IEEE VAST CHALLENGE 2009

- HOME
- DOWNLOAD
- TASK DESCRIPTION
- CRITERIA FOR JUDGING

- HOW TO SUBMIT?
- ANSWER FORMS
- RESULTS
- STUDENT SUPPORT

- DISCUSSION BLOG
- HISTORY OF CHANGES

- MINI CHALLENGE 1
- MINI CHALLENGE 2
- MINI CHALLENGE 3
- GRAND CHALLENGE

Detailed Task Descriptions for All Challenges

EVERYONE SHOULD READ THIS FIRST:
Questions? See the discussion blog or send email to challengecommittee AT cs.umd.edu

The 2009 VAST Challenge consists of three Mini Challenges and a Grand Challenge. Contestants can choose to work on one, some, or all of the challenges. To successfully respond to the Grand Challenge, contestants must tie together all data sets with an overall scenario description using data elements from each of the four mini challenges, but are not required to submit to the mini-challenges.

The datasets used for these challenges are synthetic: that is, they are a blend of computer- and human-generated data. All datasets, whether real or synthetic, have anomalies. Some anomalies may be significant, some may not. Any anomalies reported should be supported by the proposed hypotheses. For example, "all first names start with a 'M" may be interesting, but unless it is tied to the discussion of the situation, that anomaly has no place in your submission.

We have included all information necessary to form working hypotheses for the purpose of these challenges. No external data is needed to successfully perform the analysis. Be aware that using additional non-provided data may skew an otherwise successful solution.

The descriptions below provides the details for all the mini-challenges and the Grand Challenge, questions posed in each, and a description of what participants need to provide for answering each question. Each entry (Mini Challenge or Grand Challenge) is required to submit a video demonstrating how you conducted the analysis.

DEFINITIONS
There are different formats/size for providing answers to the questions:

Short Answer:

Short answers are only requested in the mini challenges.
A short answer is a text description of the answer and of how you arrived at the answer. It is limited to 150 words and a maximum of 2 screen shots.

Note about screen shots: Screen shots should be large enough that they are readable when the document is printed (you can always link to the best resolution versions in the document). Do not forget to include the legend for the visual encodings of your screen shots, captions describing what data is being shown, and what filters have been applied in the static figures presented or discussed. In other words, help us understand what we are looking at!

Detailed Answer:

Detailed Answers are requested both in the mini challenges AND the Grand Challenge.
A Detailed Answer is a longer text description focusing on how you arrived at the answer with much more details than the Short Answer.
• For mini challenges, detailed answers are limited to 1000 words, with a maximum of 5 screen shots and captions.
• For the Grand Challenge there is no size limit (but less than 5000 words is recommended with a maximum of 15 screen).

Detailed answers should provide the answer and describe in detail the PROCESS USED TO ARRIVE AT THE ANSWER. Clearly describe what was done manually or automatically, what you saw in the displays that helped you formulate or prove your hypotheses. For example don't just say "we used advanced technique X to easily see who is involved", instead be specific: tell us HOW you can see who was involved? E.g. something like "we suspected that Joe was involved because his name appeared in red when color was mapped to the number of oversee travels and he stood out as being outside his family cluster". Describe the process you used (what you had to do to load the data, how you decide to start the analysis. what worked and what didn't work, how you recorded your findings, dealt with uncertainty etc.) Make very clear what was accomplished manually, automatically, or in between. Provide estimates of the effort required (e.g. in 4 hours three team members read all papers written by Paul White; or we wrote a script to do XYZ in 30 minutes, or We found the name of the architects in about 20 minutes by filtering from January to march using the range slider, then follow all international links (colored blue) and copy and paste the new names into a shared text document

Video:

All submissions are required to include a video with voice narration.
Maximum length (shorter is better):
• 4 minutes for Mini Challenge entries
• 15 minutes for Grand Challenge entries.
NOTE: If you submitted an entry to all three mini challenges you already have three videos for them. You may reuse all or parts of the 3 videos but should also leave enough time to show how you integrated the multiple datasets and come up with the grand challenge answers.

Voice narration and good usage of the mouse to point at relevant elements of the screens are essential. Simple titles help structure the video to show steps in the analysis process. Generic advertisements of the tools' features are not useful, instead you demonstrate specifically on how you worked with the data to understand the situation and answer the questions e.g. how you interacted with the tools, what insights were revealed by the displays, how you gathered evidence to support or refute hypothesis, etc. You can show how you started your analysis, then jump to the end to see the final results.

Debrief:

Debrief are requested only in the Grand Challenge. The debrief is basically the analytic product that a professional analyst would deliver after doing the analysis.

A debrief is a maximum of 2000 words narrative describing your hypothesis about the situation at hand. Include in your narrative the relationships of the various players. If there are uncertainties, you can suggest possible next steps to clarify those uncertainties. In the debrief you should distinguish between the facts located in the data and your interpretations about the synthesis and meaning of these different pieces. Do NOT describe the tools used nor discuss the process used, instead focus on convincing us that you UNDERSTAND the situation.

See an examples of good debrief from the 2007 contest:
http://www.cs.umd.edu/hcil/VASTchallenge08/sampledebrief.htm

Two-Page Summaries:

Two page Summaries are OPTIONAL, they do not need to be submitted until after the results have been announced. They appear in the printed materials of the Symposium and also archived online.

These summaries allow you to give a general overview of your tools, significantly highlight novel features, provide references to papers and other relevant work and describe any new discoveries you made about your tools while working through the Challenge problem. Only the two-page summaries of the best entries (which are awarded an award) will be published in the VAST 2008 Symposium Proceedings. Nevertheless, ALL submitted two-page summaries will be published online - along with your answer - in a repository at NIST (the National Institute of Standard and Technology), whether or not they are awarded a Certificate of Excellence.

The two-page summary should be formatted according to the general IEEE VGTC Guidelines
http://www.cs.sfu.ca/~vis/Tasks/camera.html

CHALLENGE 1-BADGE AND NETWORK TRAFFIC

An embassy employee is suspected of sending data to an outside criminal organization from the Embassy.

We are providing two data sets to analyze. The first is a proximity (prox) card log. Prox cards are badges with electronic tags assigned to each Embassy employee. The cards are used to access the Embassy and areas of limited access within the Embassy. Each data record contains an employee number, prox card number, date/time of use and location of use. The data consists of logs covering a period of one month. The second set is a month's worth of network traffic logs. Each employee has been assigned a desktop computer with a static IP address for use in their daily duties. The network traffic log data consists of the computer IP address, the employee number of the assigned user, outgoing and incoming activity from the computer including destination site, payload (request and response data) and port number.

Table 1: Example Prox Card data. The embassy is outfitted with prox card readers on the entrance to the building as well as the door to the restricted area inside the embassy. Employees generally badge into the building (i.e., actually present their badge to the badge reader) but may occasionally go against policy and piggyback (enter the building without badging in by following a coworker who did badge in). However, employees are required to prox into and out of the restricted area and no piggybacking is allowed. No records are kept of employees leaving the building. The data consist of a CSV file with values of the event datetime, the employee id, and the type of event (prox-in-building, prox-in-classified, prox-out-classified).

User Warning	prox-in-building	employee ID	event
Synthetic Data	2008-01-01T08:03	51	prox-in-classified
Synthetic Data	2008-01-01T08:05	29	prox-in-building
Synthetic Data	2008-01-01T08:06	51	prox-out-classified
Synthetic Data	2008-01-01T08:08	2	prox-in-building
Synthetic Data	2008-01-01T08:11	23	prox-in-building

Table 2: Example IP Traffic data. Employees of the embassy have static IP addresses on their unclassified machines. Traffic on these machines is routinely monitored in case an employee is suspected of non-governmental use of their machine (reading too many chinchilla blogs). The data contains the sizes in bytes of the request (called request payload) and the response (called response payload), the port, the source IP address and the destination IP address.

USER WARNING	SourceIP	AccessTime	DestIP	Socket	ReqSize	RespSize
Synthetic Data	55.170.100.11	2008-10-01-08:02:54:985	242.82.167.209	80	1174	5370
Synthetic Data	55.170.100.20	2008-10-01-08:03:14:911	202.182.69.85	80	950	5302
Synthetic Data	55.170.100.32	2008-10-01-08:03:17:980	88.244.136.106	80	243	5478
Synthetic Data	55.170.100.32	2008-10-01-08:03:26:540	201.10.130.54	80	630	2945
Synthetic Data	55.170.100.18	2008-10-01-08:03:29:424	191.201.20.153	80	1197	5682
Synthetic Data	55.170.100.22	2008-10-01-08:03:31:445	55.170.30.100	80	225	6613

Questions/Tasks:

MC1.1 Identify which computer(s) the employee most likely used to send information to his contact in a tab-delimited table which contains for each computer identified: when the information was sent, how much information was sent and where that information was sent.

Please name the file: Traffic.txt

A sample answer would look like this if you think 2 computers were used, each one time. (Practically you just need to cut and paste the rows of the data table which are the evidence for your hypothesis).

USER WARNING	SourceIP	AccessTime	DestIP	Socket	ReqSize	RespSize
Synthetic Data	55.170.100.11	2008-10-01-08:02:54:985	242.82.167.209	80	1174	5370
Synthetic Data	55.170.100.20	2008-10-01-08:03:14:911	202.182.69.85	80	950	5302

MC1.2 Characterize the patterns of behavior of suspicious computer use. Provide a Detailed Answer.

Provide a video showing how you conducted the analysis (one video per challenge entry, mini or grand).

CHALLENGE 2-SOCIAL NETWORK AND GEOSPATIAL

Embassy employees are known to have use the social networking/micro-blogging tool, Flitter, to communicate with colleagues and friends. The Flitter network may provide a connection to a criminal ring that may have recruited an employee. We have been provided with Flitter data that we may analyze.

There will be two tab-delimited tables, one describing entities (i.e., either a user-name, a Flitter nickname and not the person's real name, or a city or a country) and one containing links. As we have only user-to-user connection information, please consider these connections as two-way links.

We also provide a map of Flovania, its major cities, and information about neighboring countries and their major cities.

Detailed information: A table of entities (listing all names and countries) is provided:

ID	Name	Type
INT	STRING	STRING
1	@Arthur	person
2	@Jerry	person

3	Kannvic	city
4	Flovania	country

And a table of links:

ID1	ID2
INT	INT
1	3
2	56
3	2
4	67
10	2
11	21

The third data set is the map provided as a jpeg file.

Part 1: Social Network portion of the mini-challenge: We believe an employee communicated with his/her handler(s) (a contact from the criminal network) through Flitter, however we do not know the Flitter name of the handler nor of the espionage organization. We believe that the associated network may take one of two forms of social structures:

A. The employee has about 40 Flitter contacts. Three of these contacts are his "handlers", people in the criminal organization assigned to obtain his cooperation. Each of the handlers probably has between 30 to 40 Flitter contacts and share a common middle man in the organization, who we have code-named Boris. Boris maintains contact with the handlers, but does not allow them to communicate among themselves using Flitter. Boris communicates with one or two others in the organization and no one else. One of these contacts is his likely boss, who we've code-named Fearless Leader. Fearless Leader probably has a broad Flitter network (well over 100 links), including international contacts.

B. The employee has about 40 Flitter contacts. Three of these contacts are his "handlers", people in the organization assigned to obtain his cooperation. Each of the handlers likely has between 30 to 40 Flitter contacts, and each probably has his or her own middle man in the organization, who we've code-named Boris, Morris and Horace. It is probable the middle men will not allow the handlers to communicate among themselves using Flitter. Each of the middle men probably communicate with one or two others in the organization, and no one else. One of the contacts for all of the middle men is the head of the organization, Fearless Leader. Fearless Leader has a broad Flitter network (well over 100 links) including international contacts.

Questions/Tasks (part 1):
MC2.1 Which of the two social structures, A or B, most closely match the scenario you have identified in the data?

Please provide your answer: A or B

MC2.2 Provide the social network structure you have identified as a tab delimitated file. It should contain the employee, one or more handler, any middle folks, and the localized leader with their international contacts. What are the Flitter names of the persons involved? Please identify only key connections (not all single links for example) as well as any other nodes related to the scenario (if any) you may have discovered that were not described in the two scenarios A and B above.
Please name the file: Flitter.txt
The Role can be any of the following: Employee, Handler, Middleman, Fearless Leader, Leader's International Contact, Related Other. (The "Related Other" corresponds to nodes that are related to the scenario but do not fall into any of the other categorizations.) Note: The fewest nodes capturing essential information is better.

ID	Role	Filter Name
100000	Employee	@Loura
100001	Handler	@Georges
100002	Handler	@Mark

100003	Middleman	@Catherine
100004	Fearless Leader	@Jean
100005	Related Other	@Heather
100006	Leader's International Contact	@Jereme
...	...	...

MC2.3 Characterize the difference between your social network and the closest social structure you selected (A or B). If you include extra nodes please explain how they fit in to your scenario or analysis. Provide a Detailed Answer.

Part 2: Geospatial portion of the mini-challenge: Please see the additional dataset linking Flitter IDs to Cities.

In addition to the above, the two social structures have geospatial implications. While a target and handler may be in a large city, a middleman might be in nearby smaller locations. A leadership role, such as the one of Fearless Leader, would likely require a presence in a larger city.

Questions/Tasks (part 2):
MC2.4 How is your hypothesis about the social structure in Part 1 supported by the city locations of Flovania? What part(s), if any, did the role of geographical information play in the social network of part one? Provide a Short Answer.

MC2.5 In general, how are the Flitter users dispersed throughout the cities of this challenge? Which of the surrounding countries may have ties to this criminal operation? Why might some be of more significant concern than others? Provide a Short Answer.

Provide a video showing how you conducted the analysis (one video per challenge entry, mini or grand).

CHALLENGE 3-VIDEO ANALYSIS
We suspect that at least one, perhaps more, meetings of persons associated with this case took place at locations captured by this security camera.

Questions/Tasks:

MC3.1 Provide a tab-delimitated table containing the location, start time and duration of the events identified above.

Please name the file: Video.txt

Use the location numbers shown in this EXAMPLE in column 1.

Location	Start Time	Duration
Location1	1:01:01	05:52
Location3	3:12:18	10:12
...	...	...

MC3.2 Identify any events of potential counterintelligence/espionage interest in the video. Provide a Detailed Answer, including a description of any activities, and why the event is of interest.

GRAND CHALLENGE
We provide several additional pieces of information to assist you in summarizing the activities of the employee and the criminal organization. First, we give you a list of IP address of machine in the embassy mapped to staff IDs. Second we give a list of Prox card IDs mapped to staff IDs. (Note: we do not give names as announced earlier)

Questions/Tasks:

GC1. Please describe the scenario supported by your analysis of the three mini-challenges: Provide a Debrief.
GC2. Who are the major players in the scenario and what are their relationships? Provide a Detailed Answer.