Big Data Analysis

*Bill Binney**

My experience in working with Big Data over the last 40 plus years involves data produced by communications or transactions between people, not machines talking to other machines or themselves, or machines taking measurements. In the realm of transactions between people the problem is content volume. If you attack Big Data by looking at content first, then you quickly become overwhelmed by the content. Instead, by using the metadata associated with content to help define how the content is related and clustered, this gives a view into the content without having to look at it. Also, the metadata associations, often referred to as 'social networks or graphs', organise content into related clusters of communities. This makes selecting content for further study a much more focused effort. These social networks or graphs are created by attributes that are associated with individuals or organisations like: phone numbers, IPV4/IPV6 numbers, MAC numbers, account numbers, passport numbers etc. These relationships in the graphs/social networks show the scope of interactions that people have in the world and help to define groups of people participating in an activity either business, social, family or other.

To get a reliable social network/graphs constructed, it's important to make sure the data used is corrected and validated. This can be achieved by first selecting attributes that can be relied on to make decisions automatically about the correctness of the data in storage. Of all the steps in analysis of Big Data this is probably the most important. There are many forms of metadata and selecting the right one to use to make decisions is critical. For example, in one Big Data case I worked, the engineers and computer programmers gave me over 200 different types of metadata. But, only 13 of these metadata elements were useful to make analytic decisions. Once you have the right metadata elements to construct the social networks/graphs, then these elements can also be used to associate corrupted equivalent elements of information from which aliases can be defined. Plus, by keeping a frequency count of these elements, this data can be validated and corrected. When the social networks/graphs have been validated and corrected, then you can address the questions you want to get out of Big Data. It may be that only a small subset of the metadata would be useful in making decisions about the data and its content for discovery of either new individuals in the communities or even perhaps new activities of these communities of individuals.

So, the steps to successfully analyse Big Data sets are:

1. Make sure to select metadata that is reliable to make decisions about the social networks/graphs and data content associated.

* Bill Binney, SHORT BIO. For correspondence: EMAIL.

2. Use frequency counts to validate relationships within the social networks/graphs.

3. Use validated metadata relationships to correlate unresolved information.

4. Correct corrupted information based on multiple matches of reliable metadata with previously validated relationships.

5. If a known target of interest (seed) is available, use its metadata to develop a community of interest, ie the social network of that specific target.

6. If no seed available, use accepted rules to identify suspicious activity. For example, if an entity frequentlyvisits web sites that advocate things such as: paedophilia, violence, or how to make a bomb, etc. Thenfocus on the community that that entity has in the social network/graph.

7. Otherwise, use things like geographic distribution of the social network/graph members to isolate communities that are suspicious because they are distributed over countries that are involved, for example, in drug smuggling.

8. Once you have developed a target/community of interest, then pull the content associated with that target/community to try and discover what activity they are about.

9. At this point, latent semantic indexing can be used against the content of the target community. With a smaller community generating content, this approach would be more effective as it would not be going against the vast amount of data available which would produce confused output as a consequence of random probability matches.

10. When discoveries are made, make sure the data is feed back into the process to keep the knowledge of the process up to date.

11. Always regularly review the processes to make sure they are still functioning properly.

12. Maintain an audit trail on all processes to show that the system is properly functioning.

13. Make sure any new discoveries that have not been previously included in data processing are added to the system. Always check new discoveries to insure they are functioning properly.

While massive data can be intimidating, the reality is that all possible combinations are not the case. For example, if one considers the total number of phones in the world to be say 3.5 billion, then every possible connection taken in order between them would produce 3.5 billion factorial variations. But, the reality is an infinitesimal of that

number – something on the order of a few hundred billion relationship combinations. This makes analysis of this scale of information a manageable problem. Similarly, other massive data sets can be analysed using these same steps.