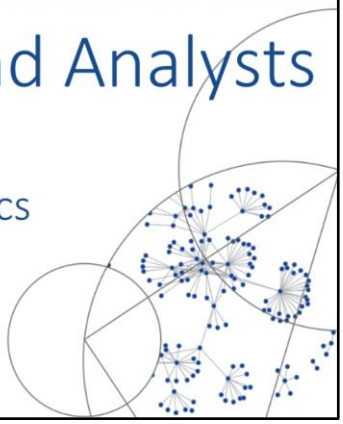
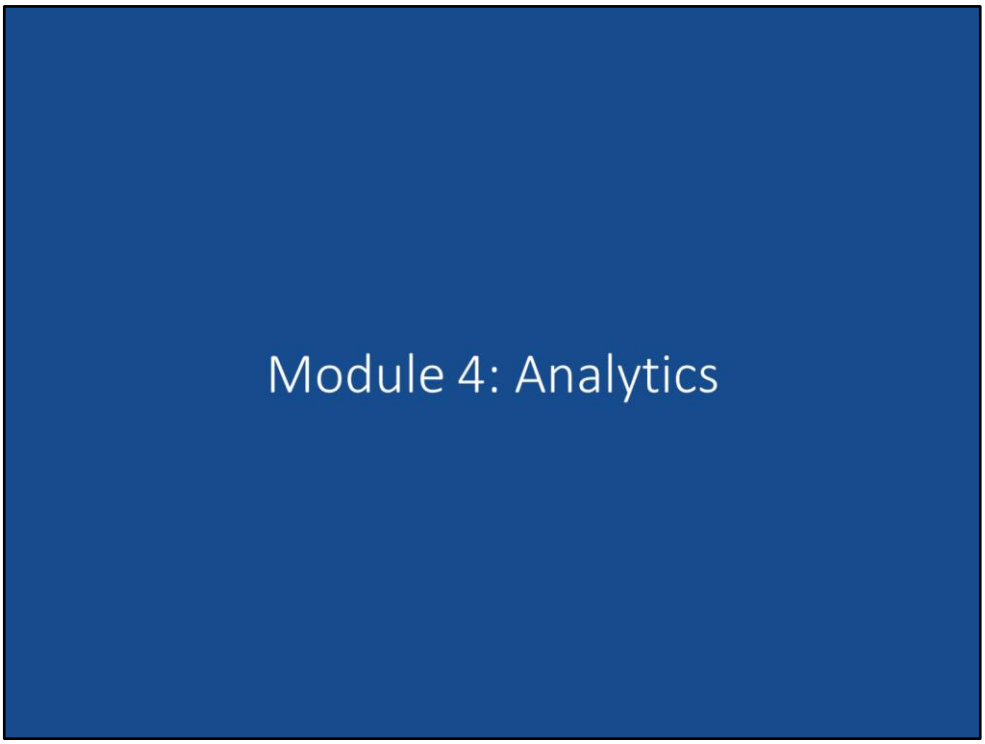
Yale University

Social Network Analysis for Criminal Justice Practitioners and Analysts

Module 4: Analytics



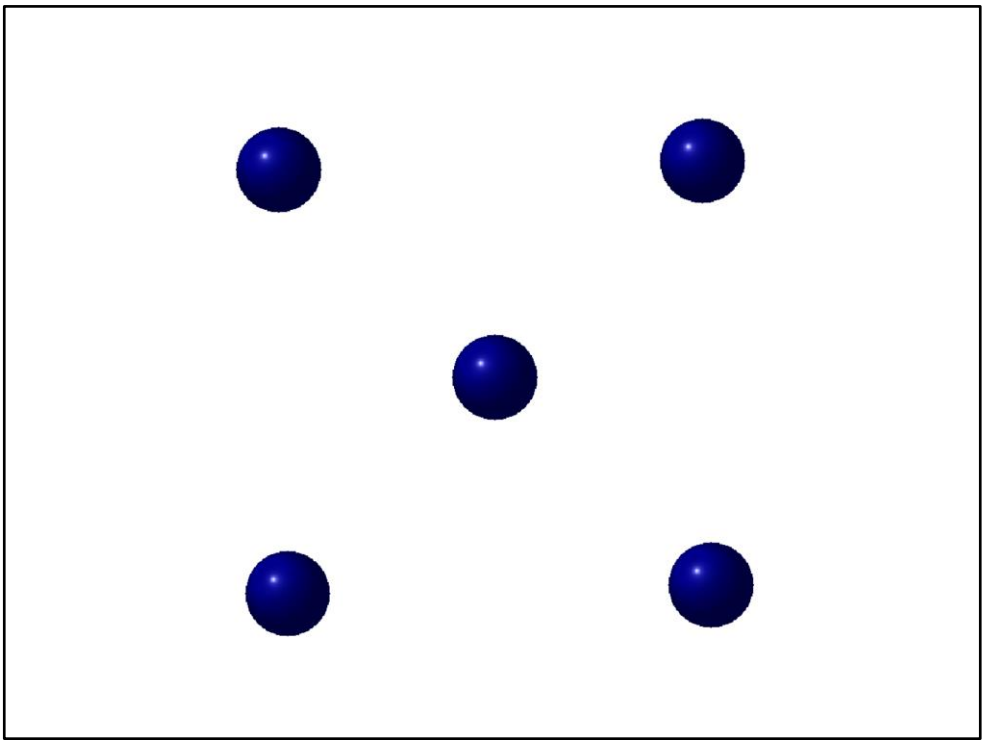
Andrew V. Papachristos



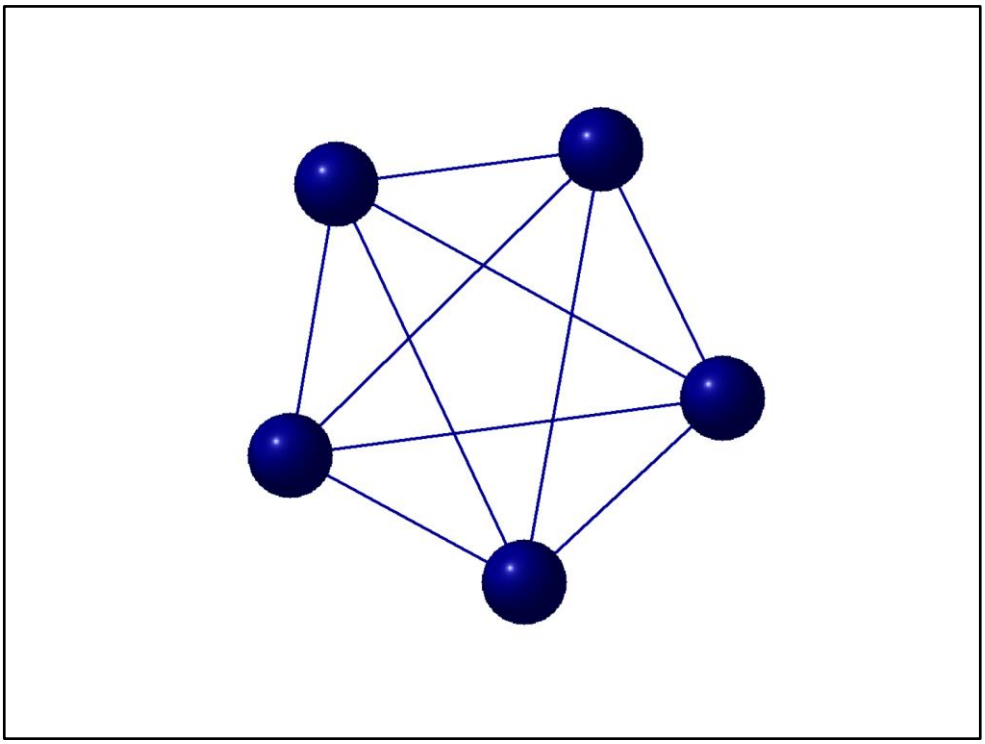
Welcome to Module 4: Analytics. This module introduces some fundamental statistics and basic measurements of social networks. The goal of this module is for participants to become familiar with some of the different ways to measure and analyze social networks. This tutorial uses illustrations, examples, and questions to introduce the following social network measures: density, components, degree, k-core, distance, brokerage, and neighborhood. Let's get started with density.

Density

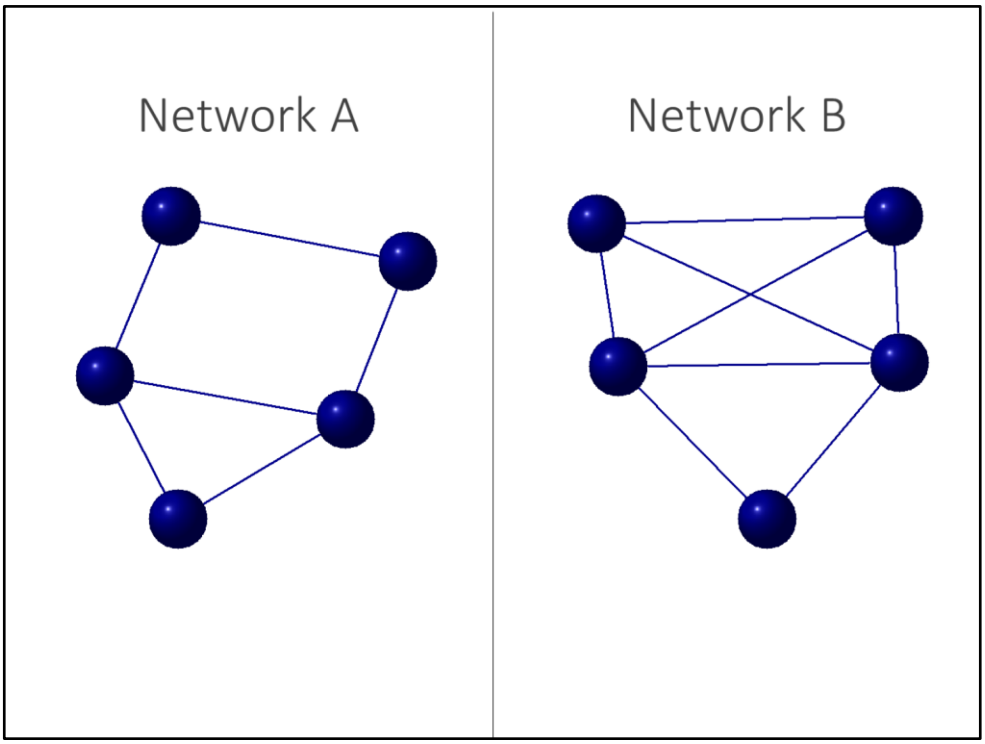
How dense or sparse is a network? How does the connectedness of one network compare to the connectedness of another network?



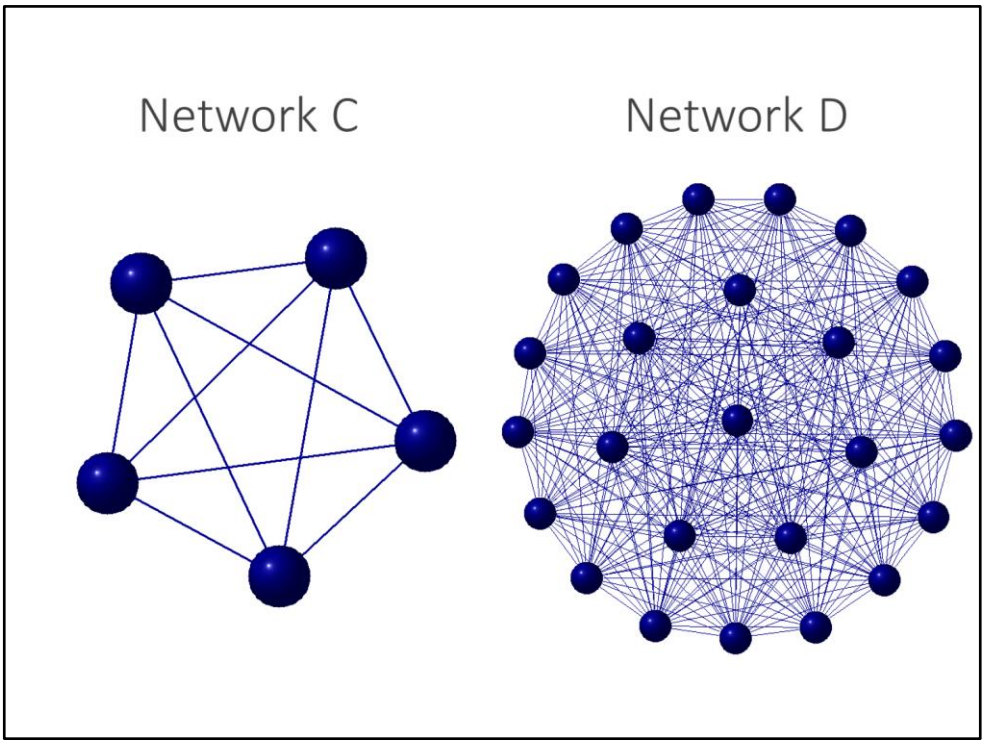
Density is a single value calculated for an entire network. The formula to calculate density is as follows: density = (the number of edges present in a network) / (the total possible edges of a network). By “total possible edges,” we mean the maximum number of edges that could exist between a given set of nodes if every single node was connected to every other node. In other words, network density is the proportion or percent of all possible edges that actually do exist within the network. Since the formula for density includes the total number of possible ties, it effectively adjusts for the number of nodes in the network. This means that we can compare the density of networks of different sizes. Above is a set of 5 blue nodes with 0 edges. The density of this network is 0 because 0% of the total possible edges are present in this network. How many possible edges could exist between this set of 5 nodes?



Here are the 5 same blue nodes arranged slightly differently. As you can see, all of the possible edges between this set of nodes are present in this network. There are 10 present edges out of the 10 maximum possible edges. 10 divided by 10 is 1, so the density of this network is 1. Alternatively, we can say that 100% of the possible edges are present in this network.



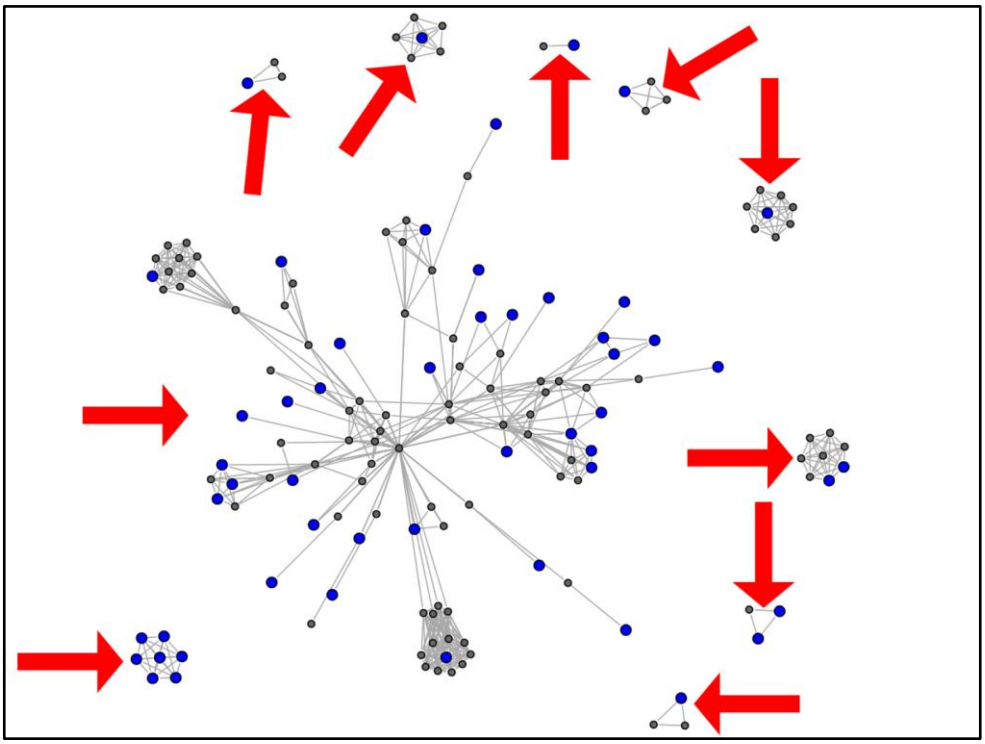
Here are two different networks with the same number of nodes. Both Network A and Network B have 5 nodes, so the total possible edges in these two networks will be the same. Network A has a density of 0.6 (6 edges present in the network / 10 possible edges in the network = 0.6). Network B has a density of 0.8 (8 edges present in the network / 10 possible edges in the network = 0.8). Network B is more dense than Network A.



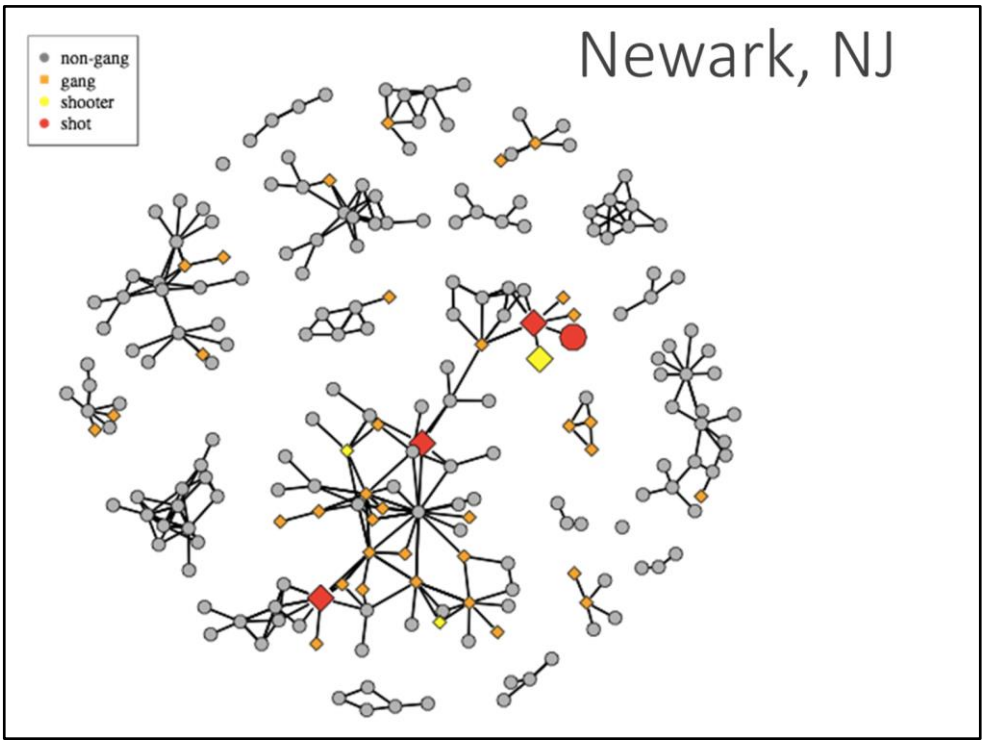
In this example, the two networks look very different. Network C has 5 nodes and 10 edges, and Network D has 25 nodes and 300 edges. Even though these networks are very different in size, they are the exact same in density. All of the possible edges between 5 nodes are present in Network C, and all of the possible edges between 25 nodes are present in Network D. That means that both of these networks have a density of 1.0, or we can say that 100% of all possible ties are present in both networks.

Components

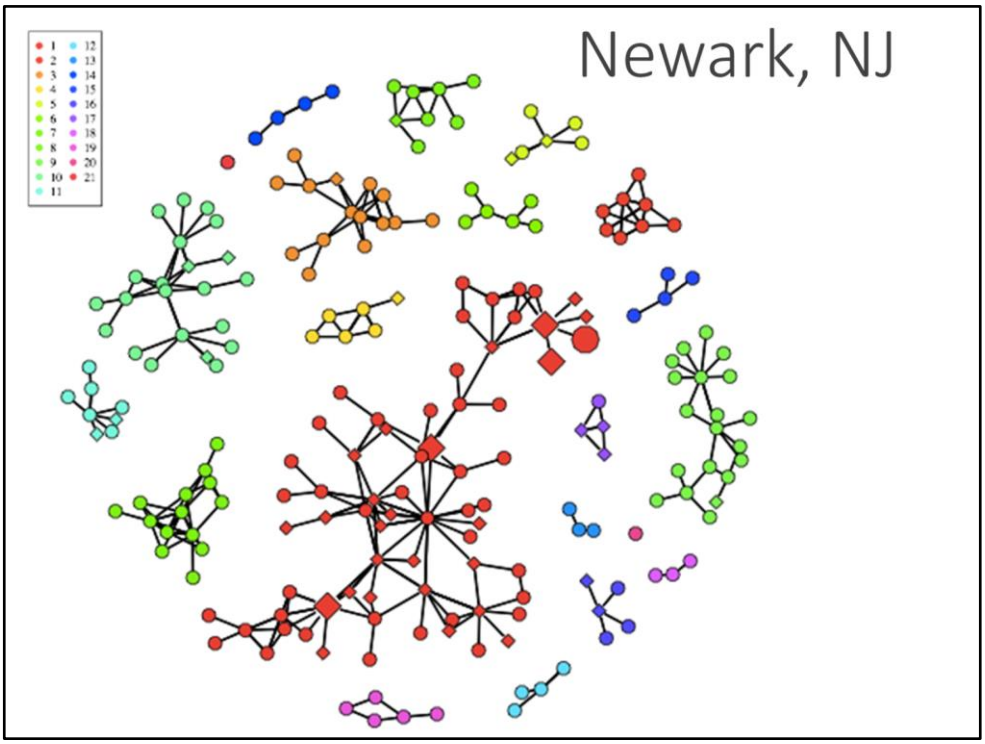
How many separate parts are there to the social network?



Components are the different sub-networks within a network or all of the discrete sections making up a network. Components are a useful social network measure because they show how connected or unconnected a network is. A single network can contain multiple components. Components can be useful for identifying subgroups within a larger network. The actions and activities within the separate components might be discreet and not overlap with the other components. Isolates (i.e., nodes with 0 ties) are their own component in a network. Analysts might be interested in only the largest component of the network, which is the component connecting the largest number of nodes. There are 10 components in this network as indicated by the 10 red arrows. There are no isolates in this network. The largest component is in the middle of the network connecting the most nodes.



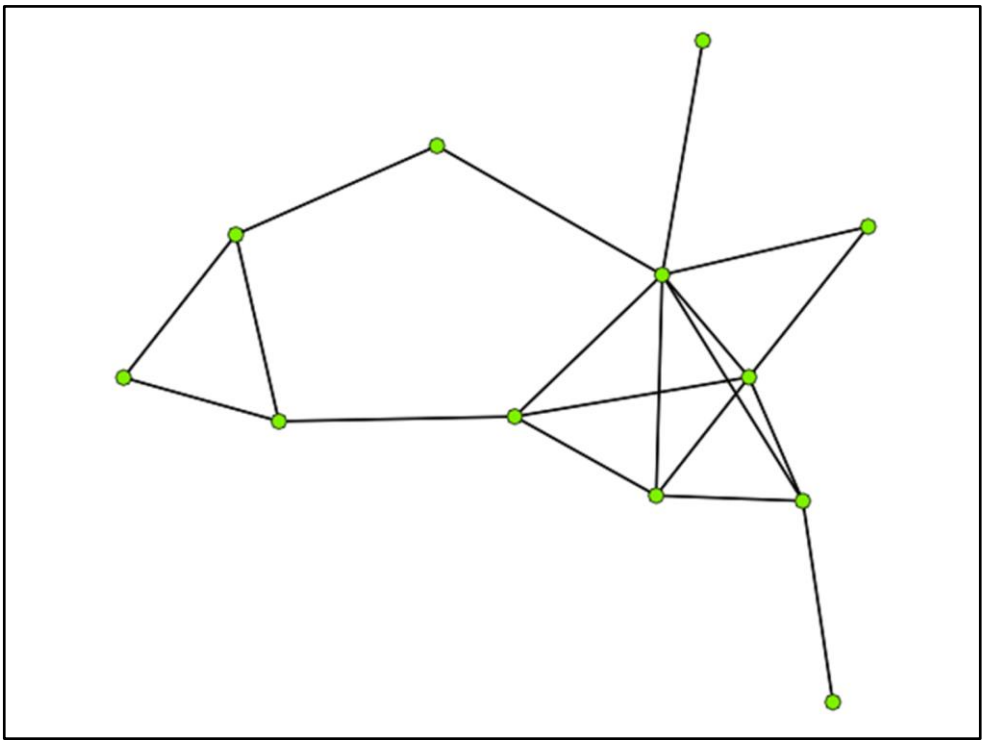
This image is an arrest network from Newark, New Jersey. The shape and color of the nodes distinguish gang members and non-gang members and individuals who were shot or were shooters. How connected is this network? As you can see, there are multiple components in this arrest network. There is one largest component that connects many nodes by their various arrests. There are a few isolates in this network identifying people who were arrested alone and were not part of any co-arrest. And there are some components of various sizes representing different clusters of arrests over time.



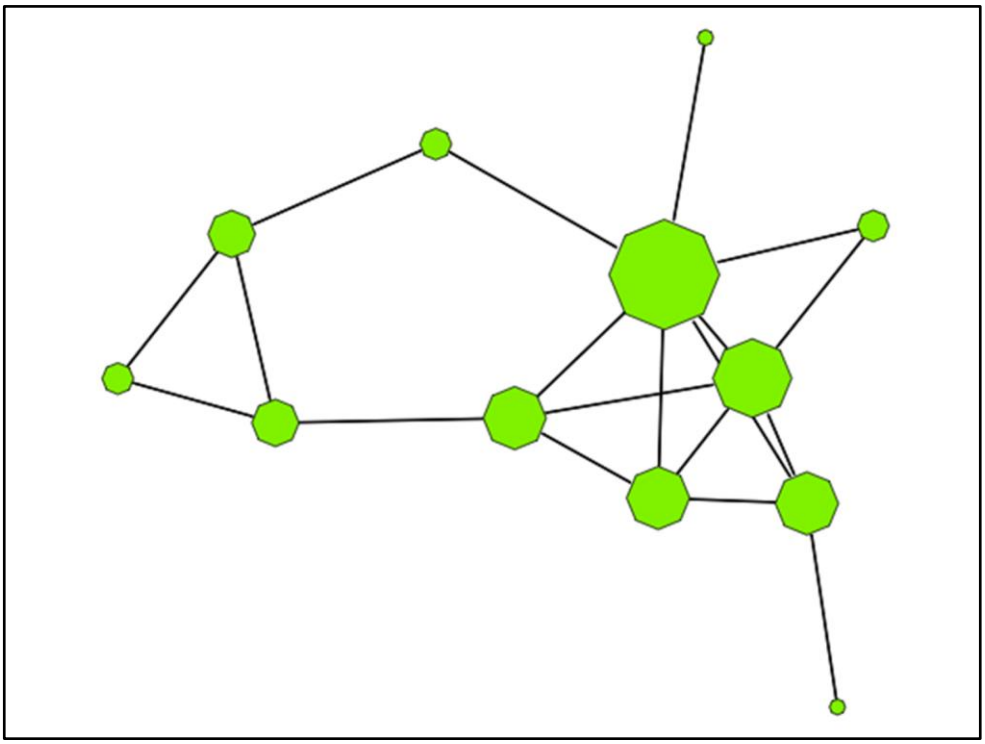
This image is the same arrest network from Newark, but here nodes within the same component now share a color. There are 21 components in all including two isolates. We can use this information to make comparisons across the different components within the network. By analyzing the components, we know that there are multiple separate pieces to this arrest network. We could compare the size and connectivity of the different components. We could compare the concentration of violence within the components. Or we could identify which components were at risk, and which components were not at risk.

Degree

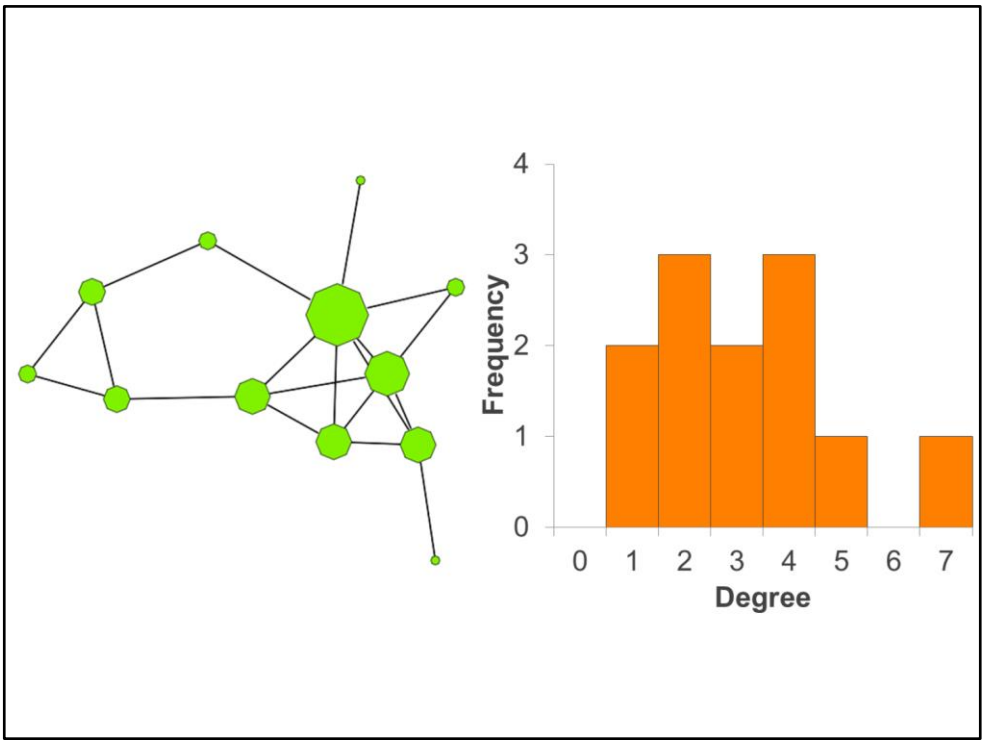
Where is the action in the network? Who are the most central people in a network? Who is popular and who is not?



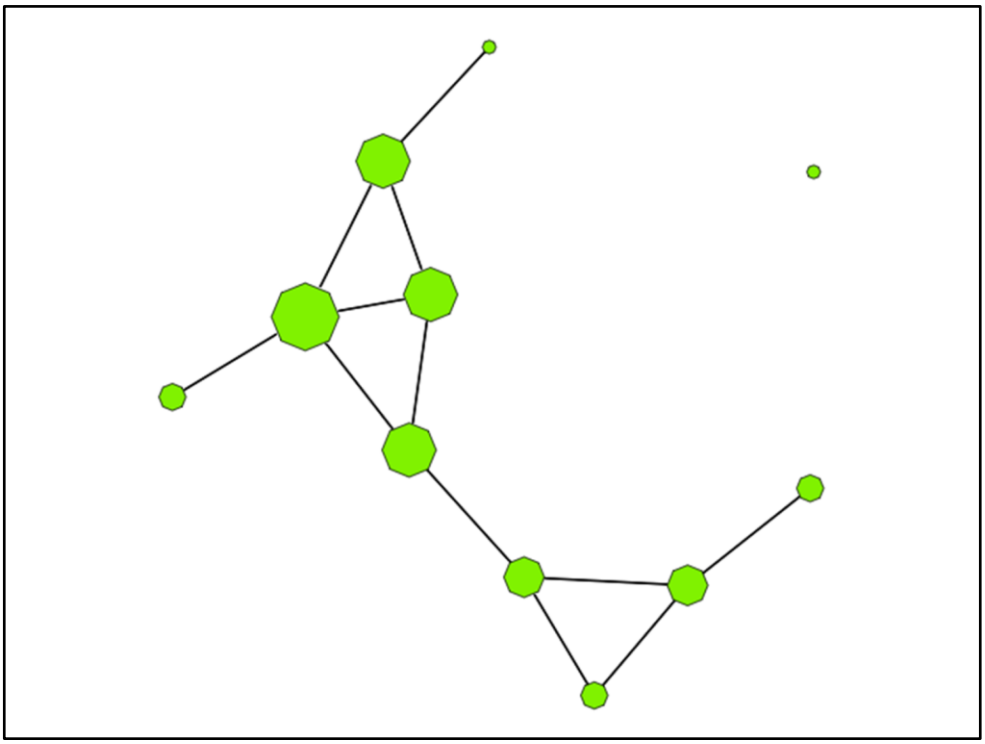
Degree is a network statistic that answers questions about activity, centrality, and importance. Degree is the literal counting of ties each node has in the network. Each node has a degree score based on its number of ties, so degree is a property of each node rather than a property of the network as a whole. Degree tells us who is the most and least popular in a network, or where the action is and is not in a network. A network has a distribution of degrees from all of the nodes. Isolates have a degree of zero, so if there are isolates in the network the degree distribution would range from zero as the lowest to the value of the highest degree. Using this example network, we can count the edges for each node to examine the degree. Which node has the highest degree? Which nodes have the smallest degree?



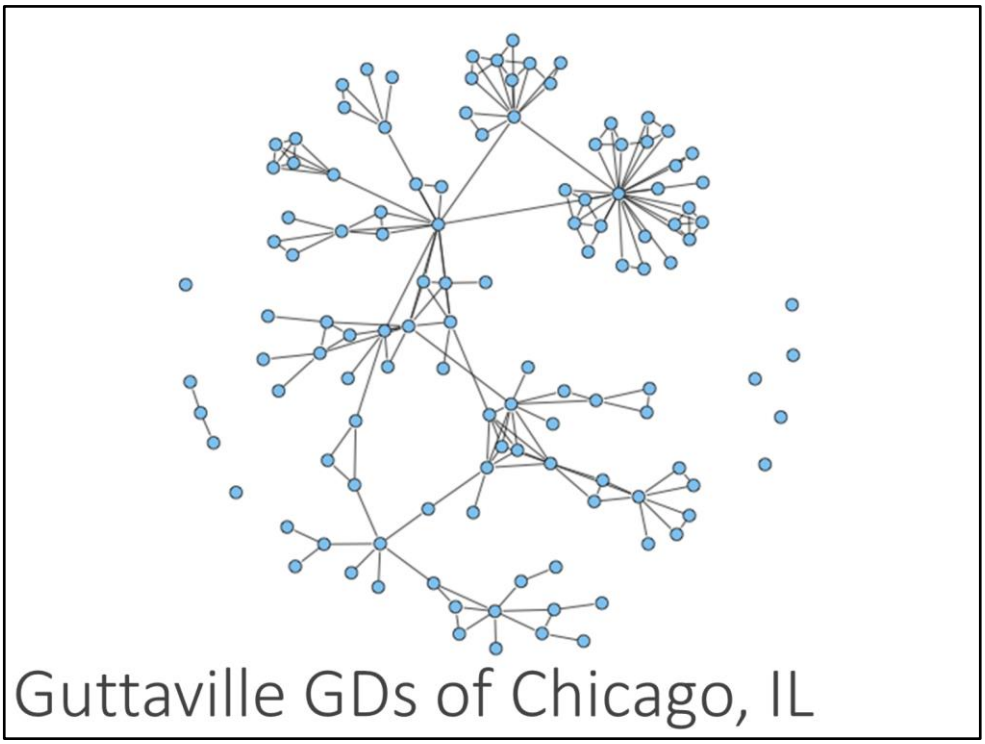
This is the same network as on the previous slide. In this image, the size of the node is proportional to each node's degree. Nodes with the highest degrees are larger in this network, and nodes with the smaller degrees are smaller. This network visualization trick makes it even easier to see which node is the most important in this network.



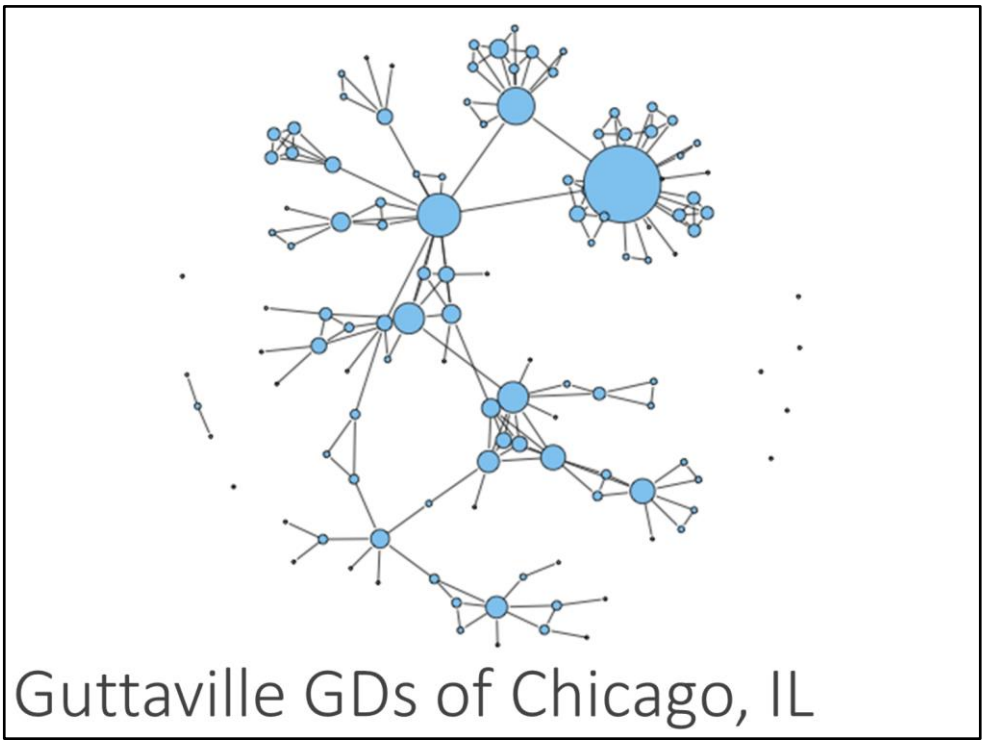
The histogram on the right of this slide shows the degree distribution for this network. The vertical y-axis shows the frequency for each degree, or the number of times each degree score appeared in the network. The horizontal x-axis shows the range of degrees in this network. There are 0 nodes with the degree of 0 or 6, so there are no orange bars for these degree values in the histogram. We can see that the degree of 2 and 4 were the most frequent in this example network as 3 nodes have a degree of 2 and 3 nodes have a degree of 4. Only 1 node has a degree of 7, and this is the maximum degree for this network. This node is the most central in the network as it has the highest degree centrality. If we wanted to know where the action was in the network or who the most important node was in the network, we could use this degree distribution to identify the node with the highest degree.



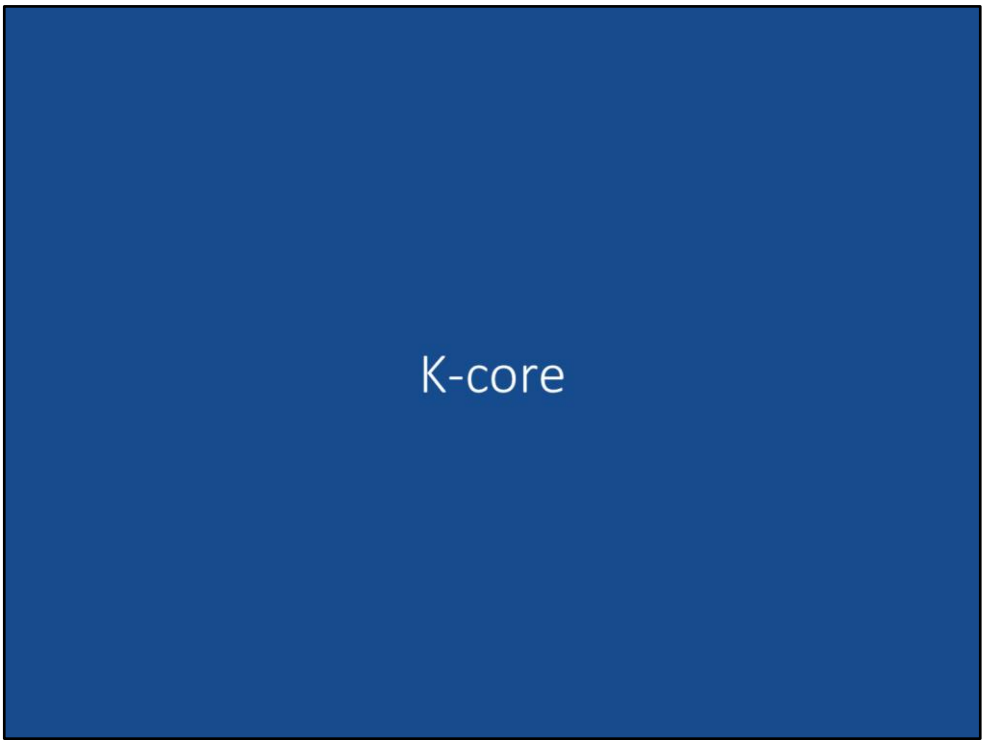
Another way to look at important or key players in the network is to see what the network looks like without the key player. Here is the same example network from the previous slide minus the node with the highest degree. Here we can visualize the impact and importance that the node with the highest degree had on the network. What major changes do you notice in this network? Notice that the network now has two components. The network is more symmetric looking or more balanced looking. Most of the nodes remaining in this network have only 2 or 3 ties.



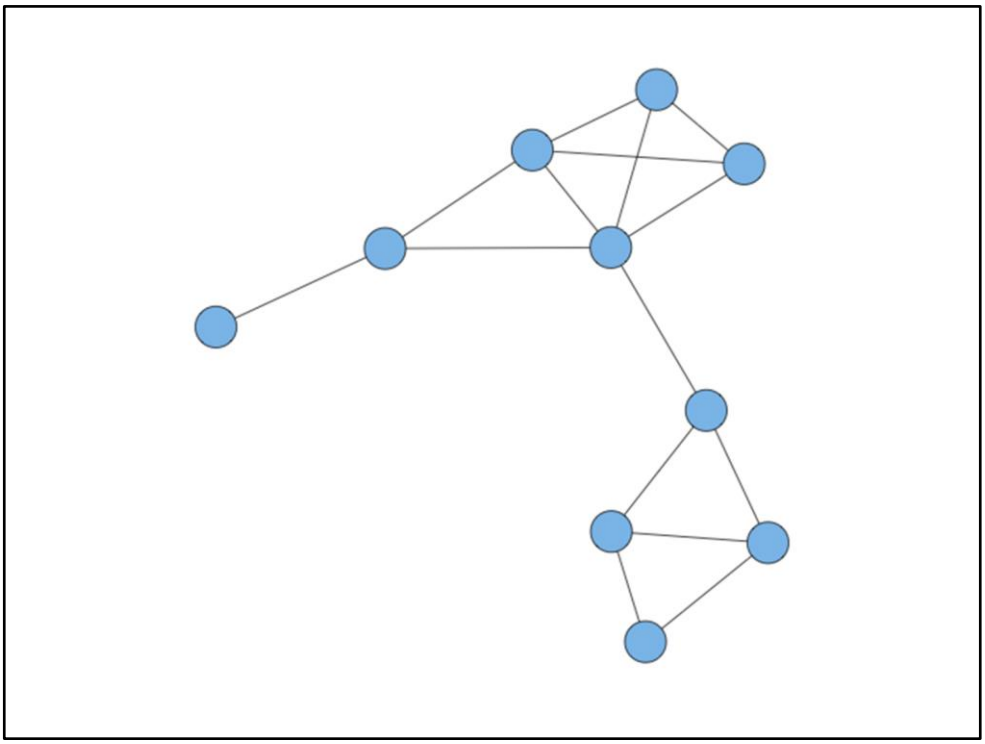
Here is a network of a gang from Chicago, the Guttaville Gangster Disciples. The first thing we might notice about this network is that members of the Guttaville GDs are not all connected. In fact, there are 9 components in this network, 7 of which are isolates. These isolates are cases of individual arrests whom the police identified as members of the gang, but these individuals were not involved in co-arrests. We also might notice that there are several nodes in this network with a high proportion of ties even though the majority of the nodes have only a few ties. So where is the action in this network? Who are the most important people in this network?



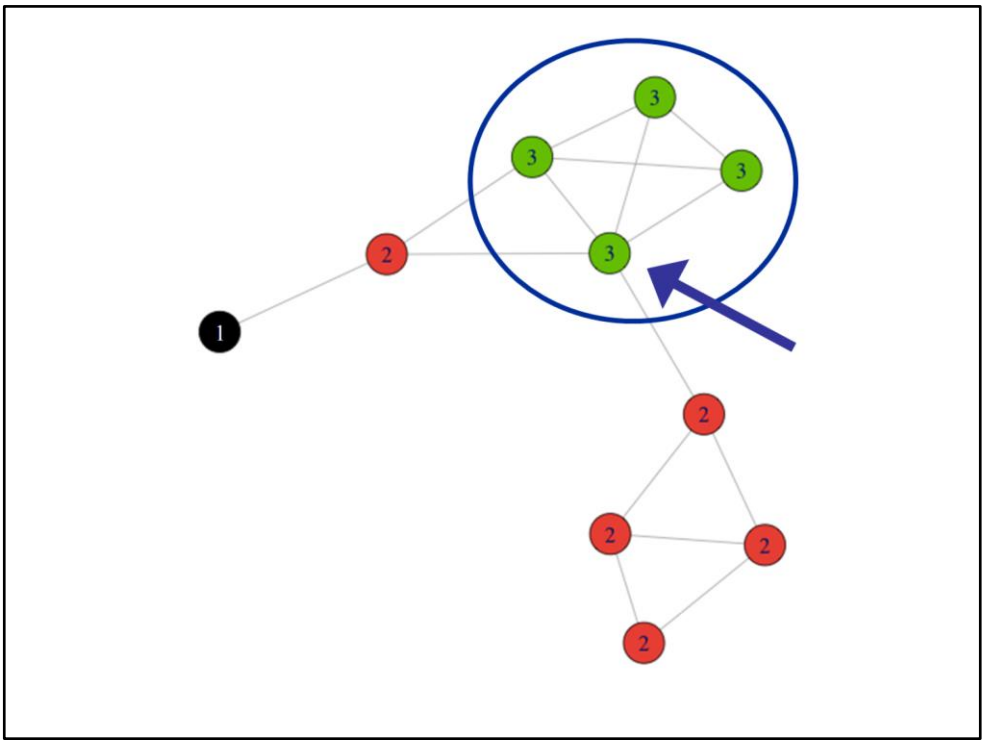
Here is the same gang network of the Guttaville GDs, but in this visualization the size of the nodes is in proportion to each node's degree. Now the isolates are the smallest nodes in the network. Nodes with a degree of 1 are also quite small. The largest circles make it easy to see which nodes have the highest degree scores in the network. Now we can easily identify which nodes are the most important or central in this network, and we could zoom in and analyze these individuals and their connections more closely.



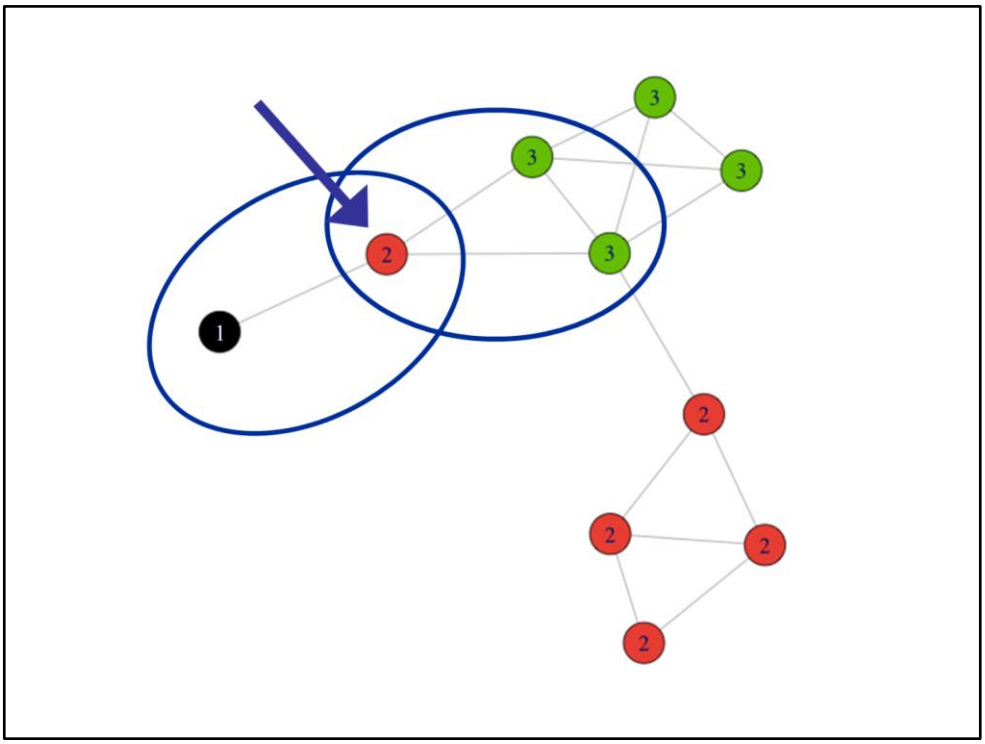
Where are the dense pockets of groups within the network?



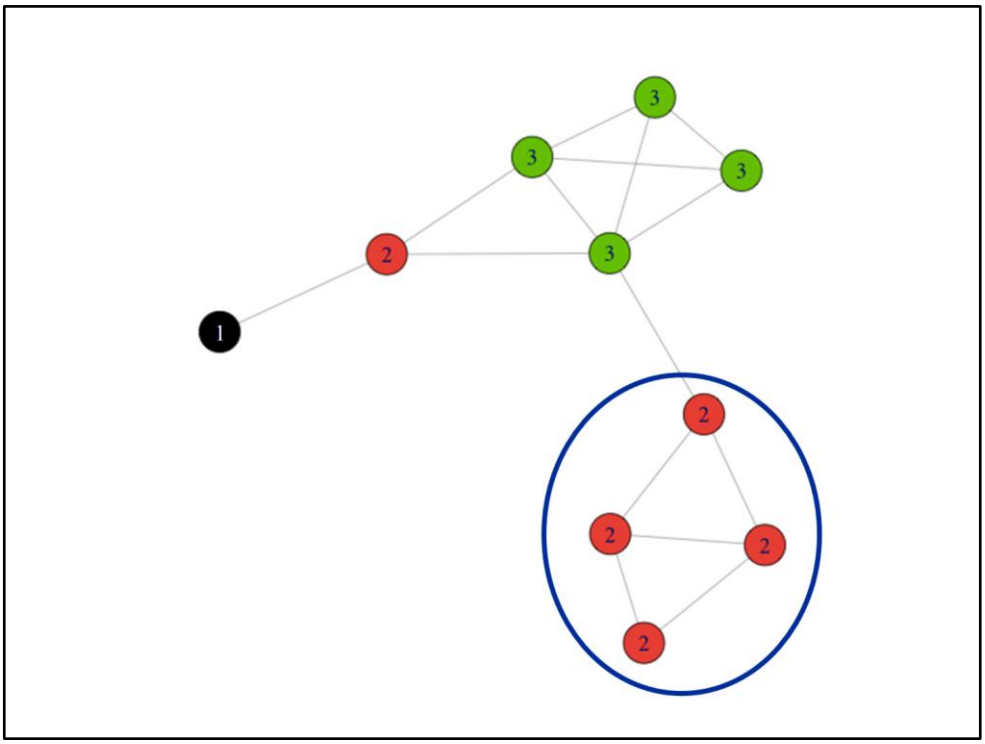
K-core uses degree to identify dense pockets of cohesion in a network. K-core measures areas of the network that have the same minimum degree score, such as small areas where all the nodes have at least 3 edges often connecting to each other. K-core will always be equal to or smaller than a node's degree. K-core can never be greater than degree. Nodes with the highest degree will not have the highest k-core if there aren't other nodes with similar degrees in their area of the network. Often k-core identifies the largest completely connected groups, but this is not always the case. Where are the dense pockets within this example network? Which nodes have the largest k-core?



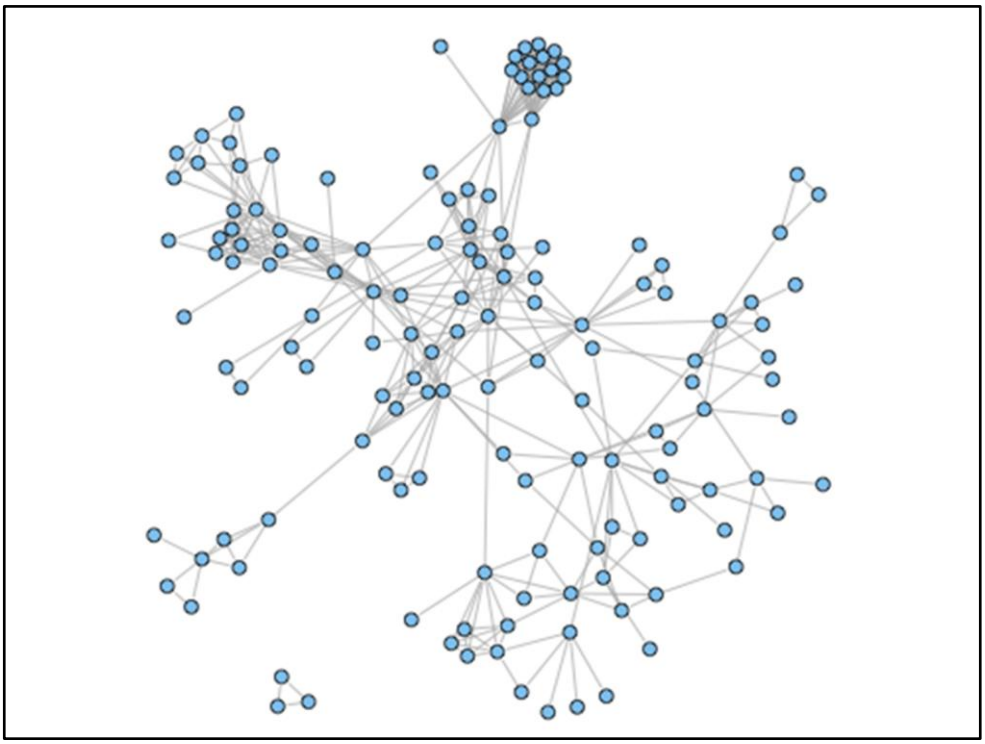
This image shows each node's k-core value, which is based on the minimum required degree within a connected area of a network. The green nodes all have a k-core of 3 because this area of the network is defined by nodes having at least a degree of 3. These 4 nodes also form a completely connected group. Notice that the green node pointed to by the blue arrow actually has a degree of 5, but 3 is the minimum required degree to be in the area of the network identified by the blue circle. Three is also the maximum k-core for this entire network, so this set of green nodes is the largest and most dense pocket of the network.



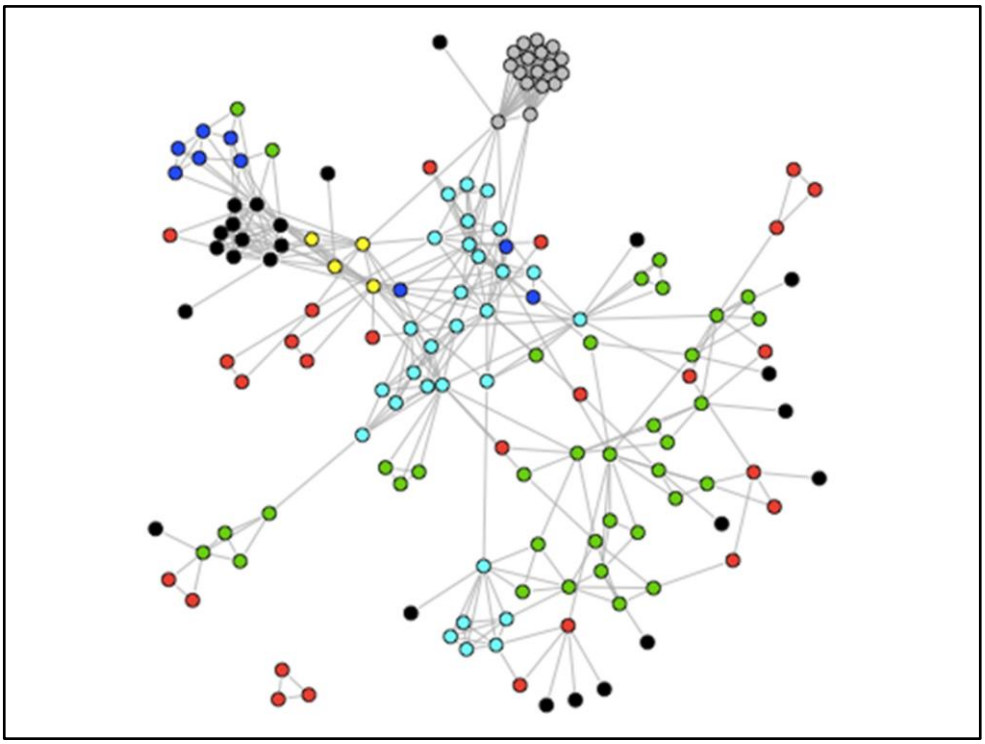
The black node has a k-core of 1 because of its degree of 1. The other node in the black node's subgroup is the red node with a number 2 pointed to by the blue arrow. This red node has a k-core of 2 because it also resides in an area of the network in which all the nodes have at least a degree of 2.



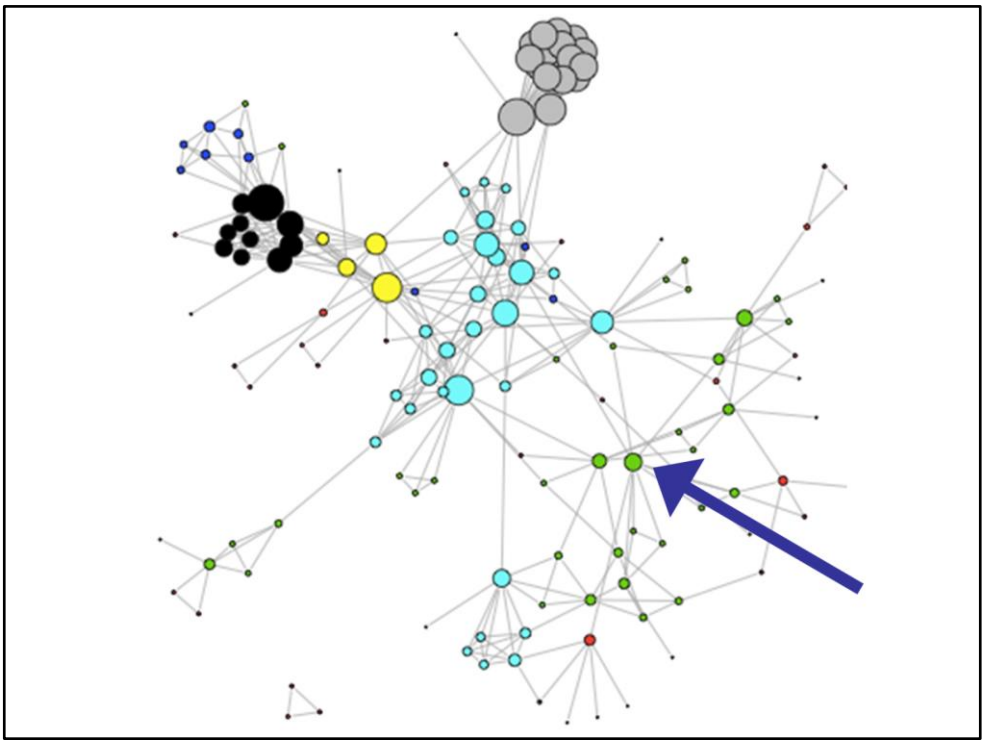
There are 4 nodes on the bottom of this network that only have a k-core of 2. Each node in the area of the network enclosed by the blue circle has a degree of at least 2.



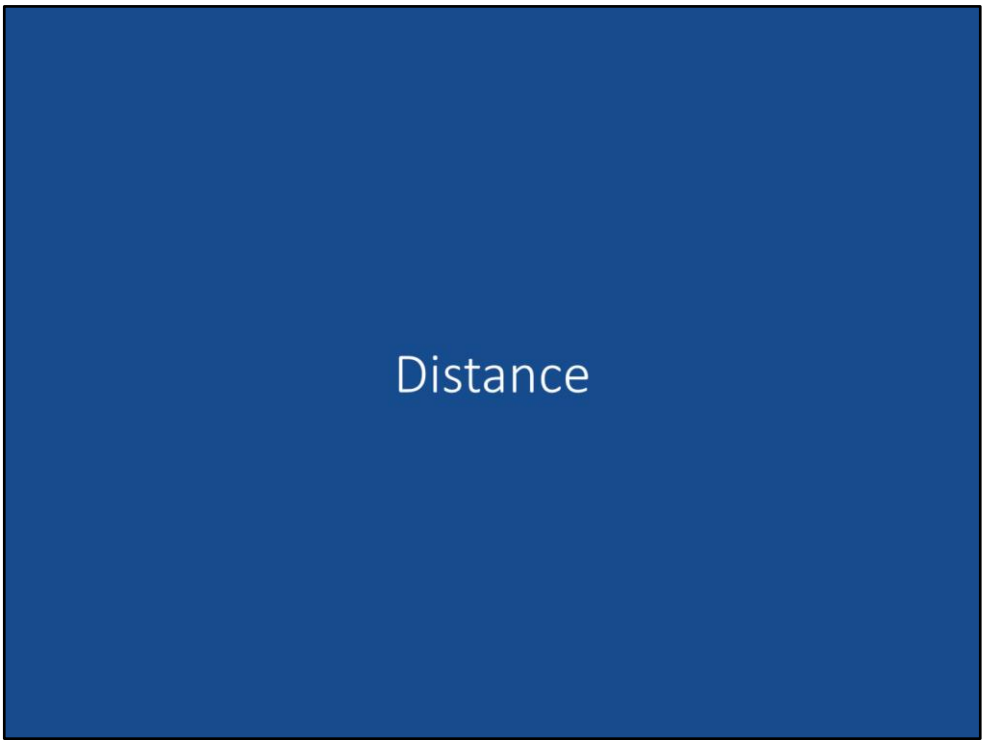
This network is of the 68th and Hamlin Latin Kings gang of Chicago. Where are the dense pockets of connected groups in this network? Where are the locations of the greatest cohesion?



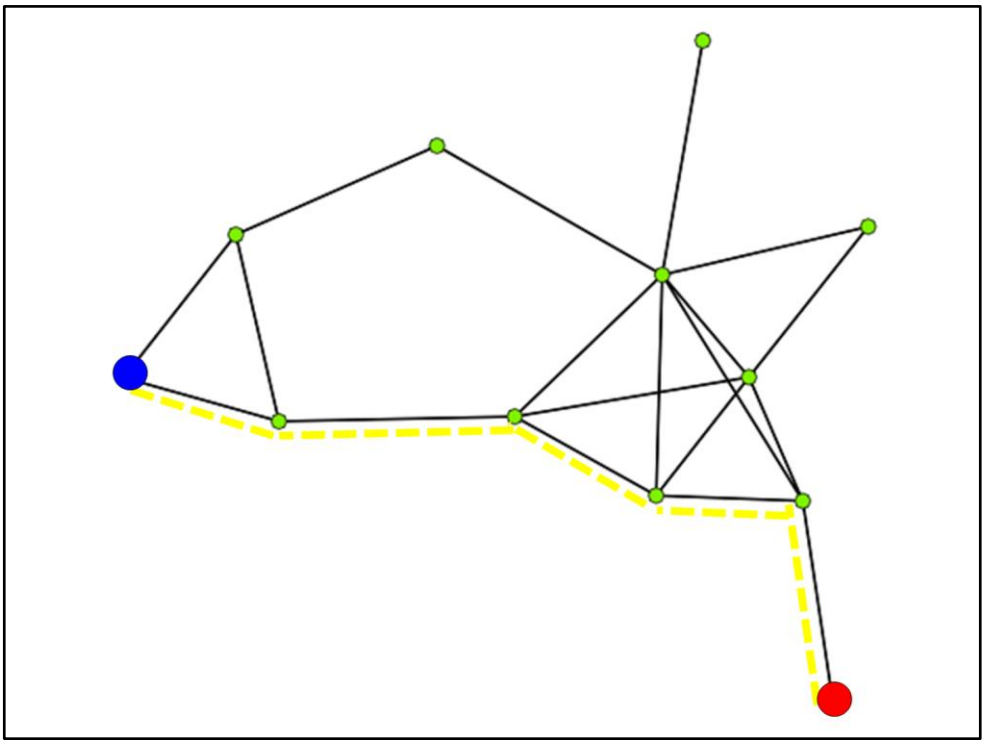
This visualization of the same network makes the questions about cohesion and k-core easy to see because the k-core sizes are color coded. Red nodes have a k-core of 2, green nodes have a k-core of 3, dark blue nodes have a k-core of 4, sky blue nodes have a k-core of 5, etc. Hopefully, you realized on the previous slide that the most dense cluster in this network is the gray cluster located at the top that connects 16 individuals all having a degree of at least 15. The palette of colors used in this network repeats itself, which means that some of the colors get used more than once. Black nodes have a k-core of 1, but, as you can see in the upper left of the network, black nodes designate a higher k-core value there.



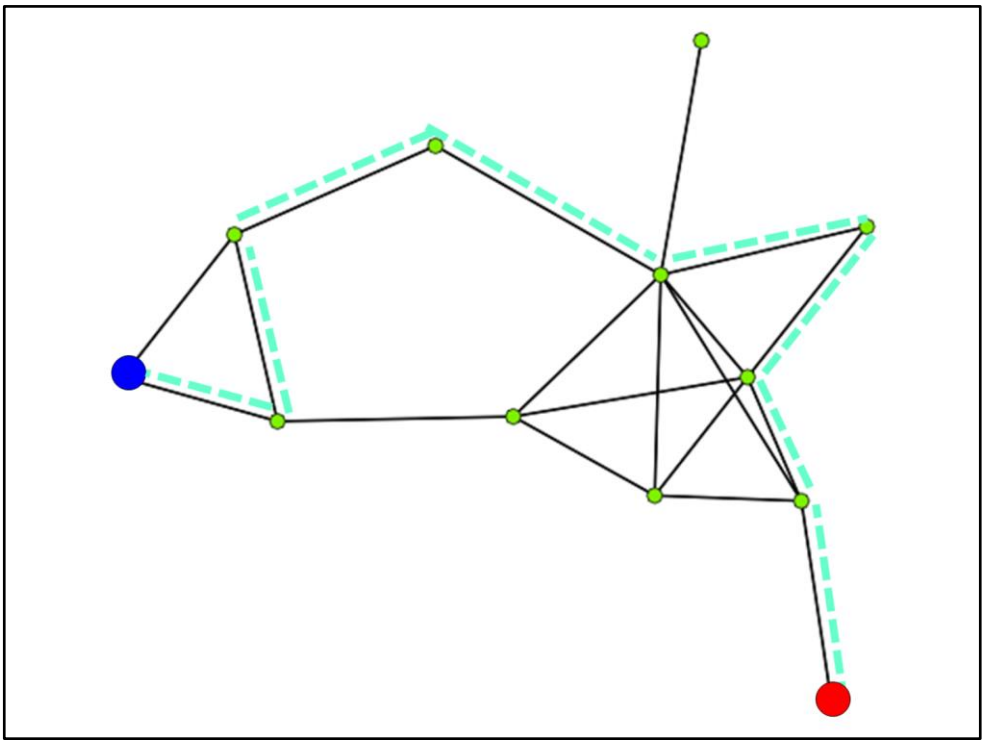
This version of Chicago's 68th and Hamlin Latin Kings gang sets the size of the nodes to be proportional to the degree of the nodes, and the colors are set to the k-core based on the previous slide. This visualization helps demonstrate the relationship between k-core and degree. Nodes with a small degree by definition have a small k-core. They don't have enough ties for large k-cores. Nodes with high degree potentially can be in areas of a network with other high degree nodes, and, thus, potentially have large k-cores. However, if none of the members of that area of the network connect, then the nodes have fewer ties and smaller k-cores. For example, the dark blue arrow points to a green node with a degree of 9. The other nodes in this area of the network do not have large degrees like this node's degree of 9. This high degree node only has a k-core of 3 because the nodes in this area of the network have only at least a degree of 3.



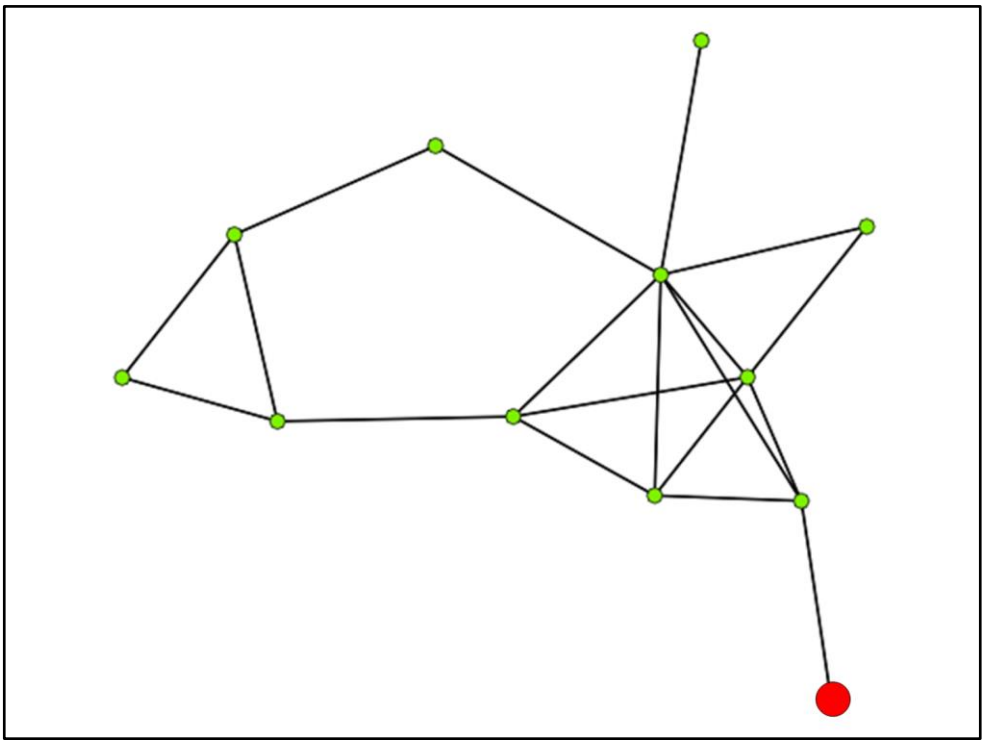
How far apart are individuals within the network? How far would something have to spread or travel to reach all the members of a network? Where are the paths through the network? How many handshakes away is someone?



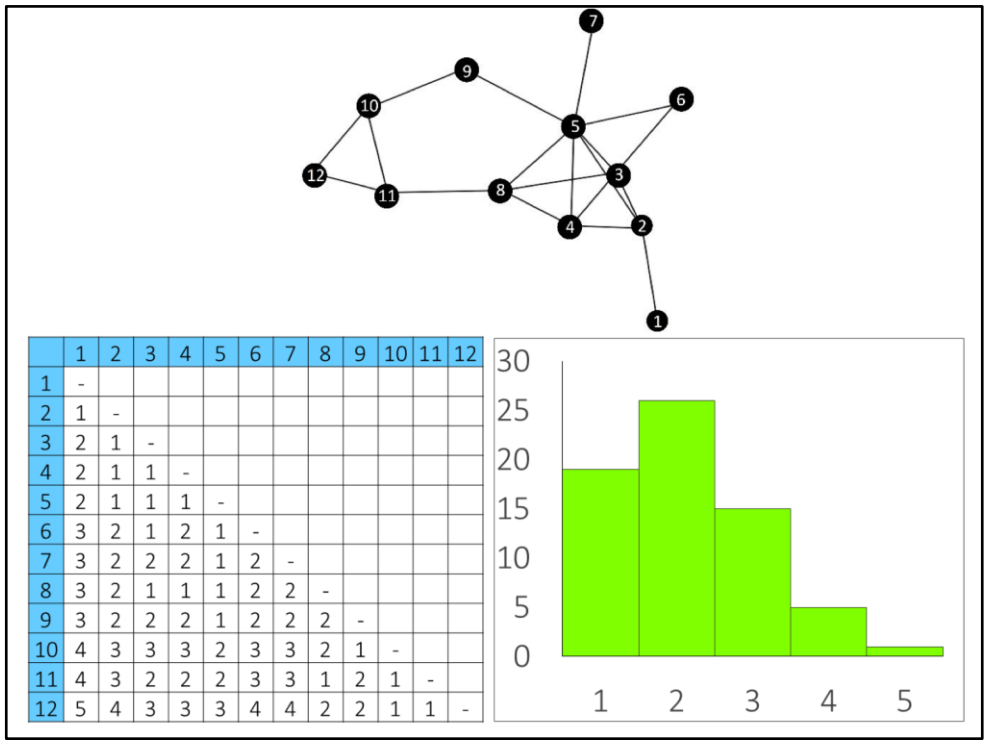
Imagine that something needed to spread across this network--maybe information, news, or money. There are multiple paths between two nodes in a network, but you would want to choose the most efficient and direct path. Let's say that the red node in the example above wants to pass a note along to the blue node. The path for passing this note from the red node to the blue node is constrained by the edges in the network, so the red node will want to use the most efficient and shortest path to send the note. The number of edges or steps along this path is called the geodesic distance. Geodesic distance is the shortest path of all possible paths between two nodes. The geodesic distance between the red node and the blue node is 5 because there are 5 edges between these two nodes along the shortest path. It is worth noting that there are several paths between the red and the blue node that equal 5 steps. One of the possibilities is marked by the yellow dotted line. Can you find some of the other geodesic paths between these two nodes of 5 steps?



This aqua colored line certainly indicates one possible path through which the red node could send a note to the blue node, but it is incredibly inefficient. This path is not the geodesic distance. It requires 8 steps to connect the two nodes compared to the yellow path of 5 steps in the previous slide.

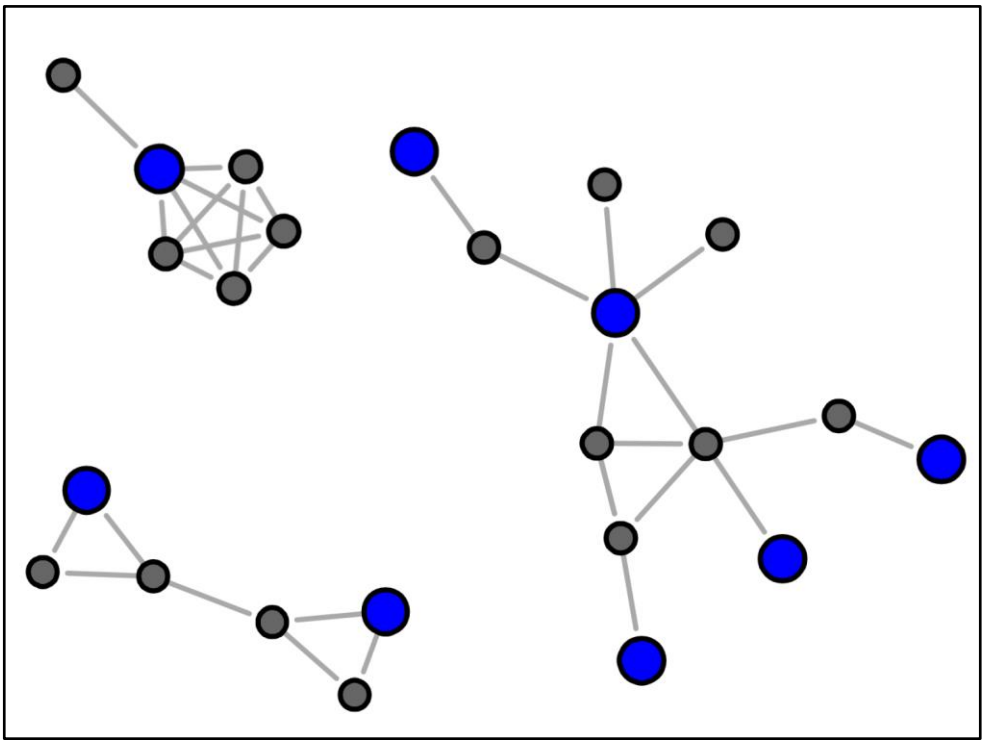


In a network of 12 nodes, like the one above, there are 66 geodesic distances. Nodes don't have a geodesic distance to themselves, but they have a geodesic distance to every other node in the component. Each geodesic distance measures the number of steps between any two nodes in the component. The red node has 11 geodesic distances, one to each of the green nodes in the component. Each green node also has 11 geodesic distances to every other node in the component. The longest geodesic distance is the diameter of a network.

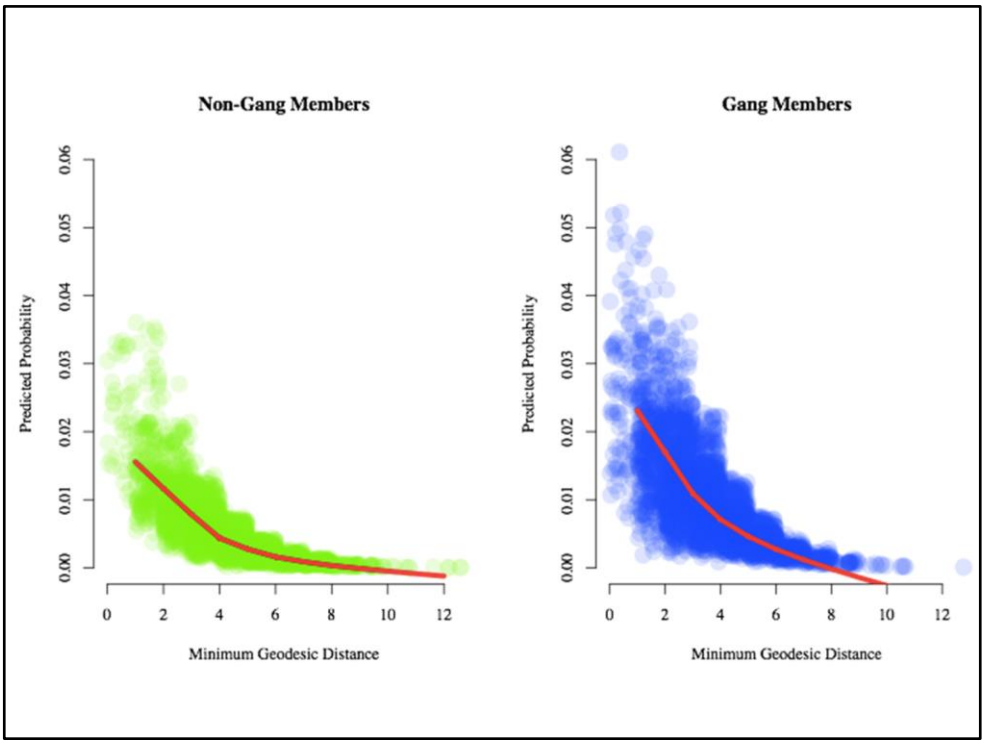


	1	2	3	4	5	6	7	8	9	10	11	12
1	-											
2	1	-										
3	2	1	-									
4	2	1	1	-								
5	2	1	1	1	-							
6	3	2	1	2	1	-						
7	3	2	2	2	1	2	-					
8	3	2	1	1	1	2	2	-				
9	3	2	2	2	1	2	2	2	-			
10	4	3	3	3	2	3	3	2	1	-		
11	4	3	2	2	2	3	3	1	2	1	-	
12	5	4	3	3	3	4	4	2	2	1	1	-

The matrix on the bottom left of this slide shows the geodesic distances between all 12 nodes. The diagonal is not defined because a node cannot have a shortest path to itself. The upper triangle of the matrix is blank because this is an undirected network. The top triangle just mirrors the bottom triangle, so we only need one triangle to list all the geodesic distances. According to this matrix of geodesic distances, what is the diameter of our example network? The green histogram in the lower right of this slide shows the distribution of all 66 geodesic distances. Here we see that the diameter of the hypothetical network is 5. We also see that the vast majority of the nodes in this network are only 1 or 2 steps away from each other. Overall, information or items do not have to spread very far to reach everyone in the network.

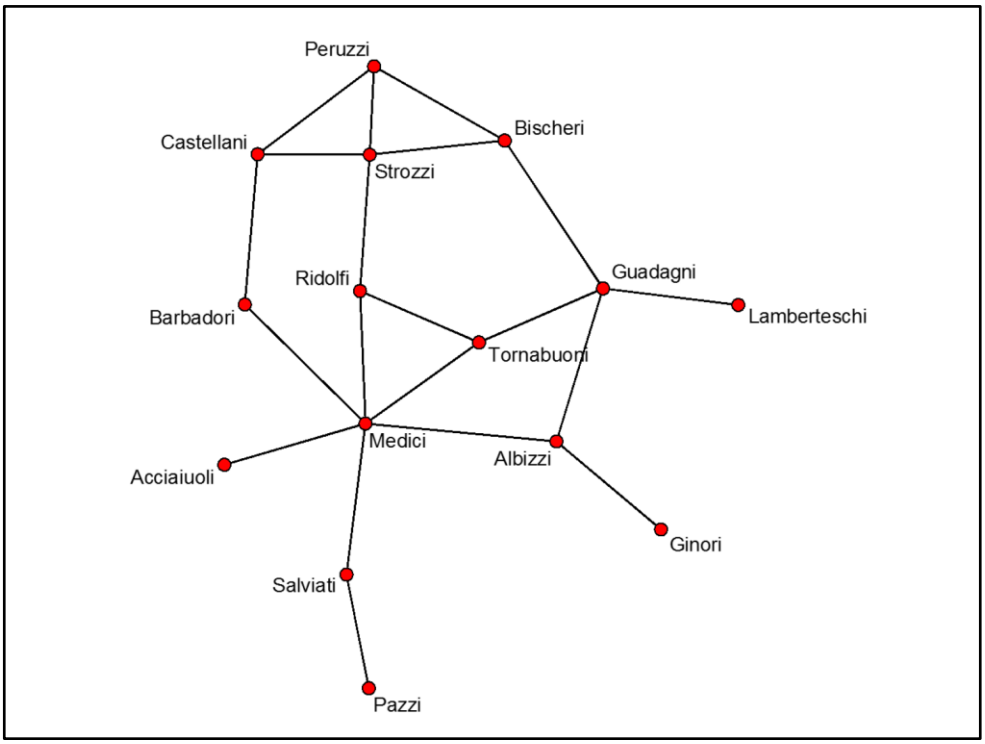


Distance can only be calculated within a single component. Distances between components are not defined. The network in this slide has three components. We can calculate the geodesic distance for each node to every other node within the same component, but we cannot calculate the geodesic distance of a node from one component to any nodes in any other components.

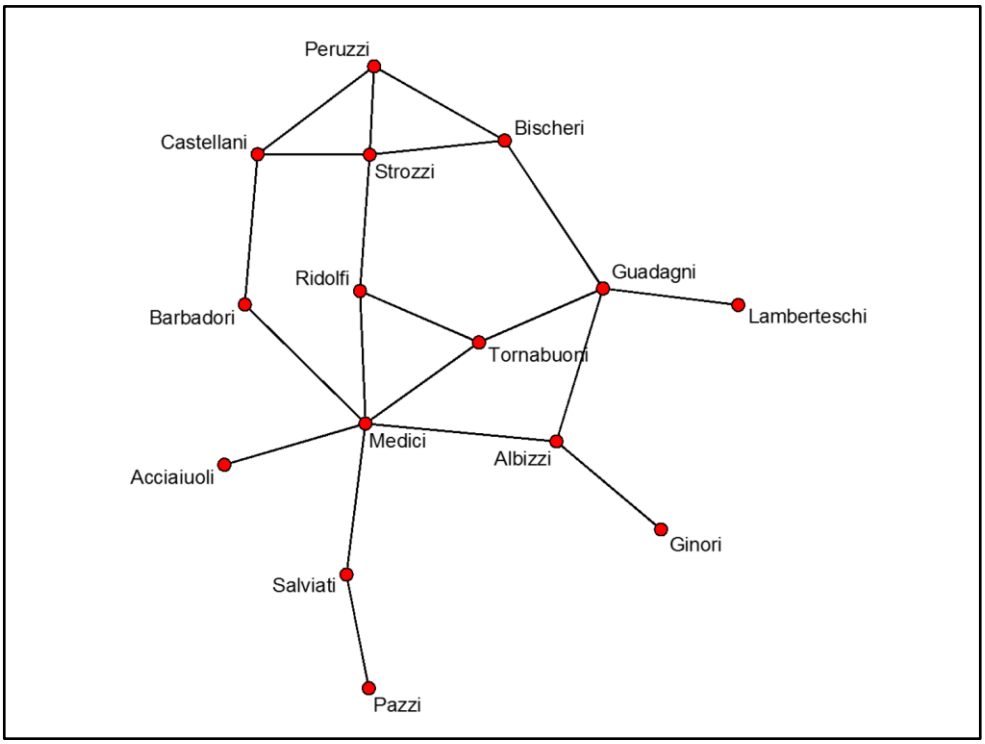


One way to conceptualize geodesic distance is to think of it as a social distance rather than a spatial distance. How many handshakes away are individuals from some sort of phenomenon or outbreak in the network? Research has looked at individuals in large urban arrest networks and measured each person's social distance to homicide victims. The finding is a social contagion effect--the closer one is to a homicide victim the more likely that person is to also be a homicide victim. In this case, longer social distances measured as geodesic distances are a good thing. Each handshake away from a homicide victim in an arrest network reduces victimization by 57 percent. The two figures in the slide above show this pattern. The predicted probabilities in the vertical y-axis measure the likelihood of homicide victimization, and the horizontal x-axis measures the geodesic distance to a homicide victim. The larger the geodesic distance, the lower the probability for victimization. These patterns are similar for both gang members and non-gang members in the arrest network.

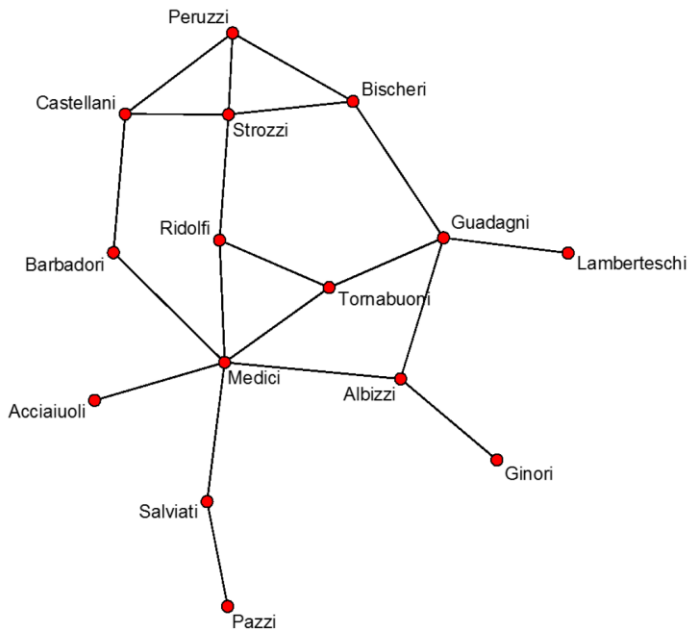
Who is in the middle of the action? Who is brokering the network? Who controls the spread of something in the network?



Now that we are familiar with geodesic distance, we can introduce brokerage. Do you remember all of those shortest paths that we calculated in the previous section? Along many of those geodesic distances resided other nodes, and those between nodes are important in maintaining the shortest paths. Nodes that fall on the path between other nodes are important. Imagine if a resource was traveling through the largest component. Certain individuals reside on many more paths than others, and so the resource has to go through them. Those certain individuals are important to making sure that the resource gets passed on to everyone else. Being on many paths becomes a rather important position within a network. Nodes located on many shortest paths can control the flow of resources, information, or ideas to the rest of the group. The number of geodesic distances that a node resides on is its betweenness centrality. A high betweenness centrality means that more geodesic distances have to go through that node, and that node acts as a broker in the network. Brokers are in the middle of the network's action and resources. A low betweenness centrality means that few geodesic distances go through that node, and that node is not a big broker in the network. It is possible to have a betweenness centrality of zero, these are the nodes that broker no ties in the network.

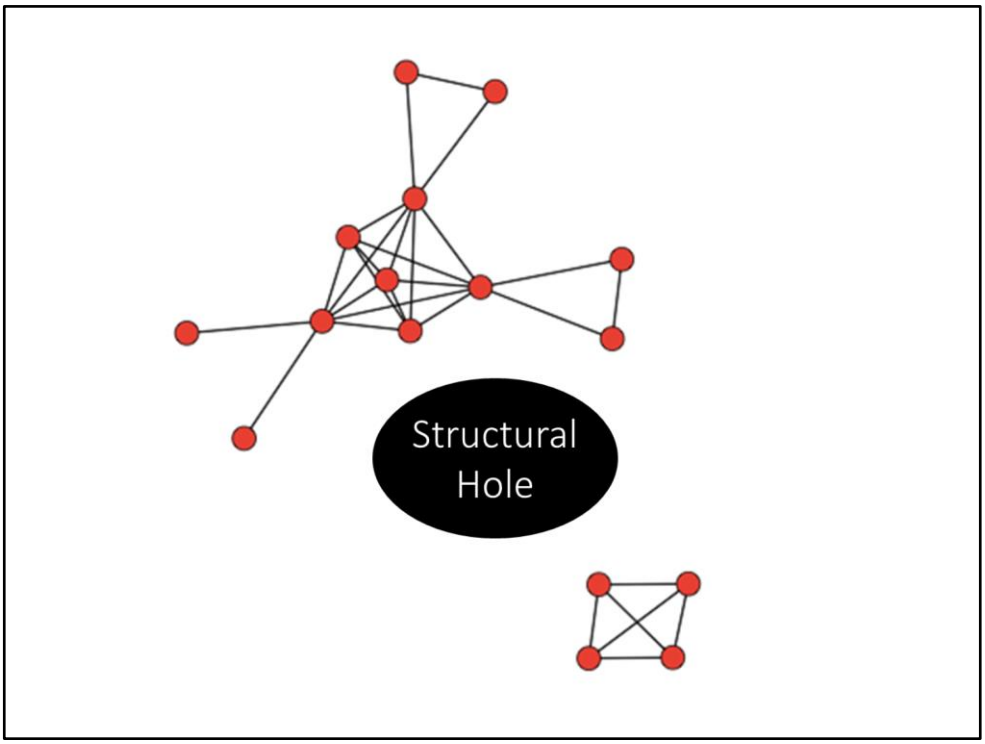


The data for this network are publicly available in one of the social network packages in R. The dataset is called “flomarriage.” Each node is an elite family from Renaissance Florence, and each edge is a marriage relationship between families. Who do you think is the most central broker in this Florence marriage network? Take a guess at what the highest betweenness centrality score might be for this network. Which nodes are never brokers? How many nodes have a betweenness centrality score of zero? You can check your answers on the next slide.

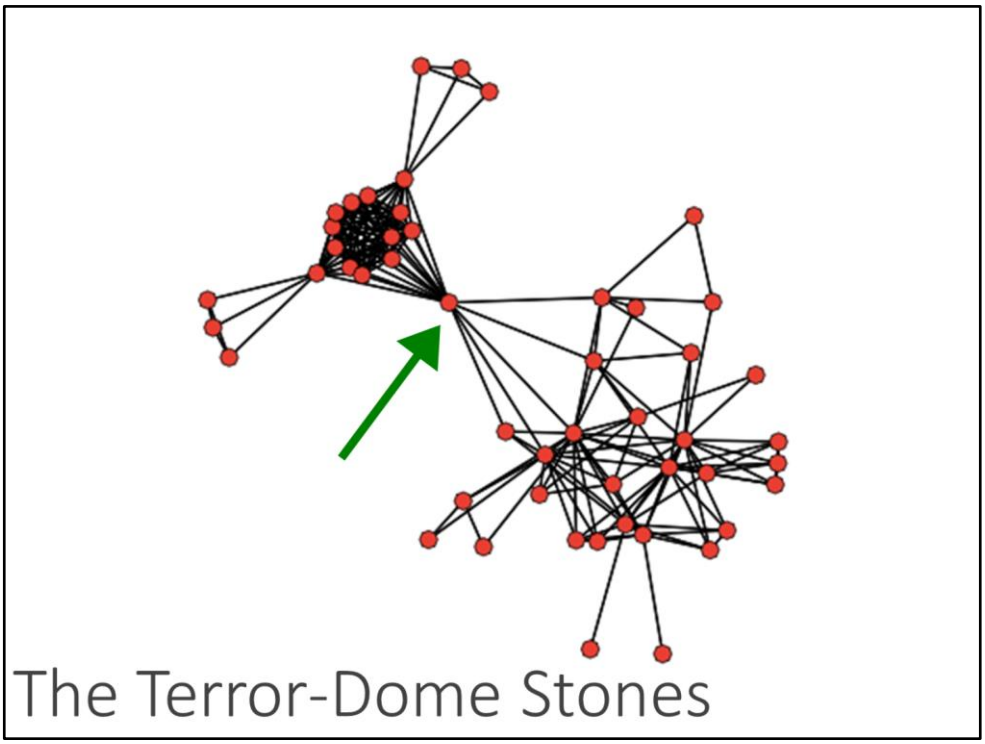


Betweenness Centrality	
Acciaiuoli	0
Albizzi	19.33
Barbadori	8.5
Bischeri	9.5
Castellani	5
Ginori	0
Guadagni	23.17
Lamberteschi	0
Medici	47.5
Pazzi	0
Peruzzi	2
Ridolfi	10.33
Salviati	13
Strozzi	9.33
Tornabuoni	8.33

The node with the highest betweenness centrality is Medici. As we can see in the table, the Medici family resides on 47.5 geodesic distances. Research on brokerage has called this position in the network “robust action” (Padgett and Ansell 1993). Notice that not all the betweenness centrality scores are whole numbers. This is because sometimes there are draws or ties for brokerage, such as cases where there are two different paths but both are the shortest path. In the cases of draws, the betweenness counts get divided between the nodes. Did you correctly identify the four nodes that are never brokers in this network, who have a betweenness centrality score of 0?

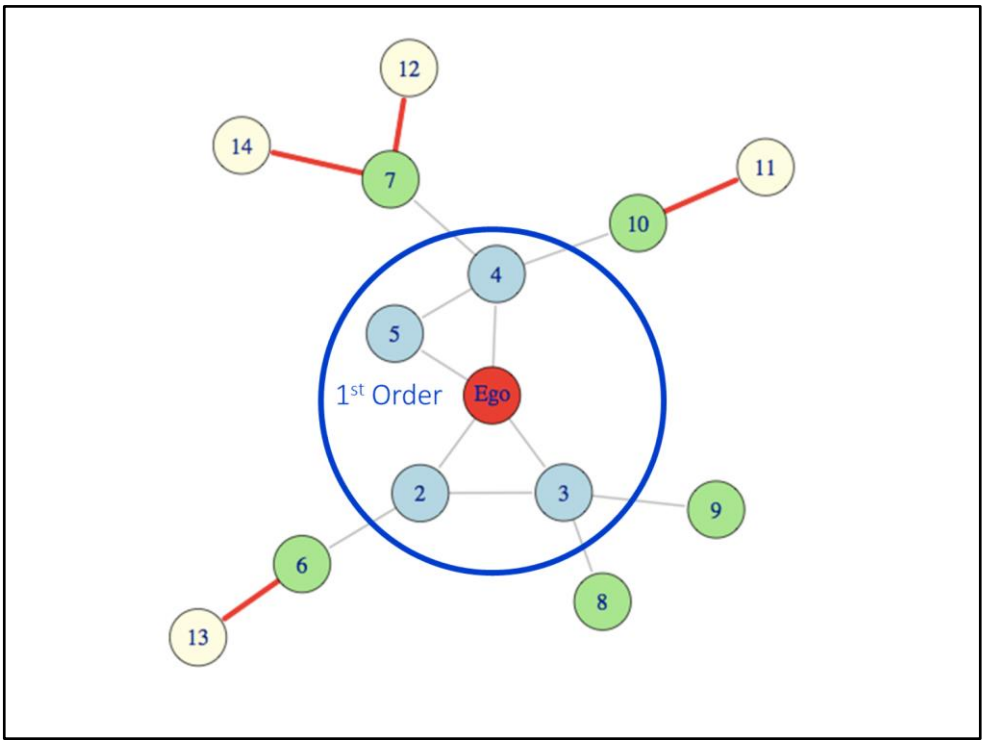


Structural holes are gaps in the network or the space between the network's components. Filling or brokering a structural hole is a powerful network position because it is a single position connecting all the nodes from one component to the nodes in the other component. A node who fills a structural hole is an important broker.

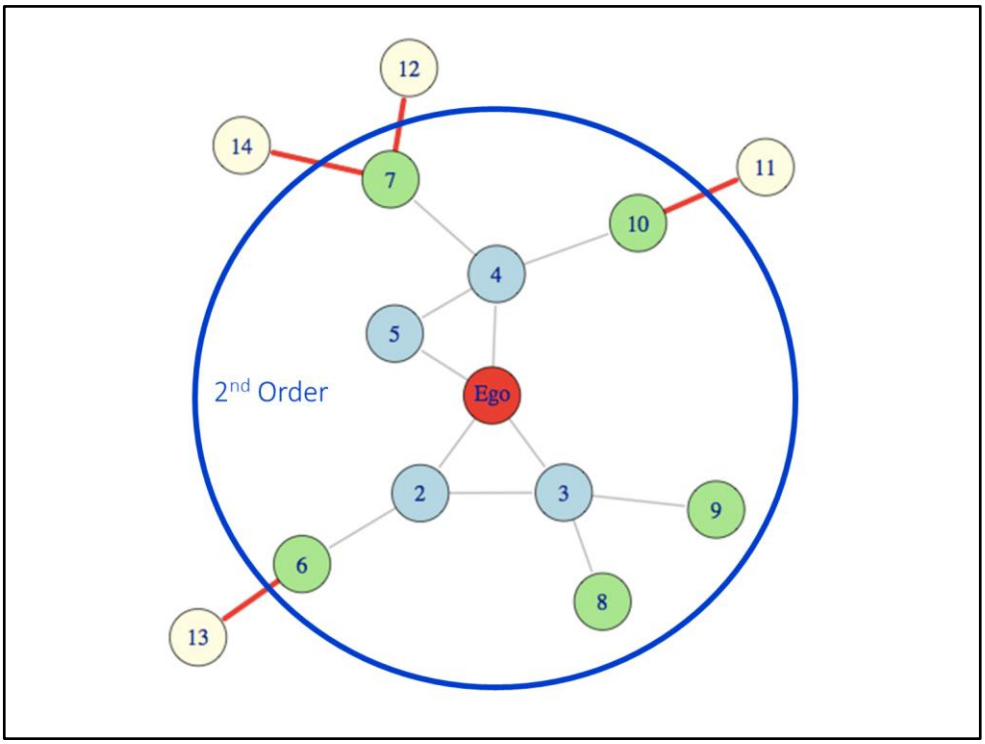


This Chicago gang, the Terror-Dome Stones, clearly has one broker filling what would be a structural hole between the component on the top left and the component on the bottom right. Identifying this broker could be important to understanding the structure of this gang. Visually it is clear who the broker is because of the structural hole that would be there without the broker. This broker would also have the highest betweenness centrality score because every single geodesic distance between the nodes in the upper left section to the nodes in the bottom right section would pass through this broker.

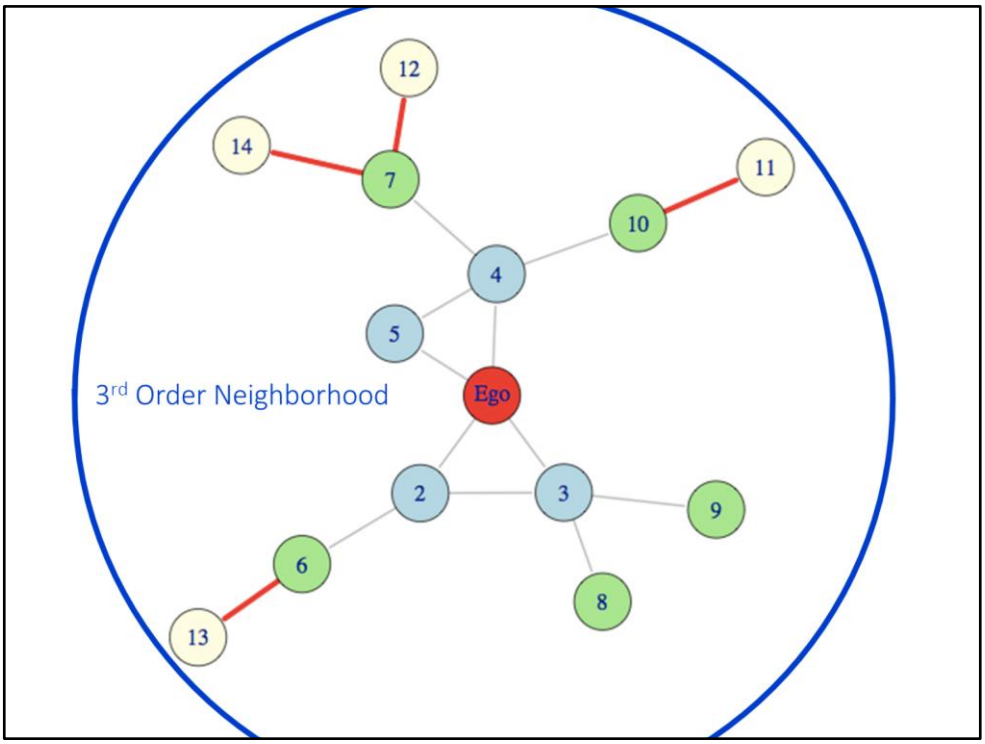
Who are the people in your neighborhood? Who are your neighbors' neighbors?



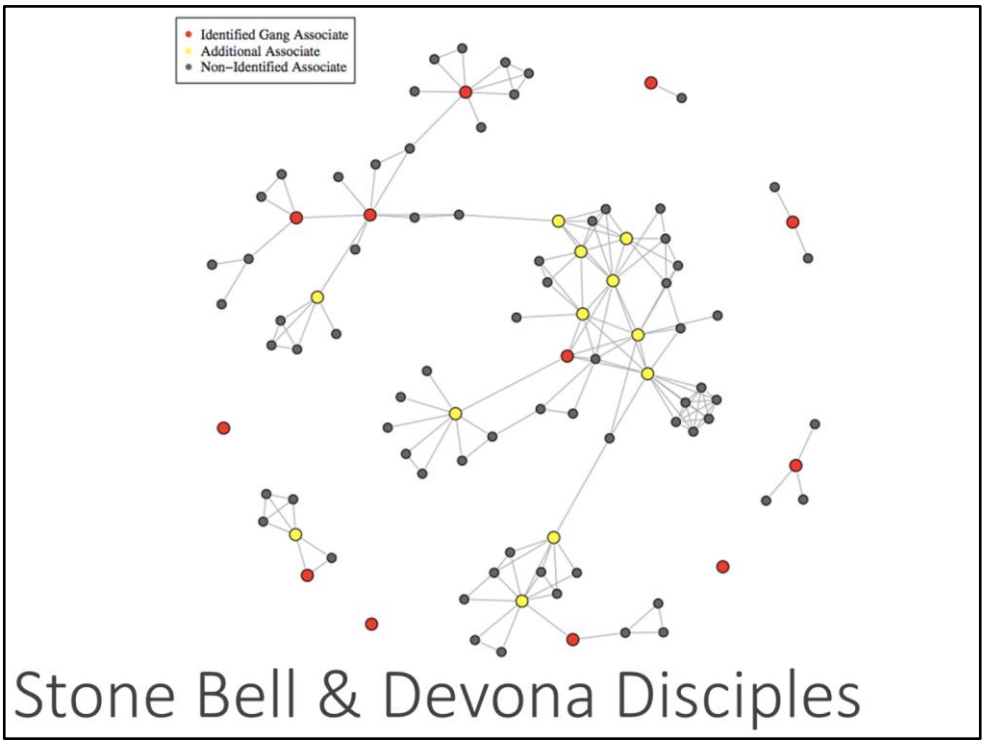
A neighborhood refers to the alters within a specific distance (e.g., 1, 2, 3, etc.) from a focal node or a node of interest. Focal nodes or nodes of interest are also called egos. A 1st order neighborhood refers to the ego plus all of the nodes directly connected to the ego. A 2nd order neighborhood includes all of the nodes in the 1st order neighborhood plus all of the nodes that are two steps away from the ego. Neighborhoods are related to the concept of social distance. The larger the order of the neighborhood the farther away the nodes are from the node of interest. Tobler's Law states that everything is related to everything else, but near things are more related than distant things. The logic here is that the closer the neighborhood, the more influence the ego has on those nodes, and the more influence those nodes have on the ego. The farther the neighborhood, the less influence. The blue circle contains all of the nodes in the 1st order neighborhood for the red ego.



Now the blue circle contains all nodes within the 2nd order neighborhood of the red ego. Neighborhoods include all of the nodes within the smaller order neighborhoods, so all of the blue nodes that were a part of the 1st order neighborhood are counted in this 2nd order neighborhood.



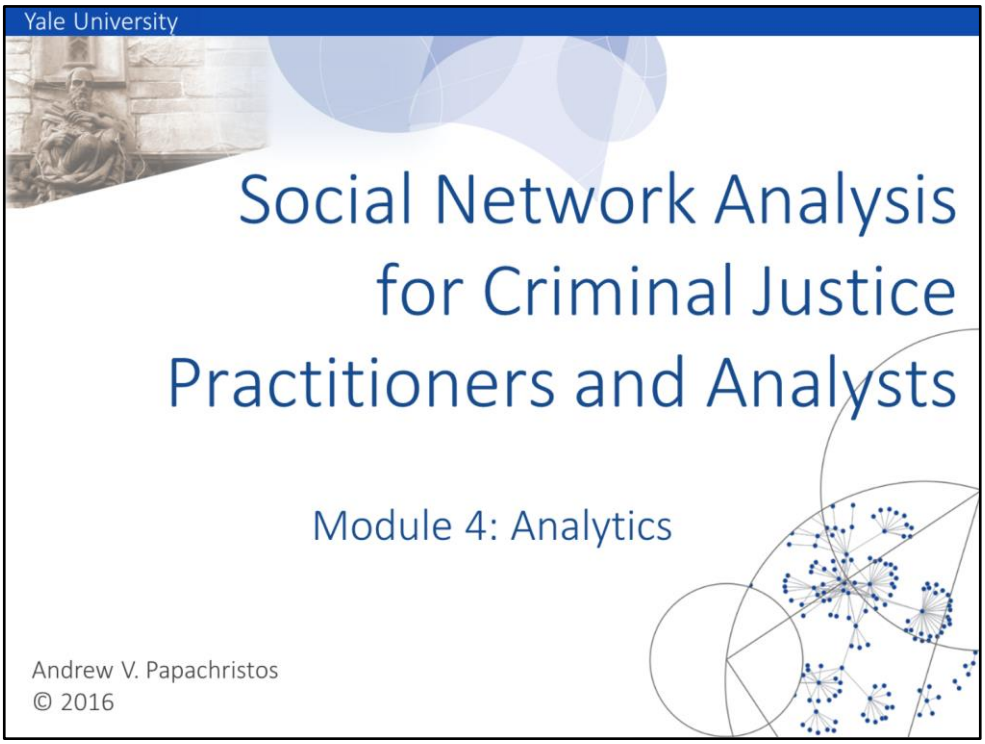
This final blue circle contains the 3rd order neighborhood, which in this example is the entire ego network.



Analyses of neighborhoods is especially useful if you have a list of names or individuals of interest and you want to see what their networks look like out in the 1st or 2nd order neighborhoods. Perhaps you had a list of all gang members who had been shot. You could assign that list of names as your egos, and then you could add in the 1st order neighborhood or 2nd order neighborhood. This method identifies smaller local structures that are sometimes more useful than a large total arrest network. Within the 1st order and 2nd order neighborhoods, you could then identify other gang members and non-gang associates.

REFERENCES

Padgett, John F. and Christopher K. Ansell. 1993. "Robust Action and the Rise of the Medici, 1400-1434." *American Journal of Sociology* 98(6):1259-319.



Module 4: Analytics covered several measures of social networks through illustrations and examples: density, components, degree, k-core, distance, brokerage, and neighborhood. Participants can review these analytics in the Module 4 labs.