# In the land of the blind, the squinter rules

(Wim Remes)

## Acknowledgements

TJ, Nick, Jayson & Robin for bearing with me and loving me unconditionally. I wouldn't be here without you. I try to be a better husband and father every day.

Raffael Marty (@zrlram) because of his dedication to bring information security visualization tools and techniques to the attention of the security community, through his work on secviz.org, by releasing the DAVIX live CD and most recently by starting loggly.com. If it weren't for Raffy, we'd still be roaming the data desert.

Eternal gratitude to Jayson Street (@jaysonstreet). For accepting me as a speaker at Excaliburcon 2009. If that hadn't happened, I would never have gotten to public speaking and I wouldn't be writing this paper. Jayson, what we achieved in 2009 will last beyond our lifetime.

For various reasons : Craig Balding, Chris John Riley, Dale Pearson, Chris Nickerson, Alex Hutton, Ian Iftach Amit, Stacy Thayer and Christien Rioux. You all rock hard.

Kevin Riggins, Brian Honan, Mark Hillick, Jimmy and Fortuné for spending some of their precious time to review and offering constructive feedback to improve this paper.

The Brucon crew and volunteers.

My employer.

## Disclaimer

All ideas and opinions voiced in this paper are solely those of the author and do not represent those of his past, current or future employer, clients or associates.

## Introduction

There isn't a lot to bring up against the year-old adage that says that the attacker has to be only right once, while the defender has to be right all the time. Moreover, the toolset available to the defender to help him achieve that goal is still very limited. The firehose effect is well-known to intrusion analysts and forensics investigators : Every product, tool or technique adds to the avalanche of data that these professionals face, every single day. Additionally, there's numerous irrelevant data to sift through to find the tracks left behind by the attacker or unveil the few breadcrumbs that indicate an imminent attack. The data is there, but it's often clouded by the sheer amount of other data that the analyst or investigator has to sift through.

On the other side, information security departments are strapped for money, trying to justify every cent they spend or want to invest in yet another effort to counter threats. Is your C-level executive interested in another report where you tell him how many viruses were blocked by your corporate proxy or who tried to download executable files to his laptop? In order to engage management with the information security challenge, we need to step up our game. We need to communicate clearly about the issues at hand and how we are contributing to the protection of the company's most valuable assets.

Both of these important challenges for information security professionals seem unrelated, but they are not. Both can profit from advanced visualization techniques:

- Techniques that move beyond the obligatory pie and bar charts that have, in most cases, failed to help us transmitting the message that's required.
- Techniques that allow us to move away from spreadsheets and other desktop applications that tend to limit the user's creativity, into a realm where we can properly analyze, understand and visualize data.

This paper will introduce you to the basics of information security visualization and how to apply it in your daily tasks as an information security professional.

It needs to be noted that I could have plastered this paper with numerous examples of visualization, however, I don't believe those would have added value to the information presented here. You will get more value out of the exhibits created from your own datasets by applying the techniques described here.

## The art of security visualization

Operating systems, applications, database systems, firewalls, intrusion detection/prevention systems and vulnerability management products all have one thing in common : They relentlessly bombard the security analyst with data. The person tasked with making sense of this data only has a limited set of tools available to dig through the data with time as his main adversary.

Just as painters and sculptors are able to bring out an emotion or a specific characteristic of a subject (be it a pear, a building or a person), data analysts are able to bring out specific information from a bunch of otherwise meaningless data.

The basic process employed in security visualization is composed of 3 tasks :

- Collecting (the right) data
- Processing
- Visualization

All three tasks are of equal importance and the data analyst should pay attention on all levels to achieve the desired result. A small mistake in collecting, processing or visualizing the data can have a very high impact on the end result.

## Collecting data

There's a variety of sources that can provide data to the information security analyst. We identify four important types of security-relevant data. Firstly, and quite well-known to most security professionals there is data gathered from the internal security infrastructure : Firewalls, Intrusion Detection/Prevention Systems, proxies, (network) access control systems, host logs, etc. all provide a plethora of information (either in real-time or historical) in various formats. However, there is a fundamental problem with this data. Influenced by regulatory requirements, most of the infrastructure solutions provide what Andrew Jacquith calls 'happy metrics' in his book *Security Metrics: replacing fear, uncertainty and doubt*. Anti-virus solutions tell you how many viruses were stopped, not how many were not stopped. IDS's let you know how many attacks were detected, with IPS functionality enabled they might even tell you how many were prevented but rarely how many remained unseen. Even though data gathered from each data source on its own doesn't provide a lot of added value in the form of usable metrics, they are not to be disregarded. The biggest challenge in creating value from these data sources is in correlating them. This subject will be covered at length later in this paper.

Secondly there is data provided through the instrumentation of business processes. Data is gathered throughout these processes with the goal to identify critical characteristics of the process. This allows the analyst to judge whether a process is running optimally and where adjustments are required. Gathering data from processes requires a higher level of maturity from the organization: Processes need to be formally implemented and one should refrain from encouraging a "pass-the-blame" culture based on process measurements. Gathering useful metrics from processes will allow the organisation to transform the Check phase from the well-known Deming cycle into a stepping stone to improvement. Where the creator of this concept clearly envisioned that principle, the implementation lacks in this regard more often than not unfortunately.

A third type of data is gathered through testing the infrastructure, whether this is through automated vulnerability scanning or targeted penetration testing, the data will allow the analyst to identify flaws within the infrastructure that are not easily identified through either of the afore-mentioned types of data.

Last but not least we point out an often overlooked type of

data. It is however, in my humble opinion, crucial data for security dashboards, especially when used to communicate security information to executives. This data is gathered from external sources and allows the comparison of the current internal situation with that of a peer group. The power that lies within this type of data should not be underestimated. As executives are, by nature, competitive they will be excited when they beat the competition by a noselength and motivated when the data points out that they're trailing behind. Again, there's more than enough data out there. Obvious sources are sites like osvdb.org, cvedetails.org and datalossdb.org. Other valuable data sources are industry reports like the Verizon DBIR and reports by McAfee, TrustWave, Symantec, Cisco, Ernst and Young and Deloitte to name just a few. There are a few issues here though. Most of the reports hold valuable data in variable amounts, but they also tend to be more or less tendentious. Where statistics can be used to prove any point, this is especially true for information security industry reports. I don't want to pass judgement on any of the mentioned reports or on any of those not mentioned in this paper. It is up to the person using the data to select those data points that are relevant for him. To separate the wheat from the chaff, one can assume that those reports using a published and verifiable methodology will provide more trustable data than those who don't.

This is something to remember when creating your own reporting. If the methodology used is not set in stone, your results can mean anything and you can assume that your consumers will come back with questions that you might not be able to find an answer to. Ideally you should be able to provide access to the data and your methodology used to create a report or a dashboard. If you are confident in doing so, you can be fairly certain that the information derived from it is trustworthy and your judgement is truthful.

## Processing data

When processing data, one needs to be careful : by applying mathematical functions, algorithms or even simply collapsing data, context can be lost easily. Without context, data loses all of it's value immediately and whatever you visualize afterwards will cloud your vision if not blurring it permanently.

It is fairly easy to get carried away when working with a varied dataset, therefore, it is important to realize that the outcome of any calculation that will be used to present data to an audience should be either a cardinal number or a percentage.

A cardinal number represents how many of the measured units there are. A percentage represents how many of said unit there are in a given dataset.

Example

Given a packet capture containing traffic in and out of a network, one could gather the following metrics :

- # of packets ingress
- # of packets egress
- total # of packets
- # of packets with dstport 80
- # of packets with dstport 22
- % of egress packets with dstport 80
- % of ingress packets with dstport 22

Each metric is represented either by a cardinal number (#) or a percentage (%) and none of them allow for any ambiguity.

A favorite quote of mine was uttered by Alex Hutton (in a talk at a BSides conference if I remember correctly) not so long ago :

*"peanutbutter times fast equals awesome"*

While Alex was talking about the quantification of risk in particular, I believe it holds true for the vast amount of metrics you can gather. Even those that don't communicate risk in the first place. When you apply complex mathematical functions to your data, or worse let the tools apply them for you, your results might be severely impacted.

## The need for a common language

As pointed out in the chapter 'Collecting data', most of the available infrastructure data sources like firewalls, IDS's, host, database and application log products, etc. provide security-relevant data in one format or another. Vulnerability scanning products (genre Qualys, Nessus or Outpost24) are another breed in this category. There is a reason why I call them products instead of solutions, a designation that would do them too much justice when it pertains to this subject.

Most of the network security products provide a more or less basic interface to access and process the data they collect. This usually boils down to the liberal implementation of the Syslog RFC 3164. Somehow, most vendors have succeeded in not keeping with the standard and they have freely added fields, modified the data model or liberally interpreted formats. The result is that, in order to process the data gathered from different sources, one has to put in the extra hours (and budget) to normalize the data in such a way that it is ready for processing. In this regard, there have been numerous efforts in the past decades to create a reporting standard. All were either too academic (detached from how it works in the real world) or vendor-centric (those that believe they if they are - or are going to be - the biggest means their 'standard' will automatically win) to succeed. Most recently MITRE backed the CEE (Common Event Expression) initiative and started publishing documents to support it. I am convinced that CEE holds a lot of promise and should be regarded as the future in logging. If we succeed in standardizing logging, we will greatly influence the relevance of all efforts to make sense of the data we gather. It is with dread that I recently noted the initiative to build a 'cloud' logging standard. I fully understand that the clouderati believe they have something special and unique at hand but I beg to differ. Cloud is just a dynamic mix of bare iron, platforms and applications that have the same basic characteristics as those you are used to. I strongly suggest this initiative to join forces with CEE and create one single standard.

Vulnerability data, as delivered by the vulnerability scanning products, suffers from the same illness. Every vendor in that space currently tries to be the biggest and assumes their data model will be the standard once they gather a large enough percentage of the vulnerability scanning market. Humbleness is advised to those vendors. It needs to be understood that your clients will most likely be using multiple products or at least need to collect data from various sources to allow for correlation and exhaustive reporting. Neither of which the abovementioned products provide the necessary functionality for. It is, without doubt, time that the vendors of these products agree on a standardized data model that allows for this to happen. Whether this is build on the existing model of one of the leading vendors or build from scratch in cooperation with the community remains to be seen. CVE, another noteworthy standard driven through MITRE, has been pushing the enveloppe for a while now. A quick glance on their website mentioning the degree of adoption shows that there's still a big gap to be crossed. It is my personal belief that the key lies in the hands of those willing to achieve an open and vendor-independent standard.

## Visualizing data

When first approaching the art of (security) visualization, it is easiest to work with your stock spreadsheet program. Whether this is Microsoft Excel, Apple Numbers or OpenOffice.org depends on your personal taste. All provide for a limited dataset to be analyzed in a fairly structured manner using build-in functions, pivotables and a limited set of visualization techniques.

I see three main issues with using spreadsheets :

1. They severely limit the user's creativity
2. They provide limited support for flexible correlation of variable datasets.
3. The limit almost always lies at 65536 rows.

This is to say you will swiftly outgrow the functionality offered by spreadsheets. While useful at first, their primary goal is not to assist you in data analysis and visualization.

The leap from spreadsheets to professional data visualization applications like Tableau (to give one example), or even specialized business intelligence (BI) tools, is quite big and there is little to be found in the space between both. Luckily the fact of the matter is that most of the audience I meet that is giving security visualization a go has a fairly technical background. While these skills are not a prerequisite to data visualization (money buys you all the tools and knowledge you need), I believe resources that possess them are a must-have on teams facing the challenge of data acquisition and processing. Python, Perl, regular expressions and bash scripting belong to their basic skillset.

While one could wonder why we need to resort to techniques that might be considered antiquated in this regard, I'm a strong proponent of this approach as it allows for great flexibility and an extensive level of automation in the long run. With the lack of a logging standard and the absence of a standardized data model for vulnerability information, regular expressions provide the much-needed mold to filter and normalize the data before consumption. By using tools like grep, sed, awk, etc. or equivalent functions in Python or Perl, you will rarely run into a limit on the amount of data you can crunch. So, while it could be perceived as rudimentary, I believe it still is the best way to learn to understand your data while not limiting the creativity of the user.

Once the data is properly structured, which should be regarded as processing rather than visualization, it is time to feed it into products that can produce the required exhibit. Unknown to many, there are ample visualization tools released under open source licenses, most of which were gathered in the excellent live CD Davix by Raffy Marty. The Davix ISO image can be found on http://www.secviz.org, a site that also provides a lot of examples on how to use the toolset. A few examples will be given in this paper, but it is by no means the goal to rehash Raffy's work here. All tools are clearly described in the Davix Manual and you are encouraged to use Raffy's book (see Sources) as a handbook to kick off your endeavours in security visualization.

Afterglow

Afterglow is a perl script to translate CSV formatted data to a DOT graph description file, using a properties files used to describe how the nodes and edges will be visualized.

```
cat source.csv | afterglow.pl > result.dot
```

the DOT file can then be used to visualize the result :

```
neato -Tpng -o result.png result.dot
```

I'm pretty sure you can use this tool in various scenario's. I've found this particularly helpful in visualizing log events, network captures and vulnerability data.

GlTail

(http://www.fudgie.org/)

glTail allows for real-time visualization of logfiles from various sources (as long as you can reach them through ssh). Create a config file and run following command :

```
gl_tail configfile config.yaml
```

This is enough to visualize real-time data from Apache, IIS, mysql, postgresql and numerous other data sources.

The video's on the website speak for themselves.

InetVis

This tool supports 3D Visualization of network traffic in a 3D scatter plot. As it is a GUI tool, its use is very straightforward.

Mondrian

Again a GUI tool, enabling the user to generate various linked graphs, including histograms and

scatterplots.

RT Graph 3D

Allows you to visualize 3D link graphs in real time.

TimeSearcher 1

This tool supports the analysis and visualization of time series data.

Again, I encourage the reader to burn the DAVIX ISO image and take it for a run. You will be amazed by how quickly you can get good visualization results.

While local tools can provide for many of the scenario's one can imagine, the power of web-based visualization tooling can not be disregarded. Users have, since long, grown used to consume information through a web browser and the use of dynamic web pages allow for exhaustive customization of the data presentation to the requirements and preferences of the particular user.

One example of a useable framework is jquery, a cross-brower JavaScript Library that helps to facilitate the client-side scripting of HTML. There are several plugins that can be used when generating custom dashboards :

- http://omnipotent.net/jquery.sparkline/
- http://www.jqplot.com/
- http://www.maxb.net/scripts/jgcharts/include/demo/#1

The last one provides a perfect segue into the last option that you might want to explore when it comes to data visualization as it is a jquery implementation of the Google Charts API. Google Charts (which can be found at http://code.google.com/apis/chart/) could be described as a powerhouse for data visualization. There is almost no limit as to what you can visualize and the depth of insight that can be gained by selecting that one striking representation of your data. Using the Google API, it is quite easy to get distracted by 3D effects and the possibility to apply an almost unlimited colour palette to your data. A good rule of thumb is to only apply 3D if the extra dimension provides additional information to the consumer of the graph. Additionally, it is a matter of courtesy to choose one palette and stick with it. Colour, while useful, is your biggest enemy in data visualization. I would recommend not to use it unless it encodes (part of) the data. You will notice that data usually pops out more in a greyscale palette then in a palette using random flashy colours with no meaning at all. The key is to draw attention to the data, a lot of visual features (like grids, labels, colour, etc.) actually achieve the opposite.

## Dashboards

As a security professional you will experience an increasing demand on the state of security in your organisation. Executives require more information on the security posture, either for regulatory requirements, to assure that security initiatives are in line with business goals and to ensure that the security processes are running as expected.

Visualizing security data is one thing, presenting it in one clear, concise and understandable dashboard for an executive, who doesn't necessarily have a security background is a whole different ballgame.

There is no "how to build a perfect dashboard" checklist to help you out here. Your success will be greatly influenced by your ability to communicate. The best bet is to sit together with the prospective audience of the dashboard and provide them with samples of data you can work with. They will be able to clearly articulate what they expect from you and that should be what you want to deliver. Also, you will have to investigate which dashboards your audience is used to. Applying drastically different techniques is probably the worst you can do. Whether you use bar graphs, pie graphs, line graphs, sparklines, bullet graphs, 3D scatter graphs, interactive world maps or the dataviz equivalent of a Rubik's cube is entirely up to you. As long as you make sure that the chosen visualization technique is relevant to the data you want to present, the odds are looking up.

## Conclusion

Security visualization is not extremely different from data visualization in general. You will be using techniques that have been used in other professions for decades. Truthfully, most of the techniques have not been perfected. How the art of data visualization has received such a limited amount of love still is a mystery to me. I personally love shaping data until the information pops out. While starting out may seem rough and your results might not encourage to delve further into the project, I'm convinced that my working on those skills that are not directly related to either the offensive or defensive security practices will help you to become the well-rounded security professional that organisations in this day and age are seeking. If 1% of the people who read this paper or attend my presentation just barely scratch the surface of this subject, my task is accomplished. I sincerely hope to see much more relevant exhibits in reports, effective dashboards and people using visualization techniques to do their job.

## Sources

This paper should not be regarded as the end all and be all for security visualization. There is a lot of information out there, either in the public domain or copyright protected content. The list provided here is not exhaustive but presents a good overview of the information sources I used to get started in security visualization. Note that not all sources are security specific. I believe it is extremely important to step outside your own field of expertise to broaden your skillset. I've found a lot of value in peeking over the wall to see how other professions do their stuff. Translating that to my own trade has, to say the least, taught me some very interesting lessons.

*Security Metrics: replacing fear, uncertainty and doubt, Andrew Jacquith. Addison Wesley, 2007.*

*Applied Security Visualization, Raffy Marty. Addison Wesley, 2008.*

*The visual display of quantitative information, Edward Tufte. Graphics Press USA, 2001.*

*Information Dashboard Design: The effective visual communication of data, Stephen Few. O'Reilly Media, 2006.*

*http://www.secviz.org*
*http://dataviz.com.au*
*http://informationisbeautiful.net*
*http://visualization.geblogs.com*
*http://flowingdata.com*