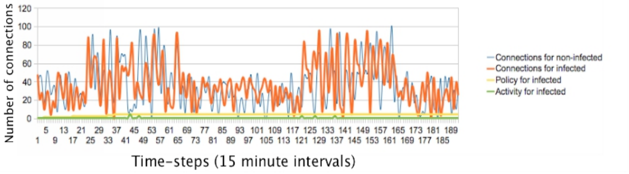
(fig. 6.9) Line graph showing number of connections of infected computers for both days

In fig. 6.9, we see a gradual increase in the number of connections till a certain time-step where the increase becomes rapid. This we believe is due to the rapid spread of the virus which we confirmed from our geo-visualization. What is interesting to see at a certain point when the working hours start ending from one time zone to another, is a significant drop in the number of infected IP addresses connections. This is normal as the computer users may be turning off their computers at night but the zig-zag nature of this fall indicates some unnatural activity. Also, a point to note here is that we don't see this trend on the first day so we can safely conclude that the IP addresses contributing to this abnormal behaviour are not among the data centers' compute and multiple machine types.

6.4 Comparison of number of connections between similar machine types:

To illustrate our point further, we compare infected IP addresses of different types of machines with IP addresses of 'safe' machines.



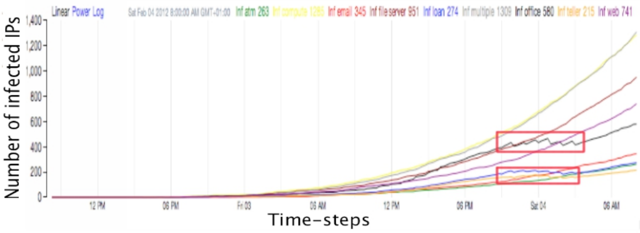


(fig. 6.10 a & b) Line graphs of multiple infected IP addresses' number of connections compared to a 'safe' IP

For both IP addresses in fig. 6.10a and 6.10b, the number of connections don't show a large deviation when compared to a 'safe' computer of their respective types. One observation we did make where an 'ATM' machine type had a large number of connections during non-working hours, we deem that as normal as possibly in a largely populated area, ATM machines would be in use at night. Again here in this case, we see that the infected IP addresses generally have a greater activity compared to 'safe' IP addresses as we can see peaks of the activity flag for the infected IP addresses. This observation may draw interest that generally computers that are used more tend to have a higher probability of being affected by the virus.

6.5 Total number of Infected IP addresses:

In order to understand the extent and spread of the virus to number of computers, we generate a count of all specific infected machine types and plot them.

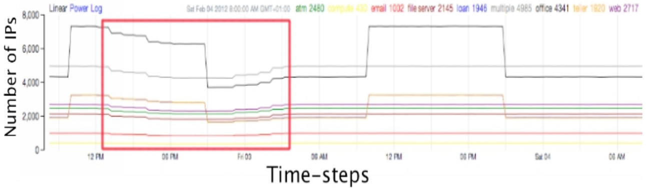


(fig. 6.11) Line graph of number of infected IP addresses of each machine type though-out the timeline

In fig. 6.11, we see that there is a gradual increase in number of IP addresses initially till the start of the second day of the data, but we see a steep rise in the number of infected IP addresses for all machine types concurrently. We have determined earlier that the first machine types to be infected are the multiple and compute machine types, we see the greatest magnitude of increase in the spread of their infection. There is a strong possibility that the steep rise in the number of compute and multiple machine types can affect the spread of the virus to other types of machines that connect to them. An interesting point in this figure is that after a certain period of time, when its non-working hours for the regions, there is a drop in the number of infected IP addresses of type office, loan and teller. This is a natural thing as the staff is encouraged to turn off their computers at night however in this graph as well we see that there is an uneven trend in IP addresses going down. The cause of this is still ambiguous but there might be the reason for the uneven drop in connections we saw earlier in fig. 6.9.

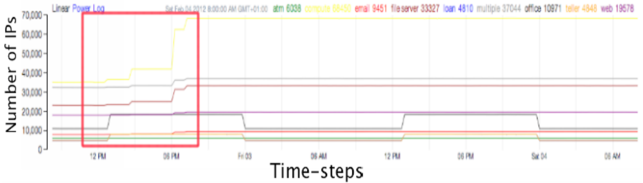
6.6 Disappearance of IP addresses:

A strange behaviour we see in the data is that 35 locations in the Bank World disappear at a certain point during working hours and then re-appear later during non-working hours. Now this may be the cause of some black-out as all 35 locations belong to the same region and the trend of their disappearance goes northwards.



(fig. 6.12) Line graph of Number of machines per machine type as for timezone 1

In fig. 6.12, we see that there is a drop in the number of machines for each machine type in timezone 1. This is seen when we observed that 35 locations are disappearing during work hours in the same region. What is strange is that they come back on at night in a similar order in which they disappear. This anomaly happens only on the first day and it seems normal on the second day.



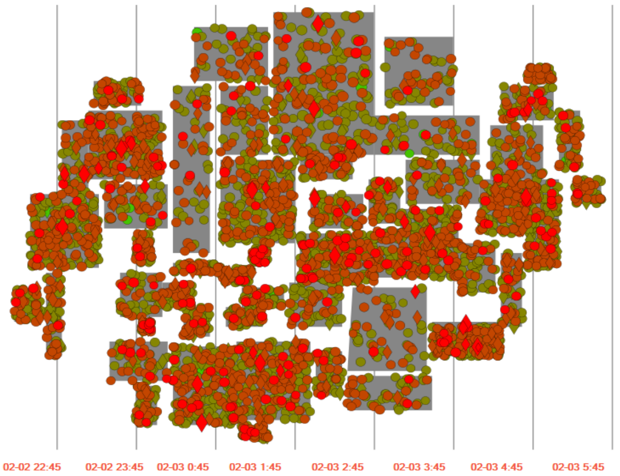
(fig. 6.13) Line graph of Number of machines per machine type as for timezone 4

What seems strange in fig. 6.13 is that there is a very large number of computers of each machine type that come on after half of the working period is over. We still don't know the cause of this anomaly however it is interesting to note that this is an opposite trend of what we see in fig. 6.12 where IP addresses start disappearing but re-appear later. This too possibly may be due to some natural or unforeseen causes but we can only speculate at this point.

6.7 Not many computers are healed after the infection:

We do see a lot of IP addresses infected with the virus however, we don't see many computers being put under maintenance and this reflects the constant upward trend of the increase in the number of computers infected. However we do see several computers of type teller, loan and office being turned off or put under maintenance after the work hours of the second day but the primary cause of the virus which we hypothesize to be the compute and multiple machine types at the data centres and regional headquarters not being treated or turned off at any point of time.

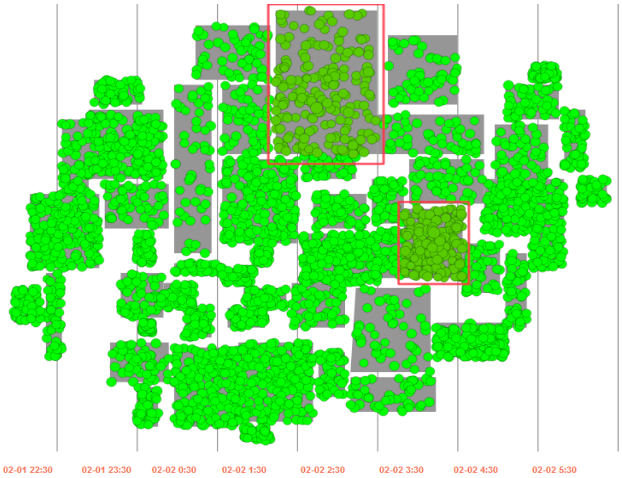
6.8 Lots of computers on at night:



(fig. 6.14) Locations indicating active computers at each location during non-working hours

Contrary to the description of the Bank of Money policy, not many computers are turned off during the night as it is encouraged but not enforced. What is unsettling about this is that we see a majority of the computers being on between the first and second day which may have helped the spread of the virus as illustrated in fig. 6.14.

6.9 Possible cause of spread of virus:



(fig. 6.15) Geo-map of Bank World showing regions 41 and 46 in dark green having higher deviation.

At this point, if we base a hypothesis that most computers working in branch offices or interconnected data centers and head quarters would have to connect to a central system which handles the data of their clients and records of transactions. There is a strong possibility that the virus might have spread to other computers when they establish a connection with the data centers while carrying on their regular transactions. Another point that comes to mind is that these computers might also have to log into a central system in the morning when the working hours begin, these connections might be made to head quarters of each region and that might contribute to the high count of invalid login attempts. A point of interest we come across from the first time step itself is that regions 41 and 46 have a higher policy average compared to all other regions. This indicates that these locations possibly have a higher workload compared to other regions. However the reason for this is still ambiguous.

7. CONCLUSION

Using our developed tools, we are able to base a strong inference on the spread of the virus as well as highlight some key anomalies that we see during the time course of the provided data. However, in our study, the data is unable to suffice the cause of the spread of the infection as we don't see any obvious indications of either the medium of spread nor the cause of the spread. These questions leave a large room for assumptions and hypothesizing on what really happens but what we do see evidently is that the spread is over the entire network of the Bank of Money and that there isn't any fixed pattern in the spread. If the data provided information on process triggers or specific activity of the user of each IP at every time step and possibly the connection map of every IP with every other IP, we may be able to pinpoint who and what process triggers the virus to the extent of drawing a graph tracing the path of the virus from the first incident till the last.

7. FUTURE WORK

A possible approach for further investigation would be the analysis of the 4th timezone to locate the exact locations and the regions where we have a sudden increase in number of all different machine types. Because its pattern reveals a possible anomaly but we haven't found its exact cause yet.

To get a more detailed insight of the data in micro level it would be better to add a zoom-in like functionality in our main visualization in order to elaborate the data in individual machine level. In other words, go from the aggregated data back to the individual machines and explore their instances through time.

8. REFERENCES

[1] "VAST challenge", Available: http://www.vacommunity.org/VAST+Challenge, Created: 10.03.2012, Accessed: 31.5.2012,14:00)

[2] Keim D., "Scaling Visual Analytics to Very Large Data Sets," Proc. Workshop Visual Analytics, June 2005.

[3] "VAST challenge 2012- Mini-Challenge 1 General Information document", Available: http://www.vacommunity.org/VAST+Challenge+2012, Created: 11.04.2012, Accessed: 31.5.2012, 13:00

[4] Michael Bostock, Vadim Ogievetsky, Jeffrey Heer, "D3: Data-Driven Documents", IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011

[5] weka http://www.cs.waikato.ac.nz/ml/weka/

[6] Ji Soo Yi, Youn ah Kang, John T. Stasko, Member, IEEE, and Julie A. Jacko, "Toward a Deeper Understanding of the Role of Interaction in Information Visualization"

[7] Christensen Ben,"Interactive Line Graph (D3)", available: http://bl.ocks.org/2657838 , accessed 15/05/2012

[8] Pino Trogu, "An archive of the Information Design 1 course taught at San Francisco State University",available: http://523informationdesign.blogspot.nl/2011/10/students-d3-graphs-scatterplots.html , accessed 5/05/2012