Graph Theory and Social Networks
Spring 2014 Notes

Kimball Martin

March 14, 2014

Introduction

Graph theory is a branch of discrete mathematics (more specifically, combinatorics) whose origin is generally attributed to Leonard Euler's solution of the Königsberg bridge problem in 1736. At the time, there were two islands in the river Pregel, and 7 bridges connecting the islands to each other and to each bank of the river. As legend goes, for leisure, people would try to find a path in the city of Königsberg which traversed each of the 7 bridges exactly once (see Figure 1). Euler represented this abstractly as a *graph*[*], and showed by elementary means that no such path exists.

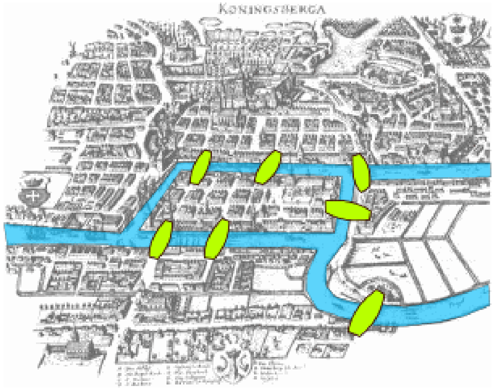


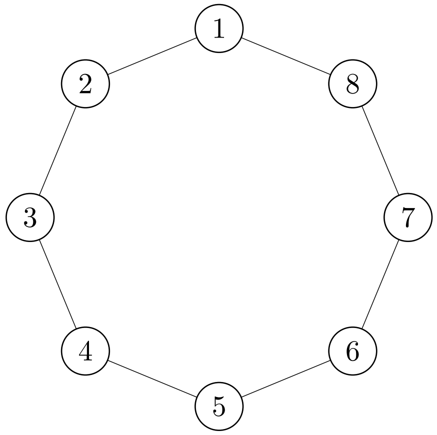
Figure 1: The Seven Bridges of Königsberg (Source: Wikimedia Commons)

Intuitively, a graph is just a set of objects which are connected in some way. The objects are called *vertices* or *nodes*. Pictorially, we usually draw the vertices as circles, and draw a line between two vertices if they are connected or related (in whatever context we have in mind). These lines are called *edges* or *links*.

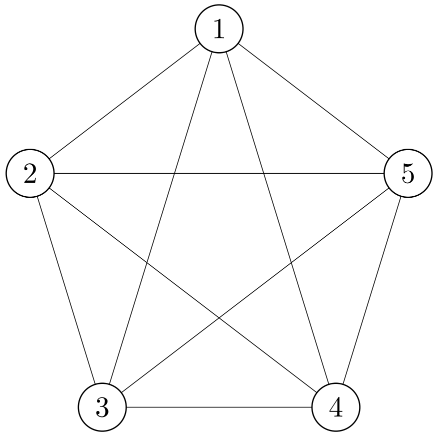
Here are a few examples of abstract graphs.

This is a graph with 8 vertices connected in a circle.

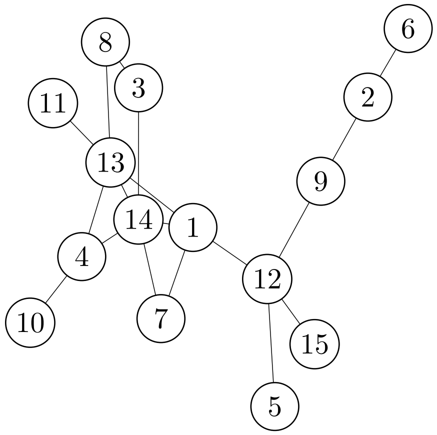
[*] In this course, graph does not mean the graph of a function, as in calculus. It is unfortunate, but these two very basic objects in mathematics have the same name.



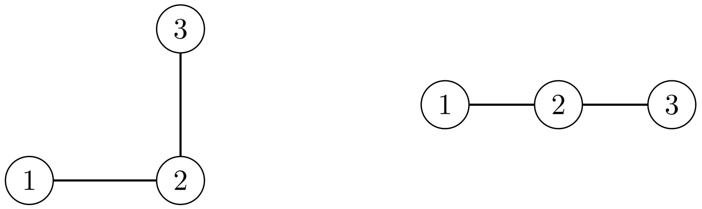
This is a graph on 5 vertices, where all pairs of vertices are connected.



Here is a "random graph" on 15 vertices generated by a computer—in this case, each pair of vertices had a 15% chance of being connected.



Note that the physical location of the vertices in the drawings are unimportant, only which vertices are connected matters. For example, the following two graphs are the same.



Graphs naturally arise in many ways, as they are a convenient way to visualize various situations or complex symptoms. In fact graphs are nearly ubiquitous in mathematics as they can be used to represent different aspects of all kinds of mathematical structures. Consequently, graphs are also prevalent in many scientific fields, where they are often called networks. Technically the two terms are interchangeable, but one typically uses the term network when one is thinking about social or technological connections.

Here are a few examples of situations where graphs arise.

Geometry/Topology: Polygons are vertices and edges in the plane, and can be thought of as graphs. Note that the graph is not the same as the underlying polygon—things like edge length and vertex angles are not considered in the graph. So in some sense, the graphs of a polygon is the polygon without its geometry. This is perhaps not so interesting for polygons, but one can do the same for polyhedra (e.g., the Platonic solids) where looking at the analogous graphs is very useful in the study of surfaces and higher-dimensional objects.

Algebra: Algebra is the study of mathematical structures, e.g., groups, rings and fields, if you know what those are. Graphs are often used to relate how various structures are related to each other (e.g., subgroup lattices), or to understand individual objects, e.g., group actions on graphs). The notion of group actions on graphs has many applications to constructions of *combinatorial designs*, which are important in error-correcting codes and cryptography.

Electric engineering: Electric circuits are graphs, and graph theory has been used in circuit analysis since Kirchoff in the mid 1850's.

Chemistry: Molecular graphs are representations of molecules as graphs—here are the atoms in the molecule and the edges are the bonds. This point of view was first taken up by Cayley in the 1870's.

Geography: Consider a map, say of Europe. Let each country be a vertex and connect two vertices with an edge if those countries share a border. A famous problem that went unsolved for over a hundred years was the *four color problem.* Roughly this states that any map can be colored with at most 4 colors in such a way that no to adjacent countries have the same color. This problem motivated a lot of the developments of graph theory and was finally proved with the aid of a computer in 1976.

Transportation networks: The US Cities and highway system make a graph, with the cities being the nodes and the highways being the edges. More locally, one could consider the Norman Area bus system, where the nodes are the bus stops and two nodes are linked if they are successive bus stops on a single bus line. Similarly, an airline's flight network forms a graph. For these networks, the distance between two nodes is quite important in terms of time/fuel costs, and one can incorporate the distance into the graph by "weighting" the edges. The typical problems in transportation networks are designing efficient networks and finding efficient ways to route traffic (e.g., what's the best way for you to fly from Oklahoma City to Sydney, Australia?).

As a different kind of transportation network, but entirely analogous, electric power grids and water supply systems also form graphs.

Communication networks: Computer systems in a local network form a graph. So do the landline telephone cable systems and internet routing systems. These can also be thought of as "transportation" networks, where now what is being transported is data. Again, design and routing and principal issues in these networks.

Social networks: In sociology, economics, political science, medicine, social biology, psychology, anthropology, history, and related fields, one often wants to study a society by examining the structure of connections within the society. This could be friend networks in a high school or Facebook, support networks in a village or political/business connection networks. For these sorts of networks, some basic questions are: how do things like information flow or wealth flow or shared opinions relate to the structure of the networks, and which players have the most influence? In medicine, one is often interested in physical contact networks and modeling/preventing the spread of diseases. In some sense, even more basic questions are how do we collect the data to determine these networks, or when infeasible, how to model these networks?

The World Wide Web: One can form a graph of all webpages, and make an edge from Page A to Page B if there is a hyperlink from Page A to Page B. In this case, one should consider directed edges, meaning each edge has a direction which is pictorially indicated with an arrow. Here some basic problems are searching the web and ranking web pages (to get search results in a useful order). Due to the incredible size of the web and amount of information, searching is highly nontrivial. Web page ranking is closely related to the problem of determining how much influence players have in a social network.

Like the web, Twitter also gives rise to a directed social network, where the nodes are the users and the arrows point in the direction of "following."

Game theory/Discrete Dynamical Systems For games (deterministic or not) where there are a finite number of possible moves at each step, you can diagram the game as a graph. Here the vertices are the possible states of the games and the edges represent the moves going from one state to another. (Draw yourself the first few parts of the graph for Tic-Tac-Toe.) Thus the course of the game will be viewed as following a certain path in the associated graph and ending at a "terminal node." More generally graphs can be used to visualize discrete dynamical systems, and some ideas from dynamical systems are extremely useful in social/technological networks (e.g., Google's Pagerank algorithm).

Neurobiology/Artificial Intelligence: The brain is an immensely complex network, and some graph theory can be used to examine the structure of the brain. Many attempts at developing forms of artificial intelligence are based on the idea of neural networks, which are simple feedback networks that can be trained to perform tasks such as optical character recognition or speech recognition.

Being a bit broader with our terminology, we might consider transportation networks, communication networks and the World Wide Web as kinds of social networks as well (they are all products of societies). However I separated them above because they have different features and one typically asks different types for different types of networks. Indeed, communcation and transportation networks motivated a lot of "classical" graph theory, whereas study of social networks has led many newer directions in graph theory. In particular, with the advent of modern computing and the internet, understanding large networks is a major theme in modernd graph theory.

Our rough plan for the course is as follows.

First, we'll look at some basic ideas in classical graph theory and problems in communication networks. E.g., how can you analyze the efficiency or robustness of a networks. This leads to notions of distance, diameter, k-connectedness and network flow.

Second, we'll give a brief overview of some key themes in social networks and complex (large) networks. Here some topics are notions of centrality, clustering, degree distributions and small world phenomena. In these first two parts we'll get our hands dirty by writing and analyzing some algorithms for these things, but we'll also do exercises by hand.

Third, we'll look at *spectral graph theory*, which means using linear algebra to study graphs, and *random walks on graphs*. This will give us a useful way to study network flow for communication networks and do things like rank webpages or sports teams or determine how influential people are in social networks.

Finally, we'll study *random graphs* to get some insight into large networks. This will be the most "experimental" part of the course, in the sense that, while we hope to do a little theory, much will be learned by generating and working with examples on the computer. Time permitting, we'll spend some time discussing dynamic networks and modeling information flow/disease transmission.

In terms of computer software, we will work initially in Python (current version 2.7.6) for writing simple code to work with graphs, then use built-in algorithms in Sage (current version 5.12, which is compatible with Python 2, but not Python 3) to do more complex things, and possibly use Sage and Python in tandem. While I don't intend to make this a course on Computational Graph Theory or Algorithmic Graph Theory (i.e., how to program everything), I think it's important to have exposure to the nitty-gritty of at least some graph theory algorithms to truly understand the main issues of modern graph theory, which in large part stem from computational complexities.

However, be at peace—no previous experience with this software, or with programming, is required or expected.

The mathematical prerequisites for this course are: (1) some familiarity with proofs, as our Discrete Math or Linear Algebra courses; (2) some linear algebra (eigenvalues, eigenvectors, etc). We'll briefly review some of the necessary linear algebra as we go along, but you really should've have seen it before to hope to be able to get a good understanding of the third part of the course. We'll also need basic probability in the latter two parts of the course. I'll introduce this when the time comes, but it wouldn't hurt to have seen a little before.

Due to both the choice of topics and use of computer software, this will be a very non-traditional course in graph theory, which is why I couldn't find an appropriate textbook and will write my own notes. Based on the goals for the course, I won't be able to cover a lot of things that are covered in a typical graph theory course. Indeed we'll be quite pressed for time with my current goals. I will include some explanations of how to do things in Python and Sage in these notes, and snippets of sample code, but these notes will not contain detailed information on how to use Python and Sage in general—this information can be found elsewhere online, and will be covered in our Computer Lab Meetings.

For fun, I made an example of a social network graph involving some people you may know: the OU Math Department. See Figure 2. Let's take a look at this before we get started in earnest, to give you a better idea of some things we'll be doing throughout the course. This is a collaboration graph. The vertices are current or semi-recent previous faculty and postdocs within the OU Math Department. The postdocs (3-year positions) are all in blue, and regular faculty (tenure/tenuretrack) are in red (current) or black (previous). Two faculty are connected if they have been collaborators (co-authors of the same paper at least once). Faculty who have not collaborated with anyone in the department are not named, but just represented the 11 isolated nodes (nodes not connected to any other nodes). In fact there should be more isolated nodes, but I got tired after drawing 11 of them.

Note, I just made this on my own one afternoon, and there may be some additional collaborations that I've missed—in any case it will probably not be up-to-date for too long. However, let's try to get a little taste of network analysis by making some comments on this graph.

First of all, I was a bit surprised at how connected it is. While our department is quite friendly and social, most mathematics research is quite specialized, and most collaborations tend to occur among people in different universities. (I've written 9 papers since being at OU, of which 3 involve an OU collaborator.) Further, a lot more research in math is solo than in the other sciences. Even taking into account collaborations at the same university, one might expect that each research area is isolated from the other ones: e.g., number theorists just collaborate with other number theorists, geometers just collaborate with other geometers, and so on. However, forgetting the 11 isolated nodes for now, this graph consists of 3 components* of size 2, 1 of size 3, 1 of size 4, and then one large component of size 23: meaning most of the faculty are connected to each other through collaborations, instead of all the different research groups being disconnected from each other. This is part of the *small world phenomena*—in many social networks, things tend to be remarkably well connected.

For the rest of the discussion, let's restrict ourselves to the large component, which looks sort of like a reindeer. In fact, this component looks quite similar to the random graph on 15 vertices presented earlier (which to me, looks like a dog). This visual similarity, in the sense that they both look like balloon animals, was not due to any intentional planning on my part, but just something I noticed after making them. However, it's not mere coincidence—organically-formed networks can be modeled quite well by random graphs. That is what we'll focus on in the last part of the course.

Back to our discussion. Another aspect of small world phenomena, besides having a large component, is what is sometimes known as *six degrees of separation*. This is a notion about "how well connected" a network is. Define the distance $d(u, v)$ between two nodes u and v to be the minimum number of edges one needs to traverse to get from node u to node v. For example, there are many ways one can get from Shankar to Forester—directly, going through Brady, going through Brady, Dani and Clay, and so on. However, the shortest route just consists of 1 edge, so the distance from Shankar to Forester is 1. Similarly the distance from Shankar to Lousma is 2, Shankar to Rafi is 3, and so on.

Six degrees of separation refers to the notion that in many social networks, even though the network may be very large, most people who are connected are not more than 6 steps (distance 6) away from each other. An alternative interpretation is that most people are no more than 6 steps away from a given individual. The number 6 is not so important here, and is not a magic number for all networks—the idea is just that if you pick two people at random, they will be rather close. The farthest apart 2 of these 23 people are is Rubin and Basmajian, at distance 11. But if you pick 2 people in here at random, chances are that their distance will be much smaller. Looking at the alternative interpretation, I'm in the middle in some sense, and I'm within distance 5 from anyone else in the reindeer.

*A *(connected) component* is the part of a graph consisting of all nodes connected to a single given node by some sequence of edges.

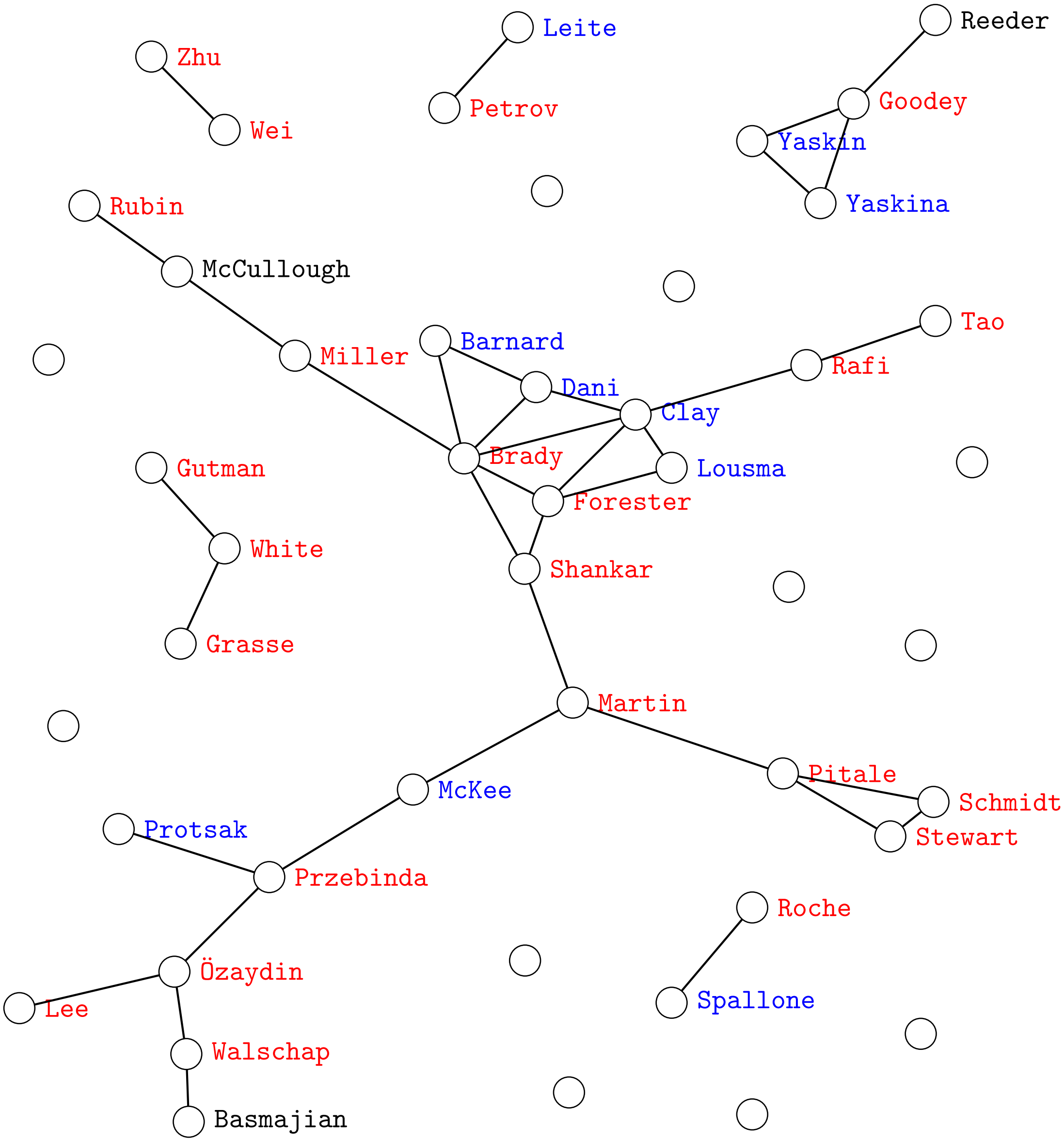


Figure 2: Collaborations within the OU Math Department

One question that's common in many social networks is: who is the "most popular"? For this, we can look at the *degree* of each node, which is the number of edges coming out of it, or equivalently, the number of other nodes it is directly connected to. Then one way to interpret "most popular" is simply having the highest degree, though for a collaboration network, "popular" isn't really a good word—I just mean who's worked with the largest number of people. In this case, Brady has the highest degree (6), followed by Clay (5), then Forester (4).

Another question is who has the most "influence"? You might first think that this is the same question as popularity, but it isn't. In fact for many things, e.g. predicting who might be able to win an election, influence may be more important than popularity, in the way I use these two terms here. Popularity (as we interpreted it) is a local trait. You can determine someone's popularity just by counting their neighbors (the neighboring vertices). Influence is a global trait. For instance, even though Przebinda and Özaydin are tied in terms of "popularity" (both degree 3), Przebinda is one step closer to McKee, and therefore most people in the reindeer, than Özaydin. In other words, Przebinda should be viewed as more "influential" as he is closer to more people in the graph than Özaydin is. Being closer to more people says you are more "central," and let's use the term central now instead of influential. (I don't want you to think our math department is full of clandestine politicking and power struggles, though some departments are! Also remember, this graph also doesn't represent the social structure of our department—only collaborations.)

There are several different ways to measure centrality. Here is a simple one. For each node u, define its centrality as the sum over all other nodes v (in the same component) of $\frac{1}{d(u,v)^2}$. (One could also define it without the squares.) Then the higher the centrality, the closer you are to more nodes. A more sophisticated way to define centrality involves the idea that you should weight these distances by how central v itself is. (Being distance 1 from a more central node is better than being distance 1 from a node on the outskirts.) It may not be clear how to make sense of a definition of centrality that already involves centrality, but this can be done with eigenvalues, and is one of two essential ideas behind Google's Pagerank algorithm.

Another phenomena you may notice is *clustering*. Look at the triangles in the graph. There are 6 in the large component and 1 in the component of size 4. However, the triangles in the large component aren't spread out evenly—5 of them are adjacent (the reindeer's head and hat). This phenomenon that triangles tend to group together in social networks is known as clustering. This just means there tend to be tight-knit groups within social networks. Often one measures clustering to say something about the structure of a network.

If you want to extrapolate qualitative information from the graph, you can, but due to the nature of this particular graph it won't be too precise. For example, let's say you want to use this graph to say something about the research interests of faculty members. (Probably a more natural use for this graph would be as an example in a study of how much being in the same institution factors into collaborations.) While my research interests are closer to those of Brady's or Forester's than say Rubin or Lee as suggested by the distances in the graph, you don't see that my research interests are closer to those of Rafi's than Miller's (or probably Brady's also), even though Rafi and I are farther apart in the graph. In fact, Pitale, Schmidt and I all work in the same area, and we are closer to Roche and Spallone than anyone in the top half of the reindeer. Even though the data is real, you should be careful not to overinterpret it as being a graph of our research interests, or how much we interact. Writing a paper with someone may or may not mean that you both work in the same are most of the time (my paper with Shankar was in an area that neither of us typically work in—in addition, research interests often change over time).

However, there are some things we can do to incorporate more information in the graph. First, one does not get a sense of the strength of the connections. For example Pitale and Schmidt have written many papers together, but Stewart has only written a couple papers with Pitale and Schmidt (these two papers consisted of all 3 authors and 1 additional one). One could weight the edges by counting the number of papers co-authored. This would make it clear that Pitale and Schmidt are closer to each other than either of them are to Stewart. It is also not clear from the graph whether a triangle represents, say, a single paper coauthored by 3 people or 3 different collaborations by each possible pair of the 3 people. This can be incorporated by allowing "faces" (in the sense of a face of a parallelopiped) in graphs, which are called *hypergraphs*. We will look at weighted graphs briefly, but not discuss hypergraphs in this course.

Another aspect not seen in this graph is that this is just a snapshot of all collaborations up to the present time. One can understand more about department research connections by looking at how the network evolves over time. Indeed, social networks tend to be dynamic networks with new nodes being added, nodes being removed and links changing all the time. This is another challenge in data collection and analysis for a given social network. (Further, not all of the collaborations between OU faculty in the graph actually occurred at OU—some collaborations began before one or both parties involved was at OU. Conversely, this might be a factor into who gets and accepts job offers at OU.)

Finally, the main reason for not being able to read too much into this graph is that it's just too small to get sophisticated information. This is the same problem as not having enough data in statistics. If you embed this graph in a larger collaboration graph of all mathematicians in the world, it should be much more clear that Pitale, Schmidt and I all work in the same area, where as Shankar and I typically do not. Even though there may be some "random connections" between people in different research areas, in a huge collaboration graph the number of these random connections will be quite small and can be basically ignored by using more sophisticated analysis techniques. While this may all seem rather frivolous, these ideas actually have immense potential for applications—for example these ideas about using collaboration graphs to distinguish different research areas can be used in things like *machine learning/artificial intelligence* and *face/pattern recognition.*