Small Scale Diffusion Testing

Claire Kelling, Ashton Verdery

August 31, 2017

Abstract

Through this study, we would like to better understand how clustering in a network affects diffusion. This will help policy-makers to determine intervention strategies in given communities, after being able to draw conclusions about network vulnerability. First, we examine isomorphic networks for small regular networks. By closely examining these smaller networks, we then are able to test our diffusion functions on larger networks and this will be able to give insight on larger network structures. For these simulations, we utilize the SI and SIR diffusion models. We took the average number of time periods until the network is either saturated or no one else is infected. For these graphs, we use the complete set of isomorphic graphs for regular, degree 3 networks of node size 6, 8, and 10. We compare the effect of the clustering coefficient of the network, and other network characteristics on the diffusion time and impact.

Keywords: clustering coefficient, network, isomorphic graphs, diffusion modeling

1 Introduction

How topological features of social and other networks can affect the realization of spreading processes such as epidemics of infectious diseases, the diffusion of information, or social influence contagion that motivate behavioral change upon those networks is a topic of great interest in a variety of fields ([Morris, 1993]; [Rogers, 2010]; [Centola and Macy, 2007]). However, the nature of complex networks means that it is difficult to vary one feature without changing others, which impedes the ability to draw meaningful conclusions about the effects of varying distinct elements of network topology on the relative speed and ultimate size of the spreading process.

For instance, levels of network clustering, the tendency for one's friends to be friends with each other, are a characteristic feature of the network topology of many human social networks, with non-trivial levels of clustering endowing networks with "small world" properties ([Watts and Strogatz, 1998]). [Newman, 2003] studied the performance of Susceptible Infected Recovered (SIR) disease models on networks with tunable clustering levels and found that as clustering is increased, the size of the ultimate epidemic declines but the epidemic threshold, the level of infectivity needed for the epidemic to take off (hence, the speed of epidemic realization), is decreased. However, other models show the opposite properties; for example, [Keeling, 2005] finds that increases in clustering increases epidemic thresholds.

Kiss and Green [2008] attempt to resolve this debate by noting that Newman's model, while varying clustering with preserved mean nodal degrees, alters levels of dispersion in the distribution of degrees such that increases in clustering lead to more degree distribution dispersion, which in turn affects the epidemic threshold, speed, and size of spreading processes. As they note, "[t]o study the theoretical effects of varying one network property (e.g., clustering), one would ideally like to generate multiple networks with all properties identical, except the property of interest. This is easier to say than do, as in practice different network properties may constrain each other, or not be independent" ([Kiss and Green, 2008], page 1). Of course, the examples they give regarding discrepancies between Newman's and Keeling's results illustrate this point: while Keeling's model preserves degree distributions and thus changes the conclusions about topological effects on spreading processes drawn from Newman's model, it does not constrain other macro-structural network features such as network diameter, the length of the longest shortest path.

In an effort to take a different look at this debate, in this paper we consider all isomorphic permutations of small, connected, regular random graphs. We use the complete set of all isomorphic networks with six, eight, and ten nodes ([Meringer, 1999]). According to [Harary and Norman, 1953], "[t]wo points P and Q of a graph are called adjacent if the line PQ is one of the lines of G. Two graphs G and G', each of n points, are called **isomorphic** if there exists a one-to-one

correspondence between the points of G and those of G' which preserves adjacency." If two graphs are not isomorphic they are called **different** [Harary and Norman, 1953]. A **k-regular graph** is a graph in which each node has exactly degree k [Meringer, 1999].

For the purposes of our work, first we look at how topological features covary with one another and then we look at diffusion processes on these networks. The focal variables we are particularly interested in are the clustering coefficient, the number of degrees, the cut set, and the diameter of the network.

We limit our analyses to the isomorphic regular graphs for six, eight and ten node regular networks, all with degree 3. We believe that this will be able to give insight on larger network structures. By examining closely these smaller networks, we then are able to test our diffusion functions on larger networks.

2 Data and Methods

In order to understand the results of our analysis, we first describe our simulation methods and variables of interest. As described in the introduction, there are several network characteristics that we are interested in relating to diffusion outcomes. For the purposes of our analysis, we also use two different diffusion functions: the Susceptible-Infected (SI) model and the Susceptible-Infected-Recovered (SIR) model. We use simplified versions of both of these models for our simulation purposes.

2.1 Network Characteristics

We are interested in the relationship between several key network characteristics. The basic characteristics we are interested in are diameter, number of nodes, cut set, and number of triads. This last characteristic, number of triads in the network, is to gain information about the clustering coefficient. This would be particularly useful as [Heath and Parikh, 2011] have recently developed a method that takes "triangle and single edge degree sequences as input and generate a random graph with a target clustering coefficient." Therefore, if we are able to better understand the effect of the clustering coefficient on diffusion, we can use these algorithms to generate random graphs with target clustering coefficients and draw conclusions about the structural risk of that network. We can also use this analysis to compare two networks and determine comparative structural risk of these networks.

2.2 Diffusion Functions

The first diffusion function we use, the SIR model, takes a random node at first as infected, and with a probability of infection proceeds to infect connected nodes in the network [Walker, 2014]. However, in this first diffusion function, there is a probability of infection and if the neighboring node is not infected, the diffusion function stops and the neighboring node can no longer be infected from that starting node. The neighboring node can still be infected from other neighboring nodes, but not from the original node. In other words, this diffusion function incorporates a level of immunity, where a node can remain uninfected at the end of the process. In other words, the SIR model, or **Susceptible-Infected-Recovery** model, incorporates this type of behavior where individuals are susceptible and may be infected with a certain probability. After they are infected, they have a certain probability of recovery. In this case, we use a very simple SIR Model where the probability of recovery is 100% after one time interval, and they cannot be infected again from other nodes.

Referring to Figure 1, below, there is an example of an SIR model. At t=1, there is an initial (yellow) node that is infected, in the lower right. At t=2, all neighboring nodes of the initially infected nodes have a certain probability of being infected. In our case, at t=2, only one neighboring node is infected, to the left of the initial node. All other neighboring nodes of the initial (yellow) can no longer be infected through the initial node. So, at t=3 two more nodes are infected from the red node in t=2. At t=4, another node is infected from the newly infected nodes in t=3. The newly infected node in t=4 fails to infect any other nodes in t=5 and all other nodes have recovered fully, so the diffusion process ends.

In the second diffusion function, many of the diffusion function properties are similar. For example, there is still a probability of infection from node to node. However, this diffusion function

proceeds until every node has been infected. Therefore, using this diffusion function, in the long run every node will be infected in this model. This diffusion function is called the SI Model, or **Susceptible-Infected**, where there is no chance for those who have been infected to recover.

We study the difference in time of diffusion in these two diffusion functions. In the first diffusion function, SIR, it may be of interest to also examine how many people are infected at the end. This is uninteresting in the second diffusion function because everyone is infected at the conclusion of the process. An example of each of these two functions can be seen below in Figures 1 and 2. In the graphs presented in these figures, there are two triangle, 8 nodes, and diameter 2.

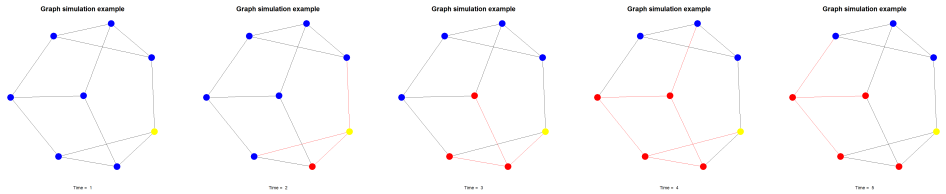


Figure 1: SIR Model

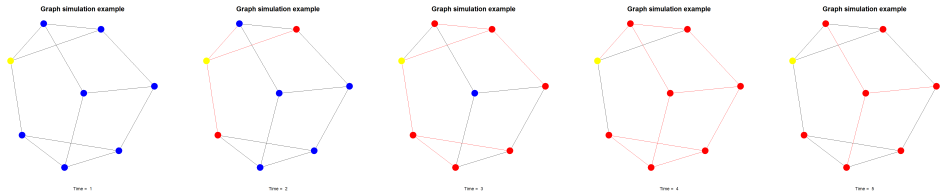


Figure 2: SI Model

For our six, eight and ten node networks, we simulate 10,000 diffusions using our two diffusion models. We took the average number of time periods until the network is either saturated or no one else is infected. We also measured the average number of nodes that were infected at the end of the simulation. For SIR models, this is an interesting measure but is simply equal to the number of nodes in the network for the SI model. For this simulation, we used a probability of infection of 50%. The first node that is infected is chosen at random for each simulation. For these graphs, we use the complete set of isomorphic graphs for each number of nodes (6, 8 and 10) for degree 3. The complete set of isomorphic networks used for this analysis can be seen in the Appendix.

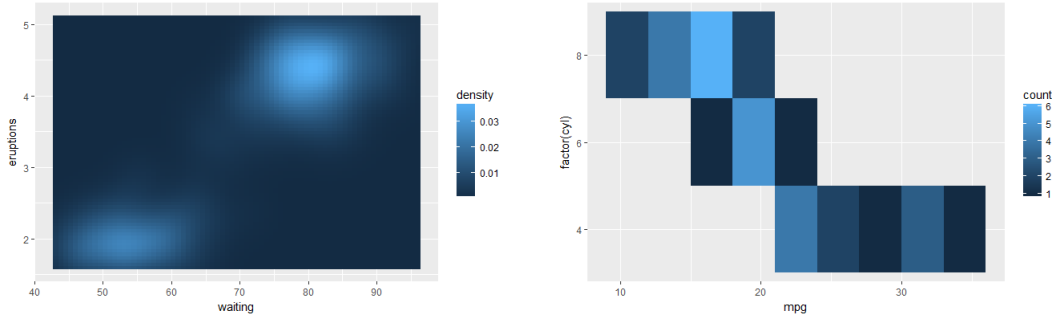
2.3 Measures

After we complete these simulations, we will assess two of the outputs, as described above: the number of time iterations until the diffusion process as completed, as well as the number of nodes that were infected at the completion of the process.

The first way we will visualize these responses is through a heat map. The most common way that heat maps are often used is over continuous space, as in Figure 3(a). We will be examining a heat map over discrete space, such as in the example seen in Figure 3(b). This figure uses a classic built in R data set and the geom_tile function in R to visualize the relationship between two variables: miles per gallon, the number of cylinders, and the number of cars in each one of these ranges [Wickham, 2009]. In our heatmaps, we will consider the Diameter of the network on the x-axis, the clustering coefficient on the y-axis, and the shade of the tile will represent a third variable: either the number of iterations, or the number of nodes infected at the end of the diffusion process. This will help us to gain an impression to the response surface and the relationship between these three variables.

An additional way that we will see the relationship between these variables is through replicated a graph similar to the one in Figure 4, which was created by [Newman, 2003]. In our example, we will consider the y-axis to be the number of nodes infected, the left side of the x-axis to be the number of time iterations, and the right side of the x-axis to be the clustering coefficient. We will also color the points on these figures so that we can also control for the number of nodes in the networks.

After examining all of these plots, we will be able to draw conclusions on the relationship between these variables.



(a) Continuous Space (b) SIR Model

Figure 3: Discrete Space

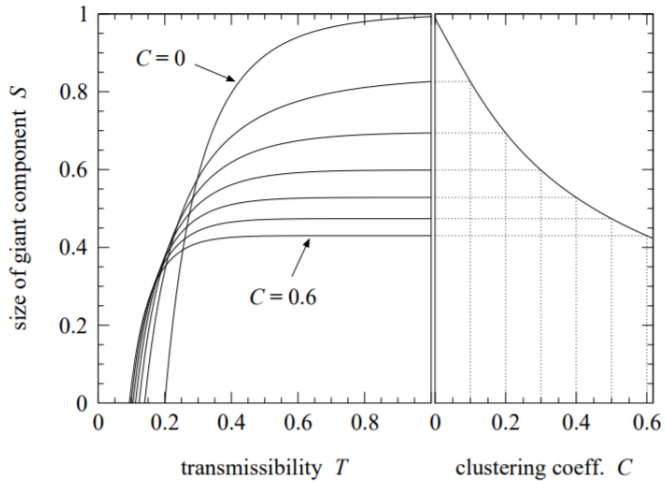
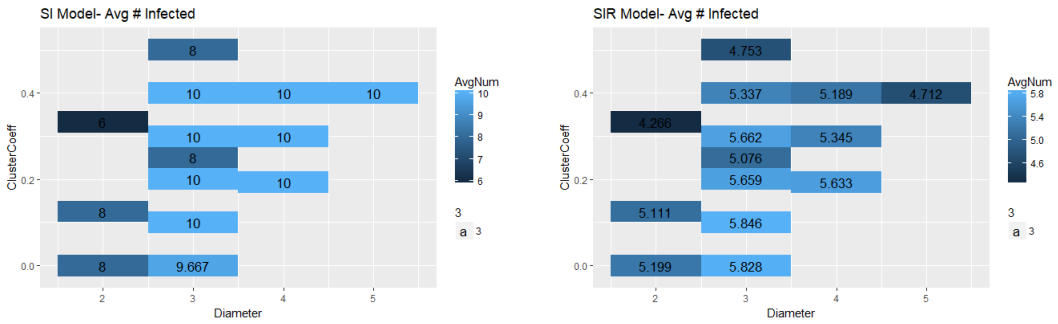


Figure 4: Example of a Dual Plot [Newman, 2003]

3 Results

First, we compare the relationship between the clustering coefficient, diameter of the network, and the average number of nodes in the network who were "infected" through the 10,000 simulations.



(a) SI Model (b) SIR Model

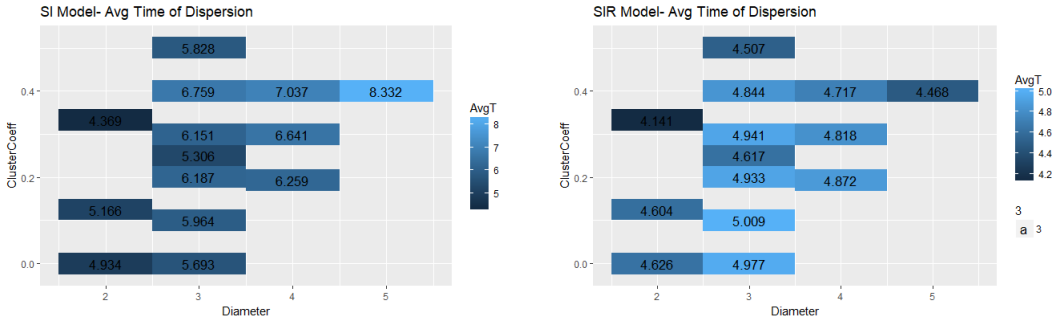
Figure 5: Average Number Infected

We notice that for the SI model, the number of people infected after the complete diffusion process is exactly equal to the number of the nodes in the network, so it is largely uninteresting. For the tiles that do not have integers, there are multiple networks with this combination of

diameter and clustering coefficient, so the number of people that were infected is averaged across both networks.

In the SI Model, we notice that there seems to be very little of a pattern moving from left to right, as the diameter increases. However, as the number of clustering coefficient increases, the number of people who are infected seems to increase.

Now, we do the same analysis for both the SI and the SIR model except for using the average number of time increments until the diffusion has completed as our variable of interest.

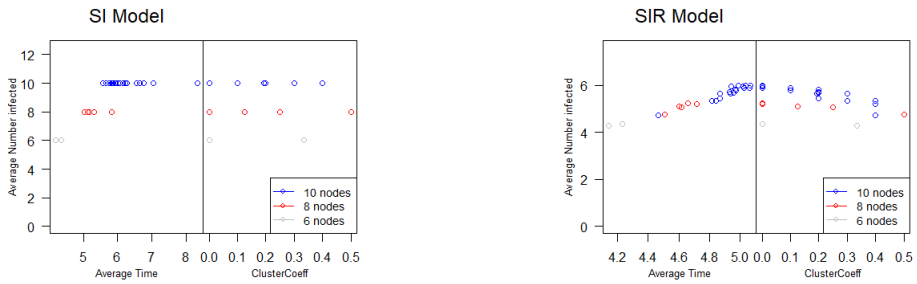


(a) SI Model (b) SIR Model

Figure 6: Average Time of Dispersion

We notice that in this representation of the data, for the SI Model, there is a clear relationship with the Diameter and the time that it takes for a disease to diffuse through a network. As the diameter increases, the time generally increases for the disease to diffuse. However, the relationship between the clustering coefficient and time is not as clear. For networks of at least diameter 3, it seems that as the clustering coefficient increases, so does the average time of diffusion. However, this is not the case for networks of diameter 2 where the average time of diffusion seems to vary inconsistently with the clustering coefficient.

Finally, we consider a different representation, where we can compare the average time of diffusion, the average number of nodes that are infected, the number of nodes in the network, and the clustering coefficient of the network.



(a) SI Model (b) SIR Model

Figure 7: Average Time of Dispersion

In Figure 5a, we see that in the SI model, these relationships are relatively uninteresting. The number of nodes that are infected after the diffusion process is the complete number of nodes in the network. However, we see that as the number of nodes increases, so does the the time of infection, similarly to our findings in relation with diameter above.

In Figure 5b, we see that in the SIR model, the relationship between these variables is a bit more complex. In general, as the time increases, so does the number of nodes that are infected. There is also as somewhat linear relationship between the clustering coefficient and the number of people infected, which is more clearly of positive slope when there are 10 nodes, as opposed to the 6 and 8 node networks.

4 Future Work

There are several important next steps in our work to further develop this research. An immediate next step is to develop this analysis for 12 node networks, in addition to the 6, 8, and 10 regular degree 3 networks that we have already analyzed. So far, we have analyzed 26 graphs, 19 of them having 10 nodes, 5 networks having 8 nodes, and only 2 networks having 6 nodes. If we were to extend this to 12 node graphs, we would have 85 additional isomorphic, degree 3 graphs to analyze. We anticipate that this would add quite a bit to our analysis. The edgelists have been made available by [Heath and Parikh, 2011].

We would also like to use this work to then simulate for larger networks and draw conclusions about the effects of clustering, diameter, nodes, and the diffusion model on different diffusion outcomes such as the number of nodes infected and the time until saturation or the diffusion process is complete.

Also, as suggested by Ashton, we will also consider the analysis of "Bounds on Clustering and Diffusion." (NEEDS DEVELOPMENT- not sure what is meant here)

References

Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

Frank Harary and Robert Z Norman. *Graph theory as a mathematical model in social science*. University of Michigan, Institute for Social Research Ann Arbor, 1953.

Lenwood S Heath and Nidhi Parikh. Generating random graphs with tunable clustering coefficients. *Physica A: Statistical Mechanics and its Applications*, 390(23):4577–4587, 2011.

Matt Keeling. The implications of network structure for epidemic dynamics. *Theoretical population biology*, 67(1):1–8, 2005.

Istvan Z Kiss and Darren M Green. Comment on "properties of highly clustered networks". *Physical Review E*, 78(4):048101, 2008.

Markus Meringer. Fast generation of regular graphs and construction of cages. *Journal of Graph Theory*, 30(2):137–146, 1999. URL http://www.mathe2.uni-bayreuth.de/markus/reggraphs.html#GIRTH7.

Martina Morris. Epidemiology and social networks:. *Sociological Methods & Research*, 22 (1):99–126, 1993. doi: 10.1177/0049124193022001005. URL http://dx.doi.org/10.1177/0049124193022001005.

Mark EJ Newman. Properties of highly clustered networks. *Physical Review E*, 68(2):026121, 2003.

Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

Ryan Walker. Going viral with r's igraph package, 2014. URL https://www.r-bloggers.com/going-viral-with-rs-igraph-package/.

Duncan J Watts and Steven H Strogatz. Collective dynamics of'small-world'networks. *nature*, 393 (6684):440, 1998.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

5 Appendix

5.1 All Isomorphic Degree 3 Graphs (6,8,10 nodes)

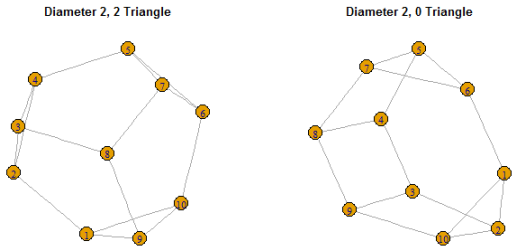


Figure 8: 6 Node Graphs

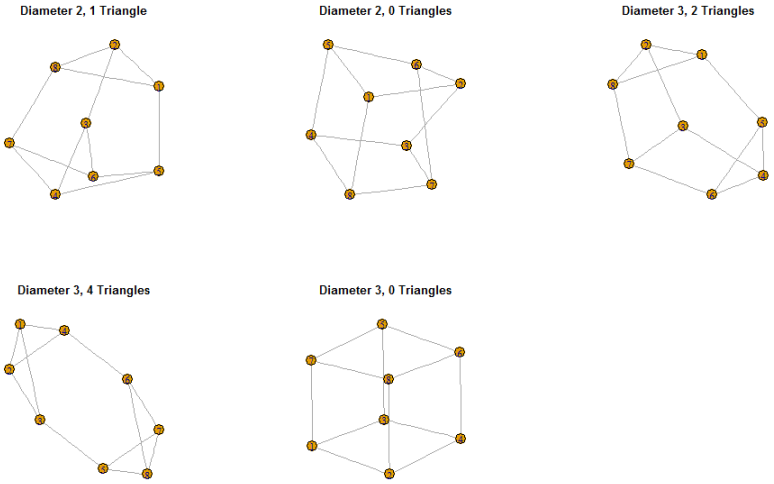


Figure 9: 8 Node Graphs

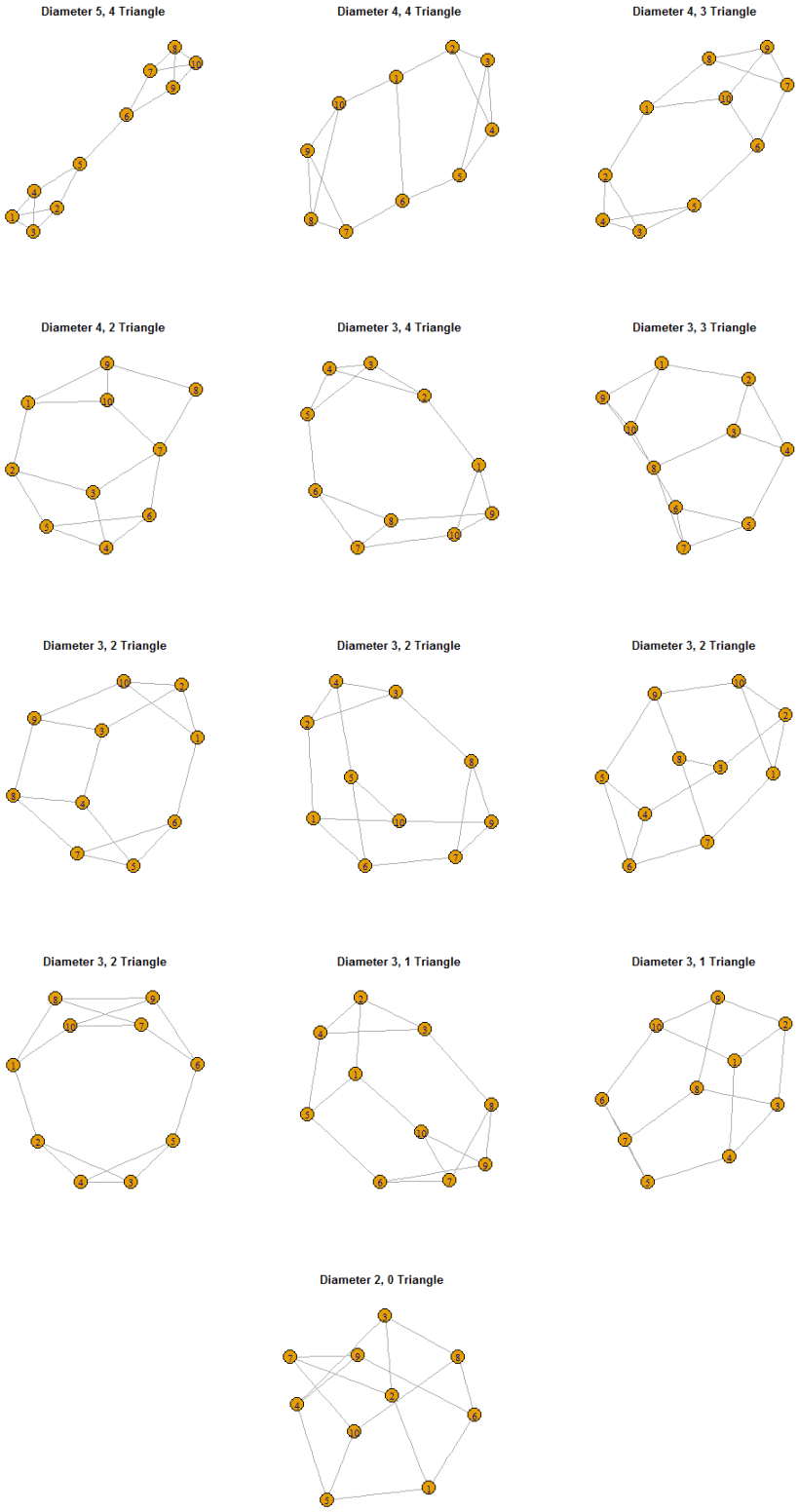


Figure 10: 10 Node Graphs