Dataset analysis

CCN: ENG-258GI

February 21, 2016

1 Introduction

In the last few years, advancements in computing and availability of large datasets have created a new field of research around large networks. The emergence of social services such as Facebook or Waze has provided large-scale user-generated data of not only abstract entities such as friendships but also of spatial information. Nevertheless, there are other networks equally important to understand, especially the ones based on real-world infrastructure, as their properties may provide additional insights into their design process. This paper analyses graph modelling Pennsylvania's road network.

2 Dataset description

RoadNet-PA dataset is available online and represents a road network as an undirected graph stored as an edge list. Each edge corresponds to a stretch of road whereas each node indicates a junction. In fact, authors of this dataset treated both dead-ends and turns as nodes as well. The network itself is quite huge, comprising $N = 1088092$ nodes, and sparse having only $K = 3083796$ edges which are negligible compared to a fully connected graph of the same size. Figure 1 shows the visualization[1] of the data which was generated using extra geographical data.

It remains unclear how highways and overpasses have been treated as they could include a node at an intersection of two roads despite being physically disconnected. Such distinction is crucial as it determines whether a resulting graph is planar or not since any underpass edge would violate planarity. Nevertheless, given the size of a network and much higher density of regular roads compared to highways in Pennsylvania it is reasonable to assume an analysed graph to be planar.

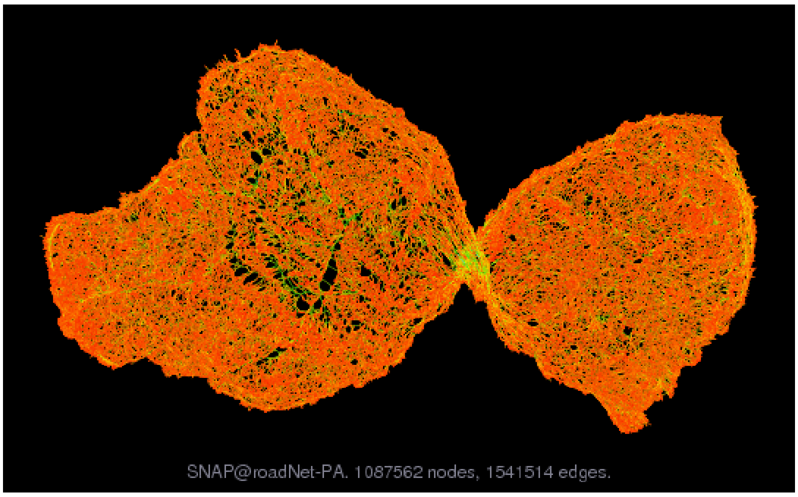

Figure 1: Visualisation of roadNet-PA network with geo data[1]

3 Dataset preprocessing

One of the reasons underlying a transportation network is the need to connect distant geographical regions together enabling the movement of resources. As such, the first step of the analysis is to check whether an entire graph is connected or not. Surprisingly, the network turns out to have 419 disjoint subgraphs although the vast majority of nodes - 1087562 - are part of a giant component. The remaining 530 nodes form mostly subgraphs of two nodes connected with one edge as well as a few larger graphs of up to 19 nodes. These disconnected components suggest that the data is either slightly corrupted or contains areas unreachable directly by road, for example, a small island without a bridge connection. Smaller components may also suggest pedestrian areas secluded from vehicular traffic. Unfortunately it is hard to give definite explanation without cross-checking nodes' geographical data with external maps. For the purpose of this report, only its giant component is used in further analysis to reduce clutter.

4 Degree analysis

4.1 Degree distribution

Degree distribution analysis usually gives some insightful information about the network as, for example, Twitter's follower graph most likely has a different distribution than a subway network due to physical constraints. Figure 2 shows the degree distribution of the road network which has a few interesting features. Unsurprisingly, it drops off sharply on a log-log scale for degrees greater than 4 with the maximal one being just 10 and represented by 3 nodes only. Such situation is reasonable for a model of a transportation network as T-junctions and intersections with 4 exits are most common whereas larger ones usually involve a roundabout and are less frequent in the US. Although there is no formal limit on their size, complexity of junctions grows along with their degree without offering any substantial gains in connectivity.

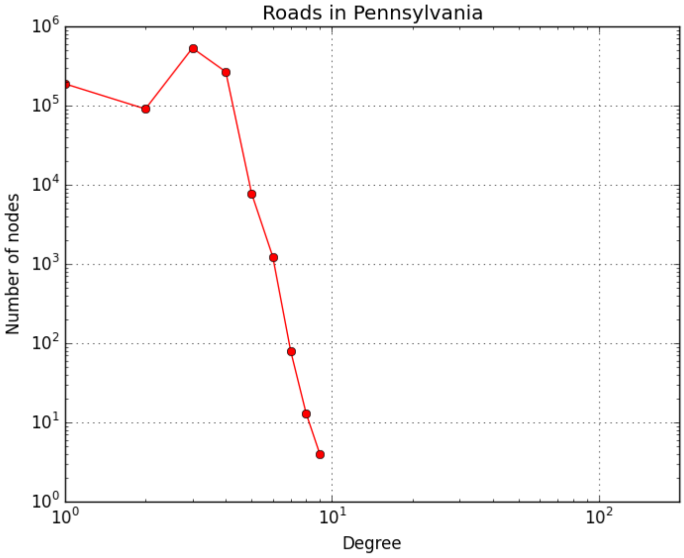

Figure 2: Degree distribution

4.2 Spatial context and planarity

It is important to look at the road network taking into account the assumption of planarity made in Section 2 as well as its spatial context. Such networks tend to have different characteristics than graphs modelling social interactions from, for example, Facebook. One of the main differences is the lack of high-degree *hubs* present in other networks. It is not the case for transportation networks as physical distance between nodes prevents edges from being too long and junctions from being too complex. Consequently its diameter is very high - 786 - compared to scale-free networks where, according to Miligram's study based on Facebook's data, most people are connected by paths of 5 steps[3].

On top of that, planar graphs have a few properties one of which is a result of Euler's formula stating that $v - e + f = 2$, where v, e, f are the numbers of vertices, edges and faces respectively. It follows that an average degree is strictly smaller than 6 which is true for the analysed data - 2.83. Also, planar graphs are *sparse* in a way that the number of edges grows linearly with the number of nodes.

4.3 Dead-ends

There are plenty of nodes with just one edge corresponding to dead-ends which is somewhat surprising for an extensive road system especially in downtown areas. However, the dataset also includes American suburbian regions where many residential units are single-family households sprawling over large areas. These are usually connected to main roads through non-transit local streets with dead-ends as shown on Figure 3.

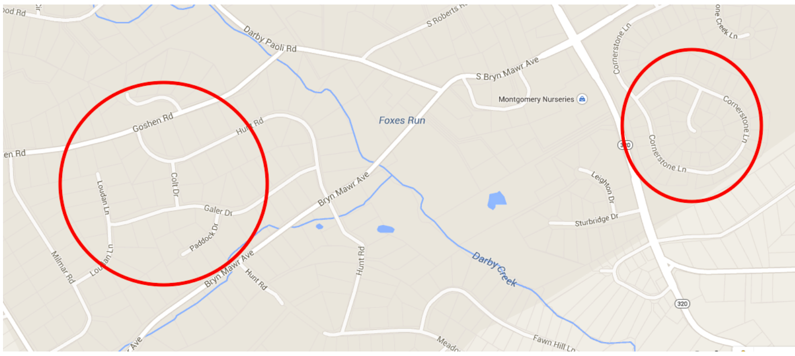

Figure 3: Pennsylvania's suburbian road network example

4.4 2-degree nodes

Section 2 mentioned that the dataset treats each turn as a separate node with two undirected edges. However, for the purpose of road network analysis, it does not matter much whether these road sections between nodes of higher degrees are treated as a single edge or a set of 2-degree node combinations. Hence, these nodes could be removed leaving only those of higher orders to improve computational efficiency. However, one needs to be very careful while doing so as the removal of some 2-degree nodes changes the structure. In fact, deletion of these *unnecessary* junctions led to a slightly smaller graph however computational savings did not justify the effort.

5 Community analysis

Community within a network is defined as a group of nodes which are more densely connected with each other than with all other nodes of the graph. Consequently, they are characteristic of a network structure and in various networks correspond to different concepts. For a road network, they may represent villages, residential areas, towns or even city boroughs.

5.1 Louvain method

The heuristic Louvain algorithm detects communities by iteratively maximising modularity of graph's partitioning. It starts from single nodes working its way up in $O(n \log n)$ time by joining separate communities. It terminates at a level with highest modularity defined as a difference in internal edges density compared to a random graph. Figure 4 shows how modularity changes as the algorithm progresses from level to level. Resulting graph is monotonic and achieves its maximum at the highest level, that is for largest communities.

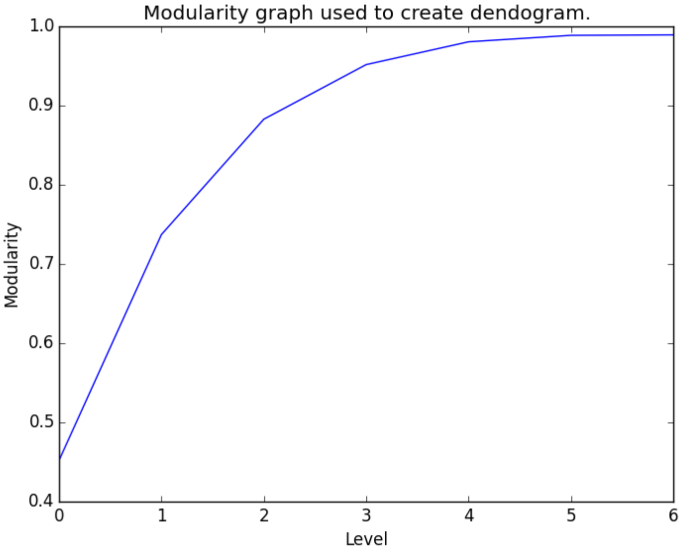

Figure 4: Modularity

5.1.1 Community sizes

Louvain algorithm terminates when it is no longer able to improve modularity by merging communities and in this case it found 206 distinct communities. Figure 4 shows that gains in modularity beyond 3rd level are small. Its high value above 0.9 suggests that communities at other dendrogram levels may be equally good. Community size distribution when partitioned optimally, shown on Figure 5, is concentrated around 3500 and follows a *roughly* normal distribution. Figure 6 shows that communities at dendrogram's 3rd level are clearly smaller with just a few greater than 500. These, as the algorithm progresses, would later be merged to compose larger ones. The question arises whether such merging is necessary since gains in modularity are small whereas community sizes are affected substantially. This issue is discussed further in Section 5.4.

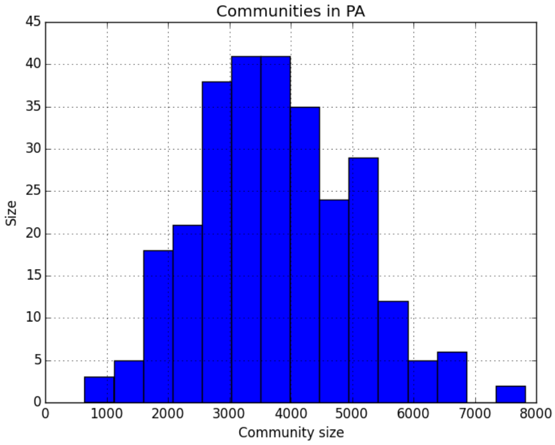

Figure 5: Optimal community size distribution

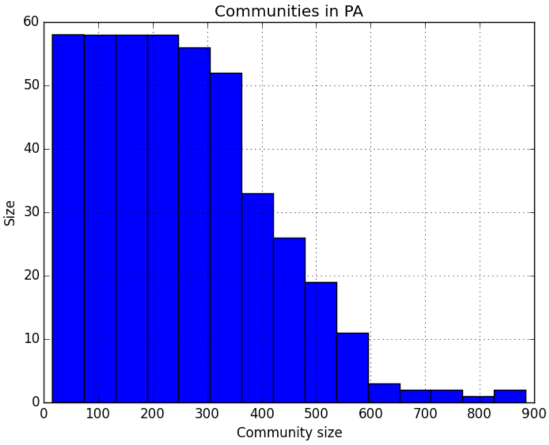

Figure 6: Community size distribution, level 3

5.2 Community layout

Communities in road networks may represent entities such as towns, villages or even local boroughs depending on the scale. Layouts of two communities found at the 3rd level shown on Figure 7 represent finer scale communities compared to Figure 8. Groups of nodes connected to each other in the middle and a few outgoing *links* resembling transit roads are clearly visible.

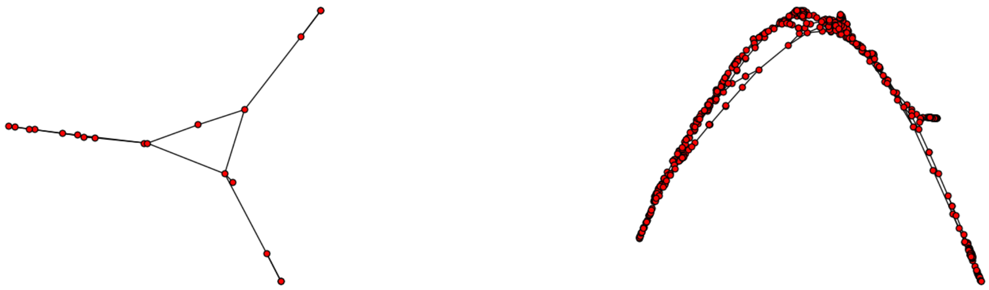

Figure 7: Smaller communities

5.3 Comparison with random graph

A random graph is a standard benchmark, however, given the size, sparsity and planarity of the network, it makes little sense to compare it against the whole graph. Nevertheless, detected communities of smaller size and higher density can be used instead. In this report, a community of $N = 817$ nodes and $K = 1153$ edges was selected and an equivalent random graph was generated using probability $p = \frac{K}{\binom{N}{2}} = 0.003$. Figures 9 and 10 show circular representations of both graphs whereas Figure 8 shows spectral representation of a community. Is clear that random graph fails to model sparse graphs well as their nodes usually have lower degrees exploiting spatial proximity and fewer distant connections as indicated on circular graphs. Spectral graph provides a very good visualisation *roughly* demonstrating planarity and local connectivity of nodes.

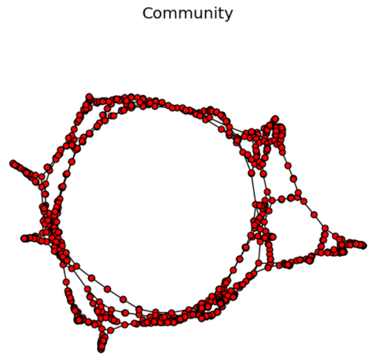

Figure 8: Community, spectral representation

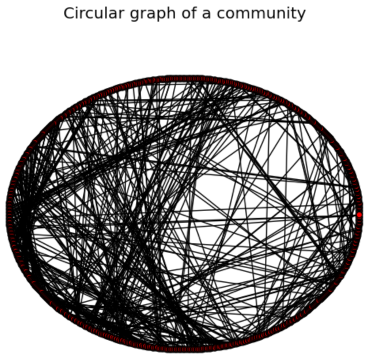

Figure 9: Community, circular representation

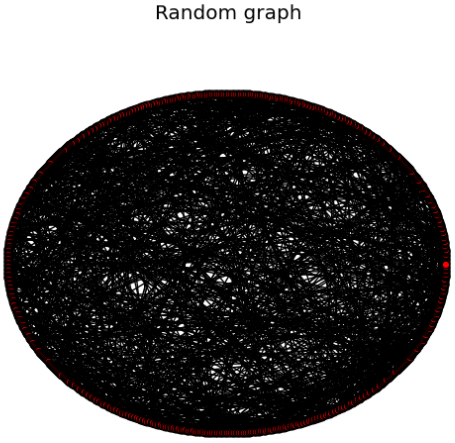

Figure 10: Random graph circular representation

It is useful to compare degree distributions of both models on Figures 11 and 12 which differ substantially. The selected community exhibits a sharp drop for degrees higher than 5 whereas a random graph has a maximal degree of 9. Similar behaviour can, although not shown here, be observed for other sample communities.

Note: Figure 11 takes into account only edges *within* the community and hence has no high-degree nodes.

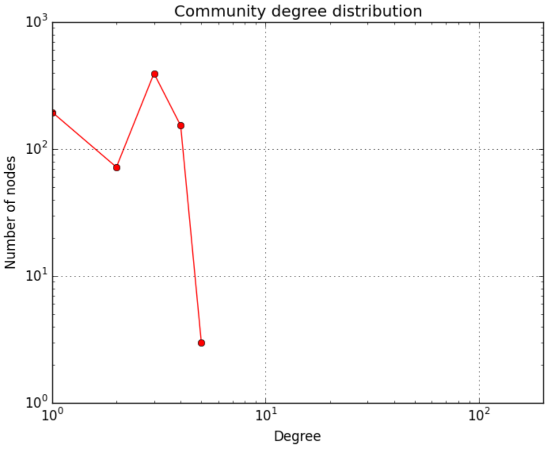

Figure 11: Community degree distribution

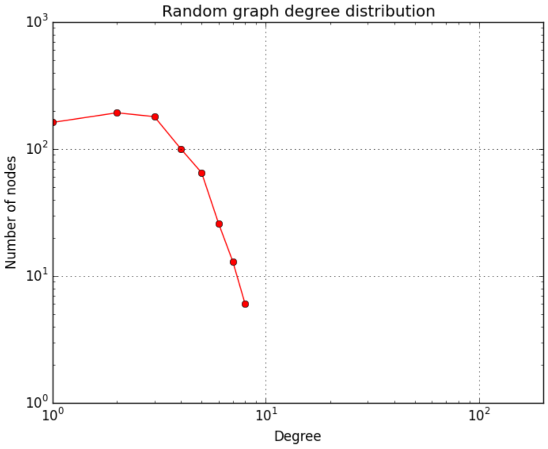

Figure 12: Random graph degree distribution

5.4 Limitations of Louvain algorithm

Really high value of modularity close to 1 suggests that Louvain method performed well and identified *communities* successfully. However, these results should be treated carefully as the method is heuristic and may lead to misleading results. In fact, Louvain method is prone to the *resolution limit problem* resulting from the relative difference in size of communities and of an overall graph [7]. It is useful to employ a different measure of assessing quality of a community [4], called conductance:

$$\phi(S) = \frac{c_s}{\min\left(Vol(S), Vol(V \setminus S)\right)}$$

, where the numerator corresponds to *cut surface* and denominator to *community volume.*

Figures 13 and 14 show scatter plots of conductances of all communities found at the optimal 7th and arbitrarily chosen 3rd level of a dendrogram respectively. It is clear that in both cases *quality* scores and community sizes lie in similar ranges within a dendrogram level suggesting that Louvain method found *comparable* communities at each iteration. One should notice, however, that mean *quality* at the 3rd level is around 0.05 whereas at the optimal one it is around 0.005. *Good* communities should be more densely interconnected within, as measured by volume, and less so with the remaining nodes indicated by the numerator. Obtained results show that Louvain method found *better* communities at a higher level with respect to conductance by an order of magnitude despite having similar modularity score.

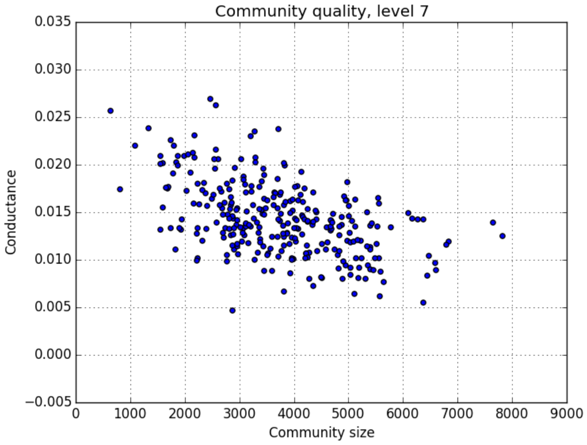

Figure 13: Community quality, optimal

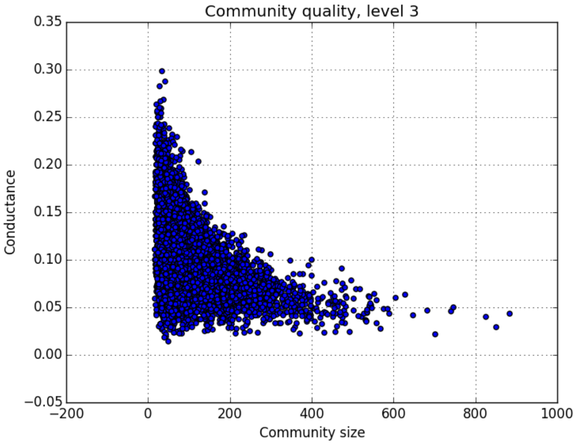

Figure 14: Community quality, level 3

6 Robustness analysis

Generic graphs have plenty of structural properties worth investigating depending on the underlying network. For graphs representing transportation systems such as roads or railroads, robustness is a critical property. Their purpose is to ensure connectivity and capacity for efficient operation. In case of this experiment, dataset has no weights rendering throughput analysis useless.

6.1 Average shortest path

A useful measure is to estimate how much *disruption* is caused by gradual degradation of a network by, for example, road-works which correspond to removing graph nodes. One could check how average shortest path between all pairs of points changes as the nodes are removed. It is, however, infeasible given the size of considered graph which would require calculation of 10^{12} paths for every few nodes removed. Also, one would have to keep track of disconnected nodes and consider just a giant component.

6.2 Giant component size

The goal of transportation networks is to ensure short point-to-point travel times but also to provide connectivity between as many areas as possible. In emergency, it is crucial to provide connectivity with lesser focus on travel time. The analysis of connectedness or, in other words, its giant component would help planners modify their designs to improve their resilience to failure.

6.2.1 Random failures

Road networks often experience many isolated failures which block a certain road or intersection. These may include local repairs, accidents or protests. Such *events* can be reasonably modelled as random and independent node removals.

Figure 15 shows how the size of a giant component changes as more nodes are removed at random. It turns out that the network copes relatively well needing approximately 22% of nodes to be deleted before its performance deteriorates rapidly. Such behaviour is expected as a lot of degree distribution *mass*, shown on Figure 2, lies at low degrees and removal of these, for example, dead-ends is not critical.

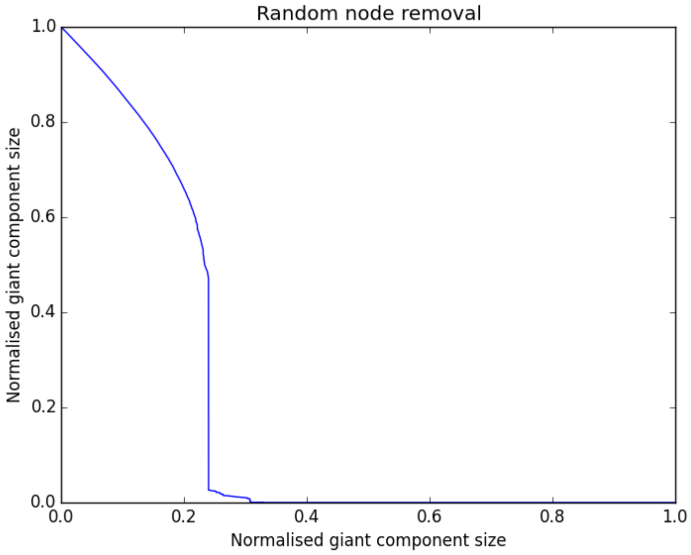

Figure 15: Network's resilience to random failures

6.2.2 Targeted failures

An increasingly more important type of a failure these days is a *targeted attack* which aims to render such systems useless. Unfortunately many networks used these days are vulnerable to attacks such as DDoS on Internet websites. A similar situation is harder to imagine for transportation networks however greater reliance on Internet-connected devices and a constant threat of cyber-warfare makes such scenario plausible.

Figure 16 shows network's *performance* as more nodes are removed from the network in descending degree order. It turns out that giant component quickly breaks down into multiple smaller ones and after removing 14909 nodes its size goes down to only 7. It is hardly surprising as its degree distribution is skewed heavily towards low-degree nodes shown on Figure 2. Critical threshold appears at approximately 1% corresponding to all nodes of size greater than 4 removed. Such result offers a stark contrast to random failures presented above showing network's *inability* to cope with a simple targeted attack. Similar results likely apply to other transportation networks revealing how easily an entire region may be paralysed.

Note: Figure 16 was trimmed to focus on smaller fractions of nodes removed as the critical threshold is really low.

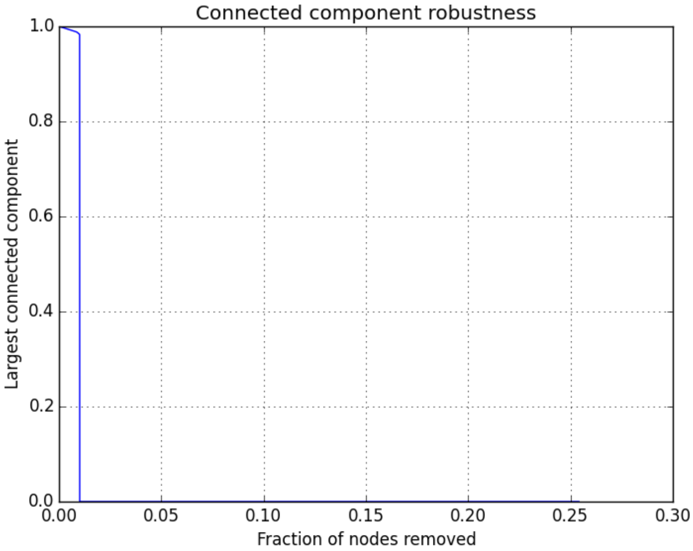

Figure 16: Network's resilience to targeted attacks

7 Future work

Although this report covers a few aspects of the dataset, there is plenty of scope for futher investigation. This section *roughly* describes approaches which were given up due to their computational complexities despite promising potentials [5].

7.1 Spectral clustering

Spectral clustering is good at detecting communities in sparse graphs tackling the problem by investigating properties of graph's adjacency matrix using eigen-analysis. Its potential has been shown in many papers [5] however it is hard to execute on a small machine given sheer size of adjacency matrix of approximately 10^{12} elements. Figure 17 offers a visualisation of adjacency matrix where each element is represented by a black pixel showing network's sparsity.

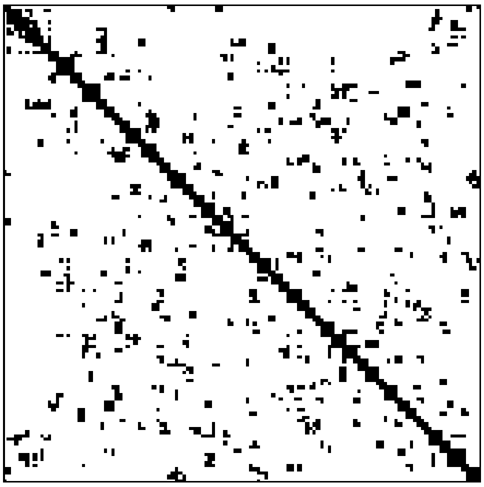

Figure 17: Sparsity of adjacency matrix[1]

7.2 Community versus core-periphery structure

Another direction of research uses a different definition of a community for planar networks which takes into account real-world examples. One could imagine that, on a scale of whole regions, single towns can be classified as separate communities however this approach fails when large metropolis are taken into account. Louvain algorithm fails to distinguish different boroughs which are usually densely connected within themselves and also with other boroughs. Such topologies may be treated more efficiently as *core-periphery structures* [6].

7.3 Network Community Profile

Real world networks are empirically known to have a particular structure of their Network Community Profiles which display *conductance* of best node clusters for each size [6]. Nevertheless, the number of possible clusters grows with a factorial and one needs to resort to heuristic approximations which still require a lot of computational power to execute.

8 Conclusions

The performed analysis revealed certain interesting features of the RoadNet-PA dataset representing Pennsylvania's road network. Most of experimental results are consistent with theoretical expectations for a large, sparse and planar graph. It has been found to have a large diameter and long average shortest path. Also, its degree distribution with rapid fall-off for larger degrees is very different from other networks studied in lectures.

Investigation also showed that a road network handles random failures relatively well as expected from experience, however it is useless when more sophisticated failure patterns are applied. The analysis has also suggested limitations of Louvain algorithm for large and sparse graphs as modularity score has not substantially changed beyond the 3rd level while community structures have. However, a different measure - conductance - confirmed that *merged* communities are better and hence conductance may be useful to include in termination conditions.

Lastly, it is important to note that there is plenty of opportunity for further analysis as many promising methods required more computational resources than there were available as described in Section 7. It would also be useful to cross-reference detected communities with geographical data.

References

[1] Hu, Y., *Matrix: SNAP/roadNet-PA.* http://www.cise.ufl.edu/research/sparse/matrices/SNAP/roadNet-PA.html, AT&T Labs Visualization Group.

[2] Leskovec, J., *Pennsylvania road network. Dataset information.* https://snap.stanford.edu/data/roadNet-PA.html.

[3] Backstrom, L., *Anatomy of Facebook.* https://www.facebook.com/notes/facebook-data-team/anatomyof-facebook/10150388519243859.

[4] Leskovec, J., Lang, K., Dasgupta, A., Mahoney, M., *Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters.* Internet Mathematics 6(1) 29–123, 2009.

[5] Shi, J., Malik, J., *Normalized Cuts and Image Segmentation.* http://www.cs.cmu.edu/̃jshi/papers/pami_ncut.pdf.

[6] Leskovec, J., *Community Detection: Modularity Optimization and Spectral Clustering.* http://web.stanford.edu/class/cs224w/slides/16spectral.pdf.

[7] Fortunato, S., Barthelemy, M., *Resolution limit in community detection.* http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1765466/.