Small Scale Diffusion Testing

Claire Kelling, Ashton Verdery

December 21, 2017

Abstract

Structural features of social networks have important implications for the spread of diseases, information, and health behaviors. For example, a disease may travel through one social network much more quickly than through another given certain characteristics of the network's structure, like the level of clustering in the network. However, the precise effects of different network structural features on different types of diffusion are unknown. Part of the reason for this gap in knowledge is that many structural features of networks are highly correlated and it is often difficult, and sometimes impossible, to change one feature without altering others, which precludes isolating the effects of different structural features. For instance, prior work has demonstrated mixed effects of network clustering on diffusion depending on whether or not dispersion in nodal degrees is fixed. In this study, we attempt to move beyond this impasse. To do so, we examine the complete sets of a series of small, regular, isomorphic networks. These networks differ in their levels of clustering and other features, but they allow us to make explicit comparisons between diffusions propagating in networks with different levels of clustering but otherwise identical features. We test SI and SIR diffusion models. As our outcomes, we analyze both the speed of network saturation or diffusion failure and the ultimate size of the network reached by the diffusion. We compare effects of network clustering and other structural features on these outcomes and find clear relationships between the variables.

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 is a topic of great interest in a variety of fields ([Morris, 1993]; [Rogers, 2010]; [Centola and Macy, 2007]). Experimental evidence, for instance, demonstrates that health behaviors spread differently in networks with different topological structure (CLAIRE CITE CENTOLA 2010 "SPREAD OF BEHAVIOR IN AN ONLINE SOCIAL NETWORK EXPERIMENT," SCIENCE). 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 spreading processes.

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 that clustering has 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 preserving mean nodal degrees, also alters the 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 degree 3, 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]. All networks that we consider are undirected.

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 nodal cut set, and the diameter of the network. The clustering coefficient is the ratio fo the number of closed triplets in the network to ratio of open triplets. The nodal cut set is defined as the minimum number of nodes that would need to be removed from the network to disconnect it into two components. And the diameter of the network is the length of the longest shortest path.

We limit our analyses to the isomorphic regular graphs for six, eight, ten and twelce node regular networks, all with degree 3 and we find that there is a relationship between the clustering coefficient, diameter, and our outcomes of the number of nodes infected and the number of time periods it takes for the diffusion process to be complete. 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

We first describe our simulation methods and variables of interest. We study two different diffusion functions: the Susceptible-Infected (SI) model and the Susceptible-Infected-Recovered (SIR) model.

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." A better understanding of how network clustering affects diffusion will help researchers draw conclusions about the structural risk of different 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]. We test a case where the probability of passing the infection in each time step is p=0.5. 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 chance of failure to transmit the infection as well as recovery from the infection and subsequent immunity to it; as such, not all nodes will be infected 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 to the first diffusion function. For example, there is still a probability of passing the infection from node to node; we test p=0.5. However, this diffusion function proceeds wihtout the chance for nodal recovery. Because the probability of passing the infection is high, it is likely to proceed until every node has been infected. This diffusion function is called the SI Model, or **Susceptible-Infected**, where there is no chance for those who have been infected to recover.

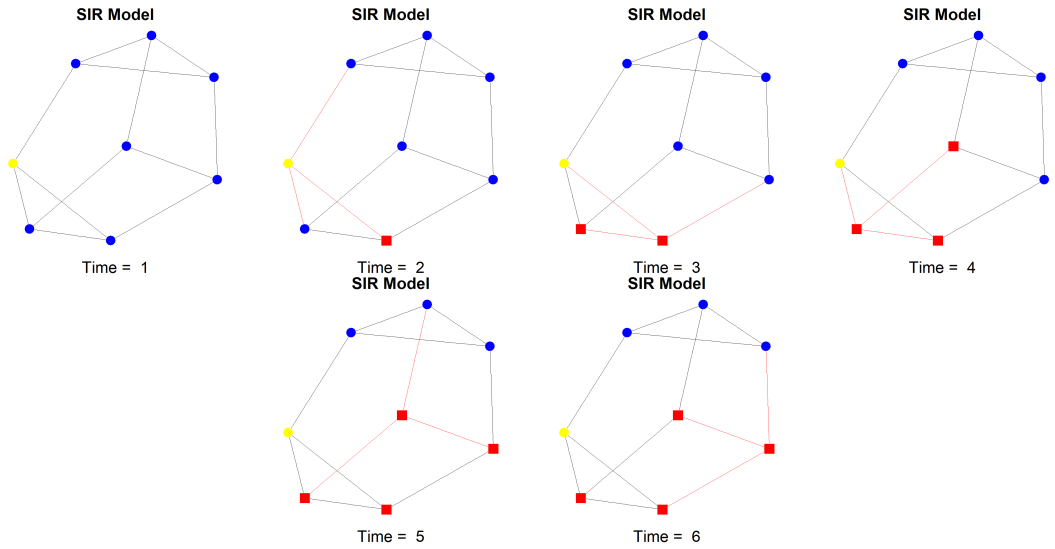


Figure 1: SIR Model

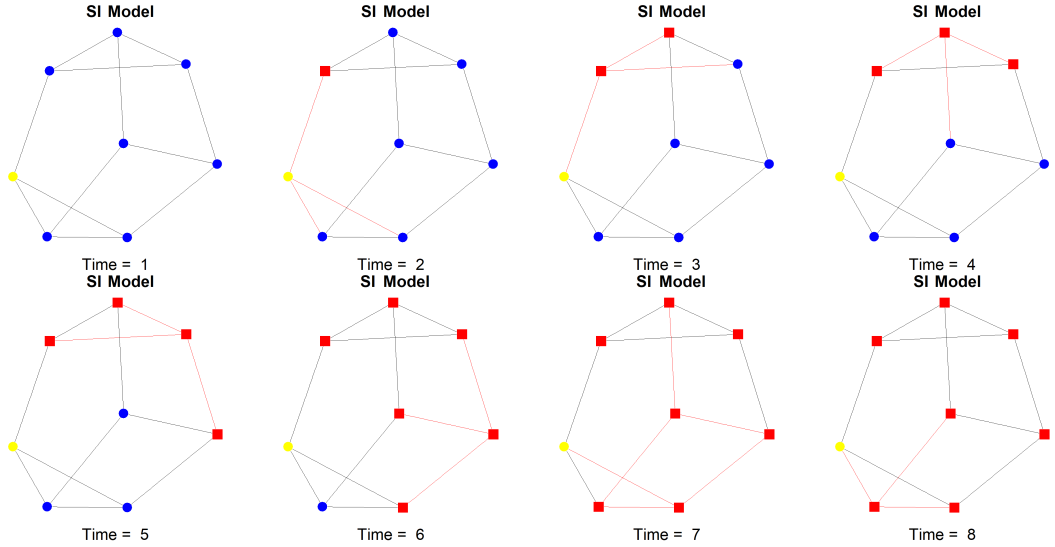


Figure 2: SI Model

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 completes, as well as the number of nodes that were infected at the completion of the process.

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 unlikely to be interesting in the second diffusion function because we expect that everyone will be 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.

2.4 Approach

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, 10, and 12) for degree 3. The complete set of isomorphic networks used for this analysis can be seen in the Appendix.

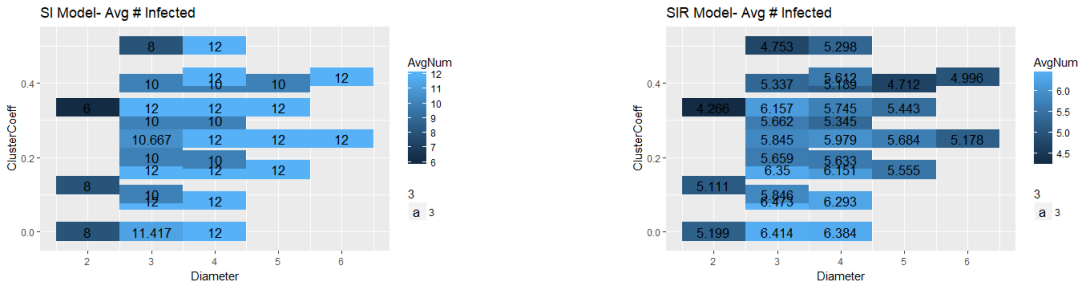
The first way we will visualize these responses is through a heat map. The most common way that heat maps are used is over continuous space, such as the examination of a response surface model over two covariates. We will be examining a heat map over discrete space. Prior work has used these approaches to visualize the relationship between multiple variables in a classic built in R data set and the geom_tile function in R to visualize [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 by analyzing visualizations of the type plotted in [Newman, 2003]'s pioneering work examining the relationships between transmissibility, clustering coefficients, and the size of the giant component. 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 see the relationship with the number of nodes in the networks and these other variables of interest.

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

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 3: 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, as we would expect. 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. So, for example, there are two isomorphic networks that have a diameter of 3 and a clustering coefficient of 0, so the average number of nodes that were infected was averaged across all of these networks to get this grid cell. Also, there are no networks in our set of degree three, and 6,8,10, and 12 node networks that have diameter 2 and a clustering coefficient of 0.25, which is why there is not a colored grid cell in that spot.

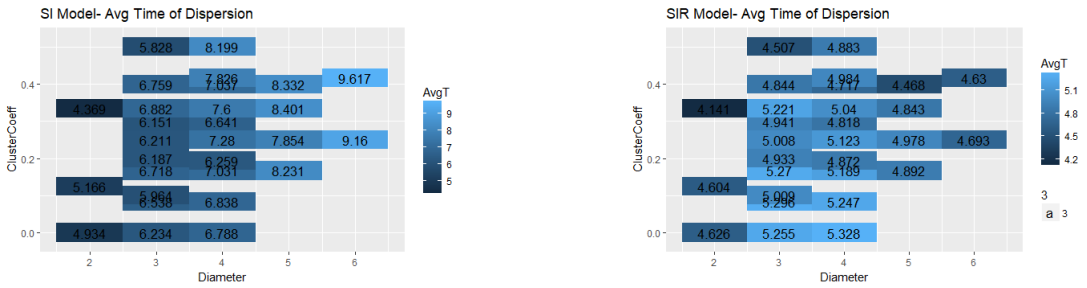
In the SI Model, we notice that there is very little of a pattern moving from left to right, as the diameter increases. However, as the clustering coefficient increases, ie moving upwards, the number of nodes who are infected increases.

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.

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, 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 varies inconsistently with the clustering coefficient.

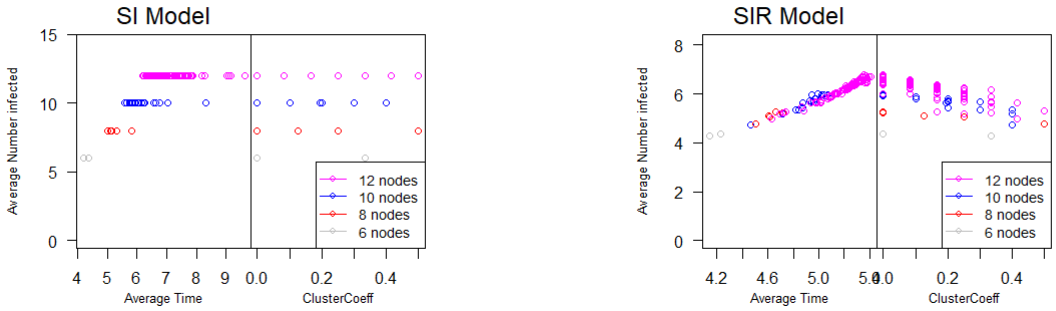
Finally, we consider a different representation, where we can compare the average time of diffusion, the average number of nodes that are infected, the number of nodes in the network, and the clustering coefficient of the network.

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



(a) SI Model (b) SIR Model

Figure 4: Average Time of Dispersion



(a) SI Model (b) SIR Model

Figure 5: Combined Figure

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 accounting for complex contagion diffusion [Centola and Macy, 2007]), where nodes only become infected when at least two of their neighbors are infected. Prior work has suggested that complex contagion models provides a more realistic model of the spread of health behaviors in social networks, whereas the SI and SIR models that we examined here more closely proxy the spread of infectious diseases.

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.

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)

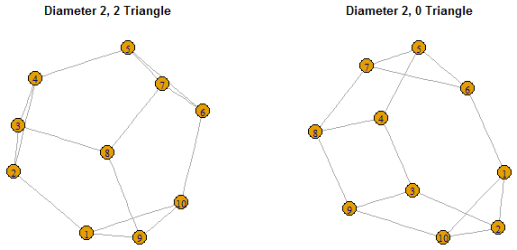


Figure 6: 6 Node Graphs

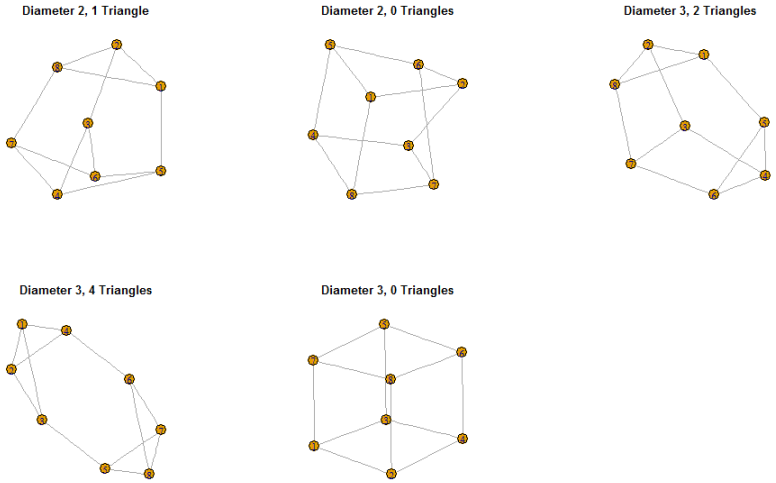


Figure 7: 8 Node Graphs

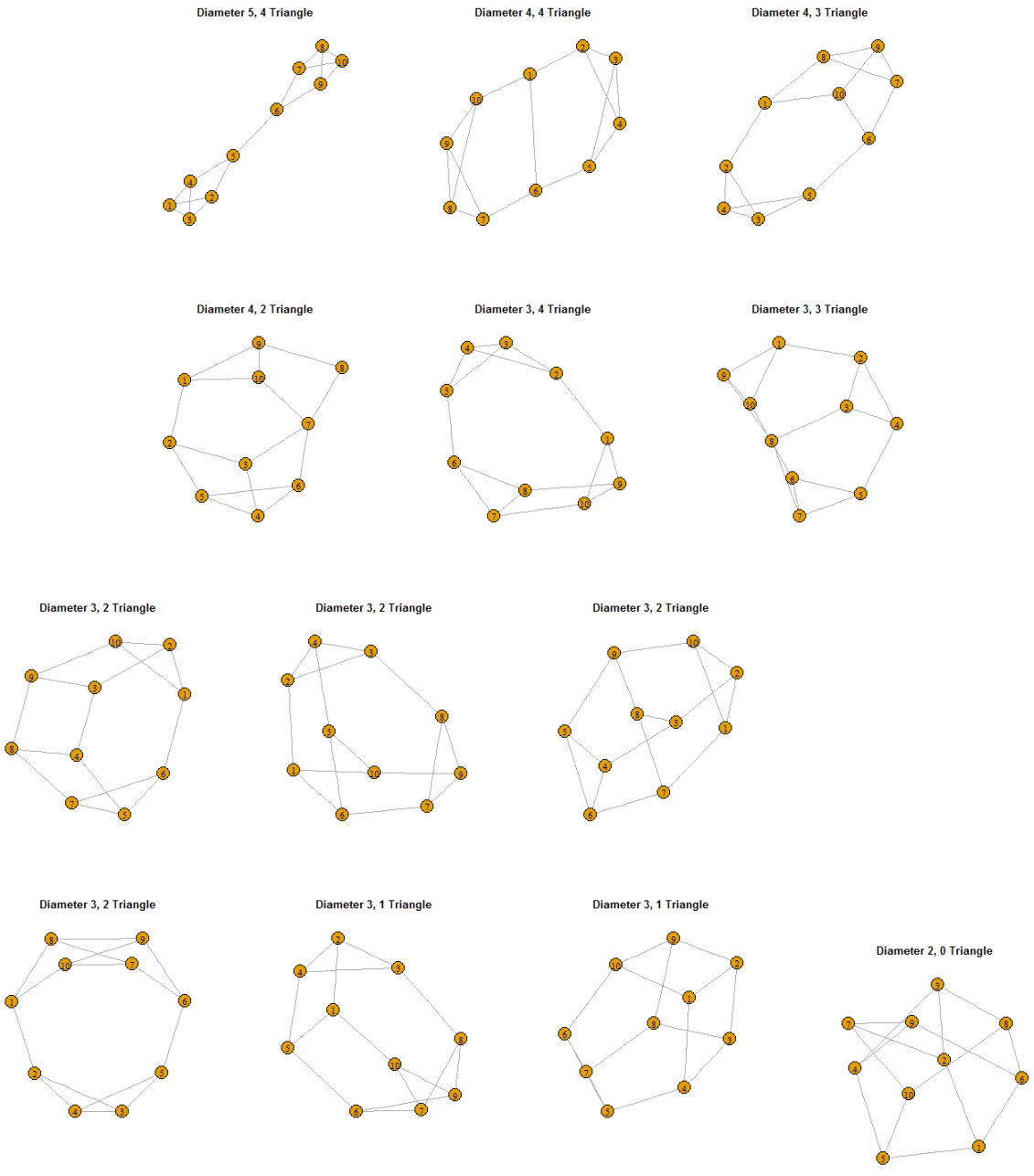


Figure 8: 10 Node Graphs