University of Waterloo

The case for the dynamic analysis of social networks

Robert Warren
University of Waterloo
Waterloo, Ontario, Canada
rhwarren@uwaterloo.ca

* The author would like to thank Mike Warnecke, Dana Wilkinson, Jonathan Fishbein and Ian Mackinnon for their help and comments with this work.

Abstract

A interesting problem in Social Network Analysis (SNA) is the resilience to interference and how information flows from one person to another. In the past, we have always approached these problems from a static or 'snapshot' perspective: all available data was lumped in the same analysis and a conclusion derived.

Our hypothesis is that since the world is a dynamic system, the analysis should either be dynamic itself or at a minimum, conclusions based on static SNA metrics should be revisited. We test our assumptions on Gnu Privacy Guard key trust databases, discuss examples where the static assumption is counter-productive and suggest possible alternatives.

Problem Statement: Network destabilization

Social Network Analysis (SNA) is used as a means of analyzing people and the relationships that bind them [1]. In several situations, researchers are interested in the flow of information within the network and in the absolute influence yielded by each node.

Another application of SNA is the analysis of the network in order to achieve its destabilization or identify weaknesses within its structure. This has direct applications in communications networks and more recently in international security [2].

The two approaches to this type of problem are the removal of either vertices (Figure 2) or edges (Figure 3) to try and isolate the elements within the graph and increase the distance between the nodes still connected to the graph.

Because locating highly connected vertices is computationally faster than selecting critical edges, many researchers prefer removing or isolating nodes rather than trying to section the graph.

In our research, we evaluate the effectiveness of node removal in destabilizing very large graphs of real-world data and determine if the assumption of a static graph is reasonable in such situations.

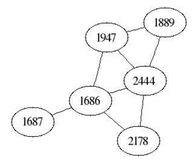

Figure 1 – A sample graph from the Debian dataset.

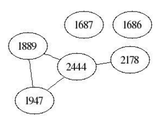

Figure 2 – The sample graph with a major nodes isolated.

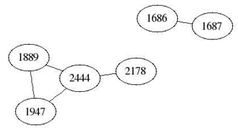

Figure 3 – The sample graph partitioned by cutting several edges.

Approach assumptions and limitations

Traditional, 'classical' Social Network Analysis is based on a number of assumptions:

- That the nodes within the system and the edges that bind them are static and unchanging.
- Because the system is static, the nodes are unaware of modifications made to their environment and do not react to them.
- There is no growth, optimization or expansion of the edges within the graph. Whatever information may flow from one node to another through the graph does not trigger the creation of new edges between the nodes.
- All nodes are present at all times within the network and are always available.
- All nodes and edges have the same cost of removal or isolation.

Research problem and experimental design

Our concern is that the assumptions are skewing research results in that the current methods do not take into account the future changes in the system. Hence, it is very possible that by the time that action is taken, the situation will have changed and the desired results will not have been achieved.

Hence, we look at two social networks extracted from public key cryptography databases and test some of the assumptions.

Specifically:

1. The graphs are analyzed from a temporal perspective.
2. The link prediction problem is revisited within the datasets.
3. The rank stability of the nodes deemed as most important within the static data is reviewed within the dynamic data set.
4. The effects on the graph of removing the most connected nodes are revisited in a large dataset setting.

Real-World datasets

As a means of testing our dataset we obtained two databases of public key cryptography keys from the Debian software project and the University of Alberta public key server. Both these databases contain detailed timestamps information and we make an explicit assumption that any trust relationship between cryptography keys implies a relationship, or an edge.

Since the datasets deal with keys and signatures, we preprocessed the datasets to record link multiple keys into single individuals and sets of signatures into relationships. Because expiry information on many keys is missing, we make another explicit assumption that a key is only valid for a year after the date of last activity.

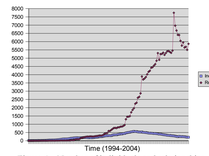

Figure 4 – Number of individuals and relationships active within the network in 1 month intervals for the Debian dataset.

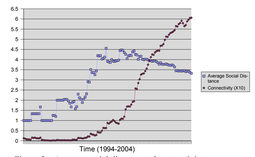

Figure 5 – Average social distance and connectivity in 1 month intervals for the Debian dataset..

The first dataset is the keyring of the Debian software project, which is manually curated. The keyring has about 1,500 individuals and 15,900 relationships within it. Figures 4 and 5 represents the evolution of the active individuals within the network. Note the sudden surge in relationships post-1999, that coincides with the first large Debian conferences.

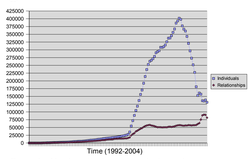

Figure 6 – Number of individuals and relationships active within the network in 1 month intervals for the Alberta dataset.

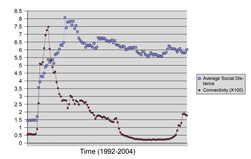

Figure 7 – Average social distance and connectivity in 1 month intervals for the Alberta dataset..

The second dataset comes from the University of Alberta public key server which is collated by 'gossip' engines. The dataset size is quite large, with over 1,450,000 individuals and a little less than 500,000 relationships. The dataset is interesting in that it gives us a representation of how a very large system behaves, since most SNA datasets tend to be very small and curated.

Figures 6 and 7 represent the population and relationship plots for the dataset over time, as well as the average distance and connectivity between the individuals in the database. Note how the social distance seems to stabilize over time.

Link Prediction

Intuitively, we accept that at some point we will come to know the friends of our friends. As such, another aspect of the problem is how to take into account the growth that is normal to any social network. Liben-Nowell et al. [3] previously looked at the problem and determined that there existed a relationship between the lengths of the paths linking two nodes and their likelihood of generating an edge.

We attempted a similar experiment with our dataset, by finding the shortest path that linked two nodes before an edge linking the two appeared. The results of Figures 8 and 9 are consistent with the results of Liben-Nowell in that nodes with a shorter indirect path tend to create direct edges. We also show the number of nodes with known indirect paths that did not create edges. The normalization of both graphs is not sufficient to build an accurate prediction model. However, we do conclude that this implies that the neighborhood around highly connected nodes has a tendency to naturally grow edges bypassing the highly connected node.

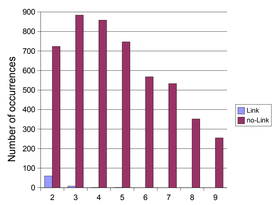

Figure 8 – Number of node that created or did not create a direct edge versus the length of the indirect path for Debian dataset.

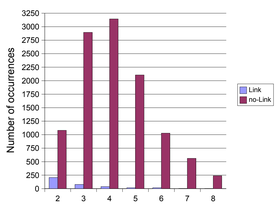

Figure 9 – Number of node that created or did not create a direct edge versus the length of the indirect path for the Alberta dataset.

Localized small worlds

An experimental observation is the existence of very interconnected small worlds that revolve around nodes with large number of edges. Figure 10 is a partial representation of the neighbors of node 893555 that is typical of highly interconnected nodes.

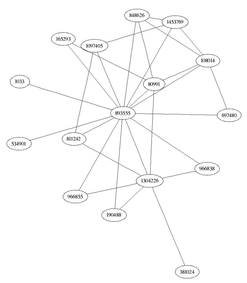

Figure 10 – Partial representation of the neighbor node 893555.

What makes this interesting is that the neighbors of node 893555 have a high number of edges within and outside the neighborhood. In effect, the 'super node' is surrounded by a ring of very connected nodes: the average number of edges for a direct neighbor of node 893555 in July of 2001 is 13.1 edges. This is a sharp contrast to the overall average of 4.3 edges per node for the graph.

Hence, any attempt to remove the 'super node' (Figure 11) does not necessarily destabilize the graph: (1) the neighborhood is acting as a highly integrated small world and (2) the link prediction problem suggests that any break has a higher probability of being overcome.

A possible solution we have been experimenting with, has been the random removal of nodes with only two edges (Figure 12). These gateway nodes are usually part of a path longer that 2 edges that is less likely to be rebuilt and which effectively reduces the connectivity of the graph.

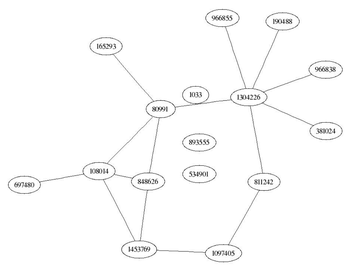

Figure 11 – Node 893555 is isolated. Connectivity is 0.86.

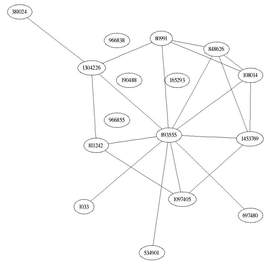

Figure 12 – Nodes 165293, 966838, 966855 and 190488 are isolated. Connectivity is 0.55.

Rank Stability

A final concern was the rank stability of highly connected nodes within dynamic graph. Many SNA researchers will assume that the dominant nodes will remain so for the whole time period.

We set out to verify the stability of this assumption by calculating the list of the top 10 most connected node within each data set and then tracking their instantaneous rank over time.

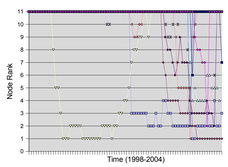

Figure 13 – Changes in overall rank for each one month period in the Debian dataset.

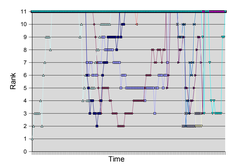

Figure 14 – Changes in overall rank for each one month period in the Alberta dataset.

The data series in Figures 13 and 14 represent the changes in rank of each of the top 10 connected nodes in the datasets. Rank 11 is used to indicate any other rank than 1 to 10. Looking at both graphs, we conclude that the actual impact and participation of the nodes varies greatly with the time period under study.

Hence any attempt at destabilizing the network should take into account the transient nature of 'super-nodes' within social networks. Too much stale data may skew our conclusion towards a solution that attempts to solve an obsolete problem.

Conclusions and future work

In this poster we reviewed some of the common underlying assumptions about Social Network Analysis, with a special emphasis on the dynamic nature of the system.

We know that people and their relationships are in constant flux, with new relationships being formed and others expiring. As such, which nodes and edges should be influenced must depend on timely information instead of an overall gross average.

The process of growth is one that we have identified as important, as it provides insight into how the graph might appear. How the nodes will react to changes forced upon them is a difficult problem that we have not addressed, but one which we feel needs further study.

Finally, the assumption that the removal of a few key nodes will result in the destabilization of a network is one that needs to be revisited. Because of the localized density where highly connected nodes are located, new strategies must be devised that do not rely on attacking strongly integrated nodes. Possibly, the selection of a number of a weaker nodes may be an effective solution.

References

[1] J. Travers and S. Milgram, An experimental study of the small world problem, Sociometry, (4)32, 1969

[2] K. M. Carley et al., Destabilizing Dynamic Covert Networks, Proceedings of the 8th intl. C&C research and tech. conference, 2003

[3] D. Liben-Nowell, The link prediction problem for social networks, CIKM 2003