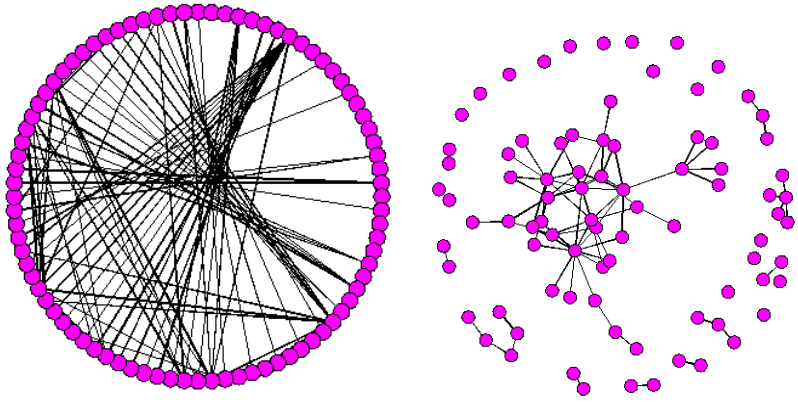
Statistical Analysis of Networks
Katie St. Clair
Prereqs: Math 275 and either Math 245 or a CS course

Suppose you collected data on study habits and GPA for your statistics project by sending survey invitations to all your Carleton friends on Facebook. You also ask your friends to ask their Carleton friends, and so on, to take the survey. The data you collected is obviously not a simple random sample from the Carleton student population, but can you use it to make (unbiased) inferences to this population? Can you use it to model GPA as a function of variables like study hours per week, major field, etc? Can we determine which, if any, variables are associated with Facebook friendship connections? Can we determine if there are any friends in this group who influence study habitats more than others?

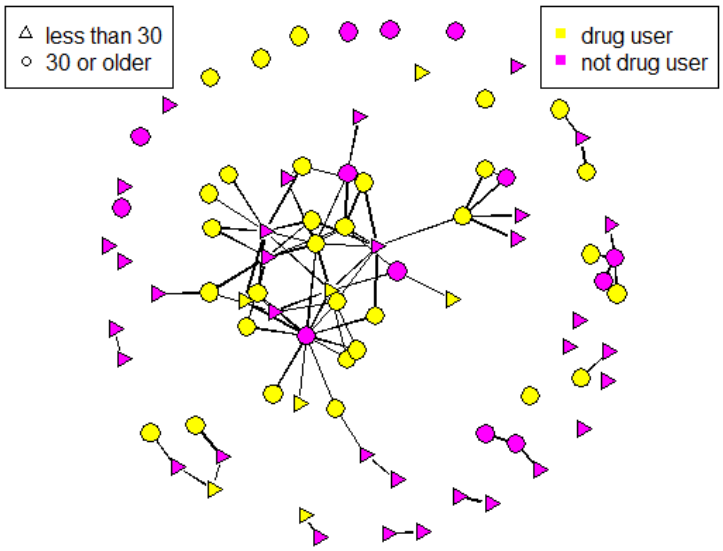
The answer to all these questions is yes (under the right conditions) if we use statistical analysis and inference methods for data collected from a network. A "network" can simply be defined as a system of interconnected objects, or we can more formally define it as a mathematical graph: for the survey data the vertices are students and the edges between vertices are Facebook friendships. We can measure variables on both the vertices and edges in this network. This social network is a simple example of a network, other examples are plentiful: WWW, Twitter, transportation and communication networks, energy grids, neuron networks, predator-prey systems, collaborative research networks.

In this comps, we will start studying statistical methods for networks in the usual place: data and EDA.

- **Data**: How do we define a network? How can we store data collected on a network?
- **Visualizing a network**: There are many methods for "mapping" a network, how do we choose a good map? Here are two views (circular vs. a force-directed algorithm) of the same network of individuals from a high-risk population (edges indicate friendship):



How can we add more data (variable information) to these maps? Here is the same network mapped above, but with vertex shape showing age and vertex color showing drug use:



- **Descriptive statistics**: *Centrality*: what are the "important" or influential vertices? Who are the influential players in social/collaborative network? What is a critical router in an internet network? *Cohesion*: to what extent are subsets of vertices are stuck together? How connected is the graph? How many "degrees of separation" (or average path length) in a film actor network? How many edges or vertices need to be removed to stop the flow of information/viruses/disease in a connected network?

We will then focus on modeling networks, both from the mathematical and statistical perspective:

- **Mathematical models**: Probabilistic models, like the random graph model, that can be used to generate a network with certain theoretical properties.
- **Statistical models**: Parameterized models that are designed to be a realistic representation of observed network data and we can estimate the model parameters from the data. These models can be used to test whether edges (e.g. friendships) are constructed at random, or whether edges associated with a vertex variable (e.g. is friendship associated with gender, major, freshman dorm assignment, etc.)

If we have time, we can explore one or more advanced topics and applications:

- Network sampling designs and inference methods: How can we obtain a (random) sample from a network and use it to make (unbiased) inferences to the larger population?
- Edge predictions: If we only observe a subset of edges (or non-edges) in a network, how can we predict edges (or non-edges) that were not observed?
- Modeling complex systems as network processes: How does information flow through a network? How does a contagious disease spread through a population?