ABSTRACT

Three members of B17 travelled to Ohio to attend the VisWeek 2008 visualisation conference. This conference, run by IEEE, is really three conferences in one:

- Visualisation (Vis)
- Information Visualisation (InfoVis)
- Visual Analytics Science and Technology (VAST)

This paper highlights relevant papers containing new ideas, useful techniques or tools for Sigint visualisation research at GCHQ.

Overall the conference is *very* relevant to visualisation research at GCHQ, and we would recommend a similar level of attendance next year. We should also consider attending the European version, EuroVis, or perhaps the CHI (Computer Human Interaction) conference, which also covers Human Factors.

Soft copy location

[redacted]

[redacted]

For additional copies of this document or for general queries please contact:

[redacted]
B17 – Data mining Applied Research
Government Communications Headquarters
[redacted]
United Kingdom

[redacted]

1. CONTENTS

2. KEY POINTS FOR VISUALISATION WORK AT GCHQ

Palantir Government	Palantir Government is a very significant and impressive commercial product, a complete "analysis and reporting solution", similar in scope to Security Service's IE domain, but in many ways much more impressive and extendable. See Palantir Government tool demo.
Building up an analysis picture	Missing from MONTE VISTA is an “analyst notebook” capability – where the analyst can keep track of the progress of their investigation and build up their intelligence report, and annotate data with notes from the analysis. Several presentations showed the importance of this (e.g. EntityBasedCollaborationTools, GraphicalHistoriesForVisualisation). This also looked to be a major feature of Palantir. This could be a major new thread of the VisFus research task or new feature set in MAMBA.
Collaborative analysis	Several presentations showed the power of allowing multiple users to collaborate on a piece of analysis. (e.g. EntityBasedCollaborationTools, CollaborationSynthesis). Related to the point above, this could be a major new thread of work for Applied Research.
Linked views at different levels of detail	Use of linked views that are at different levels of detail seems quite powerful: e.g. ExplorationOverviewAndDetail. A practical thing to try in MONTE VISTA would be 2 graph views. If you click on an aggregate link in one, the expansion into its components is shown in the other.
Graph layout algorithms	Lots of ideas relating to graph layouts: • Try constrained graph layouts (e.g. Cerebral) • Try edge-grouping techniques (e.g. GeometryBasedEdgeClustering). Might be useful for either the graph view or the geo-temporal view. Also for very large (bulk data) type graphs. • Another interesting point was that we could different layout algorithms for different sizes of graph (e.g. for very large graphs: RapidGraphLayoutUsingSpaceFillingCurves)
Large-graph visualisations	Several techniques were aimed at allowing you to understand the structure of very large graphs. The Visual Fusion team should discuss whether any of these techniques would be useful to the profiling team (e.g. GeometryBasedEdgeClustering, VisSocialAndScaleFreeNetworks, RapidGraphLayoutUsingSpaceFillingCurves). Could we add more views that would be useful for profiling/EDA researchers into MONTE VISTA? Also, several presentations had 2 linked graph views one “micro” and one “macro” on the screen at the same time.
Interesting new views we could try in MONTE VISTA	• Cross between a matrix and a graph view to really highlight clusters in the graph: SocialNetworksWithNodeDuplication (does have problems though). This might work well with dense parts of a graph – e.g. recipient-to-recipient relationships in CHARTBREAKER. • Tree maps (maybe not many applications?): SpatiallyOrderedTreeMaps • Scatter plots with transitions (RollingTheDice)
Windowing graph work	• Paper on how to explore massive time series quickly enough to maintain interactivity: MaintainingInteractivity – architecture for visualising window on very large data set. Looked relevant to windowing semantic graph work in VisFus task.

Other relevant algorithms / techniques	• Clustering by trajectories (useful for PIGS EAR style profiling): VisualClusterAnalysisOfTrajectoryData • Clustering sets of ranks: IncompleteAndPartiallyRankedData • Aggregating and visualising routes or trajectories: SpatioTemporalAggregation
Occulus geo-time	Tool for showing maps or node-link graphs in two dimensions, and time in the third dimension. Recommend we get involved in ATS evaluation of this tool – it may be a generally useful tool within B17 for exploratory analysis of graph or geo data. See ConfigurableSpaces.
VAST 2008 Challenge	The VAST 2008 Challenge invited teams to submit visualisations to solve challenge 4 problems sponsored by the US Intelligence community. The synthetic data sets were very relevant to GCHQ and Many teams contributed. It would be a good idea for the VisFus research to spend some time looking through the contributions and extracting the best ideas.

3. GENERAL VISUALISATION LESSONS

1. This table gives some general lessons about visualisations that we picked up at the conference.

Level of detail	Be able to present the information at any level of detail, to enable overview and detailed inspection. Moving up and down the scale may imply more than just re-scaling the view – for example, clustering at the higher level of abstraction.
Context and detail	Be able to get a feel for the overall shape of the data then dive in to see the detail (and vice versa). Critically, this switch should be made seamlessly, so that there is no break in the user's perception of what is happening and they can build a mental map of where they are and where they are going. Animated transitions can assist here.
Smooth Transitions	These aren't just pretty animation – conveying the flow, or linkage between two states can be very powerful.
Query = Visualisation = Result	Try to integrate asking questions, getting results and visualising the results altogether as a framework. It aids the exploration process because the user does not have to change mode each time. KNIME might be considered a reasonable example of such.
Save time not the world	It is highly unlikely that a visualisation tool can produce a fantastic insight every time it is used. However, enabling users to do their job faster is a massive benefit to the organisation and much more easily achievable. Therefore, the aim should be on user productivity and *keeping it simple*, not creating a massively complex system that may just find the needle in the haystack (but probably won't).
Data objects = domain objects	The metaphors used in the visualisation should be recognisable to the user as entities from their problem domain and not abstract data structures. This helps the user build their mental model and interact with the system much more naturally. Use organisational knowledge to create the right domain objects.
Topology	Related to the previous item about domain objects, this is about trying to use a reasonable "map" of the problem space as part of the visualisation. It might be a real map or just a logical breakdown of the domain. The topology helps the user shape the analysis and allows them to place concepts in a natural way. Use organisational knowledge to create the right topology e.g. a biological cell seen in cross section.
Collaboration	Visualisation/analysis systems can enable to people to work together in ways never possible before. Need to absolutely maximise the massive potential here.
Workspaces	Giving users a personalisable space to work in enables them to externalise and structure their private thoughts. Having shared workspaces allows collaboration, and a good link between the two allows the continuous cycle of private thinking and shared work.
Roles	It seems highly likely that when a team of analysts are working together on a complex problem their work will break down into different roles and this needs to be taken into account.
Stripy Teams	Build user interfaces in multi-discipline teams to get the best combination of the tech and domain knowledge. Learn with actual users on *real* problems not toy ones – that's where the limitations of tools are exposed.

Toolsets not applications	Requirements and analysis strategies change quickly, so better to have a set of tools than a hard-coded application.
Analysis, Communications, History, Stories and Insights	The analytical process is not a simply a case of the user absorbing data until an analysis is formed. Rather it is a complex, ongoing process of exploration where hypotheses are tested, searches are performed, ideas are externalised and eventually communicated to others. Therefore our tools should aim to support the whole process rather than just aim to convert data into pictures, which is a tiny subset of the process. Especially important are the history of the work and the story that the analysis results are trying to tell.
Report = Analysis	Do not force the user to do their analysis and then try and hack a pale, static, deficient copy of it into power point. The *whole* analysis should be the communication, and the report.
Support Iteration	Be able to move through the query/result cycle iteratively, to support the way that follow-up questions naturally arise.
Semi-structured data capture	Lack of structure gives freedom to the user, but does not enable sharing or knowledge capture. Semantic schemas avoid those pitfalls but impose a heavy burden on the user, and struggle where the data is imperfect. There is probably a middle ground which is the best of both worlds.
Comparison	Be able to see and compare two versions of a particular data set *at the same time*. Changing the parameters and redrawing the UI does not enable humans to do comparison (e.g. some of our tools). Automated comparison highlights can assist here.
History and Re-find	Be able to go back and forth through the iterations of analysis and start a new branch at whichever point.
Visualisation is not the data	The visualisation has to represent the data in the best way possible, but perhaps we too often try to represent the exact data structures as is.
Visualise something interesting	Try to show the relevant underlying causes for effects as well as the effects themselves. Predict what would happen if those drivers were changed. This is what leads to concrete actions.
Aesthetics aid understanding	The use of colour, shading and clever presentation is not just polish – it can add significant value to understanding the work.

8. PALANTIR GOVERNMENT TOOL DEMO

3. This was so significant we have created a separate wiki page: [redacted]

4. The wiki page also contains videos and presentations about the product. The following text is taken from the wiki page.

5. In summary, from the demo and questioning, Palantir looked to be an extremely sophisticated and mature system - a complete "analysis and reporting solution", similar in scope to Security Service's IE domain. B17 are in a good position to judge the system, due to related recent work on MONTE VISTA, and we were *very* impressed. You need to see it to believe it.

6. The tool has a very polished "thick-client" user interface with multiple views (graphs, tables, geo). It is supported by a scalable back-end server architecture (90% of the code) which federates to customer databases, and stores working data sets in a fused semantic graph model. In many respects (such as analysts working on collaborative projects, the backend data federation system and tools for working with entities extracted from text documents) it would take an enormous effort for an in-house developed GCHQ system to get to the same level of sophistication. Unlike other systems in the same field (i2 or IE Domain), it has open APIs which allow customers to easily plug in their own Java software.

7. Clearly, adopting this commercial system for visual analysis and reporting would conflict with the current plan in Better Analysis Agility to develop our own desktop integration framework based on Eclipse in collaboration with second parties. However, we feel it is important that the System Engineers in Better Analysis are aware of the product and its feature set.

8. Links to videos and reports are below. The VAST challenge papers and video give a good overview of the tool, and include screenshots and worked analysis examples of an interesting, but fictional, intelligence data set from the VAST 2008 challenge.

Company Background

9. Palantir are a relatively new Silicon Valley startup who are sponsored by the CIA. The company was started as an offshoot of the team developing the fraud detection system for Paypal. They claim to have significant involvement with the US intelligence community, although none yet at NSA. They have approximately 150 employees in the USA (with a current policy of US citizens only, presumably because of clearances, so no Europeans yet) and are looking to double this number over the next year. They sponsored the IEEE VisWeek 2008 with the aim of recruiting some good people during the conference.

10. Their largest customer is somewhere been 100 and 1000 users, but likely in the 100-200 mark. They claim to be keen on getting more customers and adapting

[redacted]

the platform to new things - but how this might play out is unclear (eg if X agency with 100 users buys in, and Y agency with 10000 users buys in then which ones features get developed first?). There was a throwaway comment at the end of their VAST competition presentation about wanting to try new areas outside Finance and Intelligence - this might lead to abandonware.

Palantir Government

11. Palantir Government is a sophisticated integrated analytics platform. It provides a very rich Swing based GUI, and potentially a thin client too. Some instances of this are apparently forward deployed to US interests, and the tool has been developed closely internally with intelligence community users (unspecified, but likely to be the CIA given the funding) over the last two years.

12. The platform is developed on 1 month internal cycles, with every third internal release being a public release. Must-have functionality or fixes can be delivered as hotfixes to clients typically within days (or faster if the business case demands).

Platform

13. It is a complete open (but proprietary) platform. They give full API details for their backend API, and every call available within the backend is available within the GUI APIs. It is possible to completely replace their GUI product with a custom one if desired, although this seems pointless given the richness. Their aim is to ship a template product plus a platform to develop things further - and they expect customers will do this to add support for their business.

14. The API exposes a number of points for integrating tools - either as clients that can request / manipulate data themselves (e.g. possibly for mining algorithms), or as GUI plugins that sit inside their Swing tool. This allows complete new views of the data, similar to how this is done in MONTE VISTA, to be created. The server connections operate over HTTPS.

15. There is substantial Microsoft Office integration, including some sophisticated export to PowerPoint (see "History" below)

16. Everything is fully Unicode compliant, and we saw data in Arabic alongside other scripts.

17. Schemas are possibly based on OWL from what is said in their literature. There is a concept of a dynamic ontology - their videos explain this better, but it is essentially the same concept as the MONTE VISTA "Semantic Model", but with the ability to add and remove types at runtime (as long as they aren't in use).

Desktop system requirements

18. This depends on the deployed functionality:

19. The financial sector and advanced government sectors have dual / triple head machines, multiple core machines, high powered graphics and loads of memory.

20. At the other end of the scale some intelligence agencies have slimmer deployments but with most of the capability (processing is largely serverside) running on single 15" monitors on a single core machine. Obviously it gets harder to see relationships between larger amounts of data / views in this configuration.

Views

21. Currently a couple of views are provided - a simple tiled imagery geo view and a graph view. However, there are also lots of controlling panels / subviews in the system, such as histogram / line plots, as well as providing attribute / aggregate counts for drill down / graph element selection. The line/histogram plot was very similar in concept to the valuebar in MAMBA but perhaps a little more comprehensive.

22. Data from one view can be visualised in another view by drag and drop into that other view - this was a deliberate design decision from working with users as it allowed them to work on something in detail in one view while foraging for data in another view, then manually add that new data into their other working view.

23. There was no automatic synchronisation of data between views in this release, but it was being added in the next major drop as a toggle option.

24. The graph view had a very powerful "split links to bipartite" function, which looked like it could be quite powerful. It also allowed merging of entities, eg adding a telephone number as a logical attribute of a person by simple drag and drop of one entity over the other, followed by a wizard.

25. Brushing and dimming were supported in all views, giving a good indication of data in the context of the wider picture.

26. KML export was provided, which appeared to be via network link. This meant that changes in the data to be displayed in the geo view could all be managed within the tool itself, and google earth just became a simple viewer. We didn't see anything as sophisticated here, but this was relatively new capability.

27. The graph view had the ability to overlay "flow" information - ie a red pulse travelled down the links between nodes, which is useful to see how data, money etc are flowing around a graph / social network (essentially this is adding another complex dimension onto the visualisation without overloading existing metaphors such as link direction, colour, labels etc). This was really powerful, as the movement really stood out against the static graph and background.

Data import

28. It is possible to set up persistent stored queries against datasources (eg the RAPTOR federator) which can return data as it appears on the underlying systems.

29. There are four main import mechanisms:

- Copy and paste as document - this creates a new document (file) entity within the graph and immediately allows tagging to take place to extract information as other nodes/links. This is very polished, intuitive, and easy to use.

- Open existing document, e.g. .doc, .txt, .xls - for structured formats this is quite sophisticated and uses a simple bayesian method to learn the most likely field mappings into your semantic schema. Can also load their interchange xml format.

- Direct JDBC connection to your database - you define a mapping between fields and attributes / types.

- RAPTOR / federation backend - you provide semantic mappings and connections on the backend and it queries your stores.

Project / Investigations

30. Current work is stored in investigations (could be thought of as similar to a project filtered graph, or an i2 chart of stuff known for a particular operation).

31. Each investigation is a subset of a larger backend semantic graph, and these are stored on the servers (as is all data) and shared on a publish / subscribe model. It is possible to essentially fork investigations to support multiple users going in different directions, or collaboratively work together (though we didn't see this).

32. Each investigation conforms to the security model in the system, and users that don't have the right credentials wont see those entities or relationships in the graph they do not have permission to see.

33. It is also possible to "export and lock" investigations (or fork, then export and lock) to allow users in the field or on poor network links to carry on working on a standalone workstation with new data. The standalone version is currently limited to 4GB of data as it uses Oracle XE as the backend database in the absense of the network connection

Analytics

34. Some basic graph analytics are provided in the tool as standard, and we saw things like centrality, shortest path between nodes, etc.

35. A more sophisticated plugin was in development for a customer and this provided much more in the way of social network metrics.

History

36. The system tracks all changes to data through an "online history". This allows very rich auditing, but also allows users to undertake multiple lines of enquiry. This was very similar to some of the visual history talks that have been presented over the years at VisWeek, but was much more friendly in that it doesn't delete the redo histories when a different track is taken.

37. The history stores "what we knew when" - ie snapshots of the current knowledge - and can be used to automatically generate Powerpoint or export to i2 Analyst Notebook charts.

Structured Text Extraction

38. The platform can connect to any of the major entity extractors for automatic tagging of imported documents. There is also a very sophisticated and easy to use interface for manually tagging data and (really cool) linking data together (eg associating a telephone number with an identity). Phantom entities can automatically be created, eg when creating a telephone number this can be associated with a "dummy" person or entity and the rest of the details added as they are discovered.

Backend details

39. The backend stores things as a semantic graph, similar to MONTE VISTA, and is backed with Oracle. They do not use RDF for performance reasons. The rest of their backend runs on a MapReduce architecture to allow rapid and huge scalability, and they charge on a per-core basis.

40. There are three main types of server involved in Palantir:

- RAPTOR federator (for data import from customer systems - this is an analytics platform, not a bulk store)

- Versioning server (for the investigations)

Search and the RAPTOR federator

41. The RAPTOR federator is a query mechanism that can connect to legacy / corporate stores and analytics and query them for data. The returned data is adapted into the graph representation by a set of mapping functions very similar to the transformer/adapter mechanisms in MONTE VISTA and Eclipse.

42. This would be one way that users could easily access data from BROAD OAK, HAUSTORIUM, SALAMANCA, and IIB etc. RAPTOR can be scaled horizontally by adding more boxes / CPU licences.

43. Data ingest goes through a custom data validation and transformation mechanism for each source, where data can be transformed into approximations for faster fuzzy searching later.

44. Data can be searched with "metaphone" approximations, or any other "approx" method you choose to add to the search and import facilities. This was quite powerful, but not as powerful as the B14 / NSA method of full phonetic space edit-distance matching.

45. There is a "google for knowledge" box - simply type something in a-la google and it goes against the current graph, backend store and and the raptor federator to find matches to your query and then import them.

Data model / security

46. The data model supports custom metadata, which works at both the entity and the attribute level. This allows tracking of security information, as well as the source of any information that is imported, entered or adapted within the tool.

47. The security mechanism is pluggable, to work with most standard security schemes (eg it ships out of the box working with Microsoft Active Directory, but can easily interface with many PKI products).

Licencing

48. US export regulations apply, and UK intelligence would have. Applied Research contacts have indicated that their evaluation of the product was not really dropped because it was not technically suitable, but rather due to a wider project being cancelled, so we shouldn't draw negative conclusions from this.

Trial costs

49. Approx USD $190,000

50. 12 cpu core licences

51. 12 months support and maintenance

52. 90 hours integration effort with datasources etc

53. 15 users basic training

54. They demonstrated the system on November 4th to our partners in London. We were invited, but felt that it would be better to discuss the system here first and request an onsite demo where we could invite more people.

Conclusions

55. Clearly adopting this commercial system for visual analysis and reporting would conflict with the current plan in Better Analysis Agility to develop our own desktop integration framework based on Eclipse, and it would have a massive effect on our ability to collaborate with second parties. However, the product provides a lot of capability in a very well integrated product, and it is interesting

to see how such a product works - especially given the similarities to inhouse development.

56. We should perhaps consider an on-site demo, but such a demo should be carefully contained so as not to undermine the programmes of work, as these will likely provide more long term benefit. We have also been invited any time to their offices in Palo Alto for more demos and discussions, but they also have a fairly large presence in Maryland and Washington D.C.

Pros

- It looks good and appears easy to use - but this could be partly marketing spin from very good sales people.
- It seems very powerful - for example, there are important capabilities (such as collaborative workflow, analysis history, and use of extracted entities from text documents) that would require effort to integrate quite so smoothly into a GCHQ in-house tool - but nevertheless could be done.
- It seems to scale well with federated access to very large databases (but note, this is purely based on questions we asked rather than solid evidence)

Cons

- This is not in Eclipse RCP. If we produced our own tool in RCP then we would be throwing away their GUI and any functionality from other agencies using the tool. I.e. we would have to adopt this instead of Eclipse rather than merge the two.
- Adoption would have huge monetary and IPR cost (ie its no longer our IPR, other agencies would need to buy in too in order to share "plugins")
- We would be buying a "complete architecture" and therefore would become utterly dependent on a commercial product.
- There are no British staff which has caused issues over integration during talks with (unspecified) sister agencies in London.
- This is a small company who initially would be very keen to please new customers by adding new features. However, as their customer base grows we might get frozen out.
- It is possible there may be concerns over security - the company have published a lot of information on their website about how their product is used in intelligence analysis, some of which we feel very uncomfortable about.