MAE-253: Homework 2

John Karasinski

May 3, 2016

Problem 1: The Cayley tree

A Cayley tree is a symmetric regular tree emanating from a central node of degree k. Every node in the network has degree k, until we reach the nodes at the maximum depth d that have degree one and are called the leaves of the network. For a Cayley tree of degree k and depth d calculate:

a) How many vertices are exactly distance one from the center?

3

b) How many vertices are exactly distance two from the center?

6

c) How many vertices are exactly distance l from the center?

By induction, $k(k-1)^{l-1}$ for $l > 0$. For $l = 1, 2, 3, 4, 5$, for example, this leads to $3, 6, 12, 24, 48$.

d) What is $n(l)$ the total number of vertices contained within distance l from the central vertex? (Include the central vertex in this count).

$n(l)$ is calculated by simply summing to distance l and adding 1,

$$n(l) = 1 + \sum_{l=1}^{l} k(k-1)^{l-1} = 1 + k\frac{1-(k-1)^{l}}{1-(k-1)}.$$

e) Present an argument that the Cayley tree has a small world diameter by showing that $d \sim \log(n)/\log(k)$. (It does not have to be rigorous, but show your reasoning.)

Since the distance of any leaf node from the central node is l, the diameter of the Cayley tree is simply $d = 2l$. We then have

$$\begin{aligned} n(l) &= 1 + k\frac{1-(k-1)^{l}}{1-(k-1)} \\ &= 1 + k\frac{1-(k-1)^{(d/2)}}{1-(k-1)} \end{aligned}$$

Plugging in $k = 3$ and solving for d leads to

$$\begin{aligned} n &= 1 + 3(2^{d/2} - 1) \\ (\frac{n-1}{3}) + 1 &= 2^{d/2} \\ \ln((\frac{n-1}{3}) + 1) &= \frac{d}{2} ln(2) \\ \frac{2}{\ln(2)} \ln((\frac{n-1}{3}) + 1) &= d \end{aligned}$$

For $n \gg 1$ we then have

$$\begin{aligned} d &= 2\frac{\ln(n)}{\ln(2)} \\ d &\approx 2\frac{\ln(n)}{\ln(k-1)} \end{aligned}$$

which suggests a small world ($d \propto \ln n$).

Problem 2: Finite size scaling

Consider a network where node degree k follows a power law degree distribution $p_k = (\gamma - 1)k^{-\gamma}$, with $\gamma > 1$. We will approximate k as continuous and then let P_K denote the cumulative distribution function (CDF) which is the probability a node will have degree less than or equal to K,

$$P_K = \int_1^K p_k dk.$$

Here we will work out an estimate for the maximum node degree, K_{max}, that one would expect to see in a network of size N with a power law degree distribution. Operationally, we define the expected value of K_{max} for a network of size N to be the value of degree when we expect only one node bigger than this value:

$$N(1 - P_{K_{max}}) \approx 1.$$

Using this show that $K_{max} \approx N^{1/(\gamma-1)}$ and evaluate the expression explicitly for $\gamma = 2, 3, 4$.

Integrating

$$\begin{aligned} P_K &= \int_1^K (\gamma - 1)k^{-\gamma} \\ &= \frac{(\gamma - 1)k^{1-\gamma}}{1-\gamma}|_1^N \\ &= 1 - K^{1-\gamma} \end{aligned}$$

Plugging this expression in to $N(1 - P_{K_{max}}) \approx 1$, we have

$$\begin{aligned} N(1 - [1 - K_{max}^{(1-\gamma)}]) &\approx 1 \\ K_{max}^{(1-\gamma)} &\approx \frac{1}{N} \\ K_{max} &\approx \frac{1}{N}^{1/(1-\gamma)} \\ K_{max} &\approx N^{1/(\gamma-1)} \end{aligned}$$

Plugging in 2, 3, 4 into the formula for K_{max} results in

$$\begin{aligned} K_{max}(2) &\approx N \\ K_{max}(3) &\approx N^{1/2} \\ K_{max}(4) &\approx N^{1/3} \end{aligned}$$

which indicates that larger values of γ lead to lower values of K_{max}.

Problem 3: Analysis of a real-world network

For this problem you must find a data set of a real-world network. It could be a recommendation network of books constructed via amazon.com, a flight network for an air-line, a collaboration network of scientists or movie actors, a protein-interaction/gene-interaction network, a piece of the Amtrak rail network, a Facebook network, etc. The network should have somewhere between 200 to 1000 nodes.

a) Describe your data set and where/how you obtained it. Is this a directed or undirected graph? Are there several components, or is it all one connected component?

My dataset describes a network of all astronauts that have ever flown together in a spaceship that was either leaving or returning to Earth. Each node is an astronaut, and each edge relays that the two astronauts flew together. Since the astronauts flew together, this results in an undirected graph. I obtained the dataset from Wikipedia by crawling a list of all astronauts, and then did some minor manual cleansing of the data.

b) How many nodes and edges are present? What is the average degree? (If it is a directed graph give values for both average in and out-degree.)

There are 519 nodes and 2497 edges. The average degree is 9.6.

c) Plot the degree distribution (again, if directed, plot both in and out-degree distributions). Identify the distribution that best fits your data, choosing from Gaussian, exponential, power law. (If you want to get more sophisticated, consider also power law with a cutoff and log normal distributions.)

Three different functions were fit to the resulting degree distribution, and had the RMSE from the best fit and the true distribution was calculated. Though the Gaussian function performs slightly better than the exponential function, it requires an extra degree of freedom. As a result of this, the exponential function is considered to perform the best here.

Name	Function	RMSE
Gaussian	$Ae^{-\frac{(x-B)^2}{2C^2}}$	7.70
Power law	Ax^B	10.17
Exponential	AB^x	7.73

Table 1: Distributions fit to astronaut degree distribution with resulting RMSE.

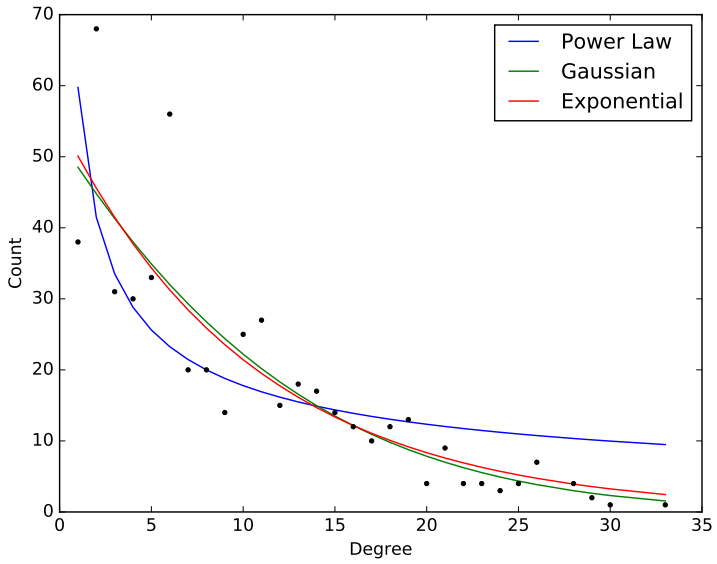

Figure 1: Degree distribution with different distribution fits. The power law performs the worst, while the Gaussian and exponential fits perform nearly the same. In this case, the Gaussian performs lightly better.

d) Visualize the network. Try to use color or size to display interesting attributes of your data (degree, age, high-clustering, etc). You may want to label the nodes with their identities.

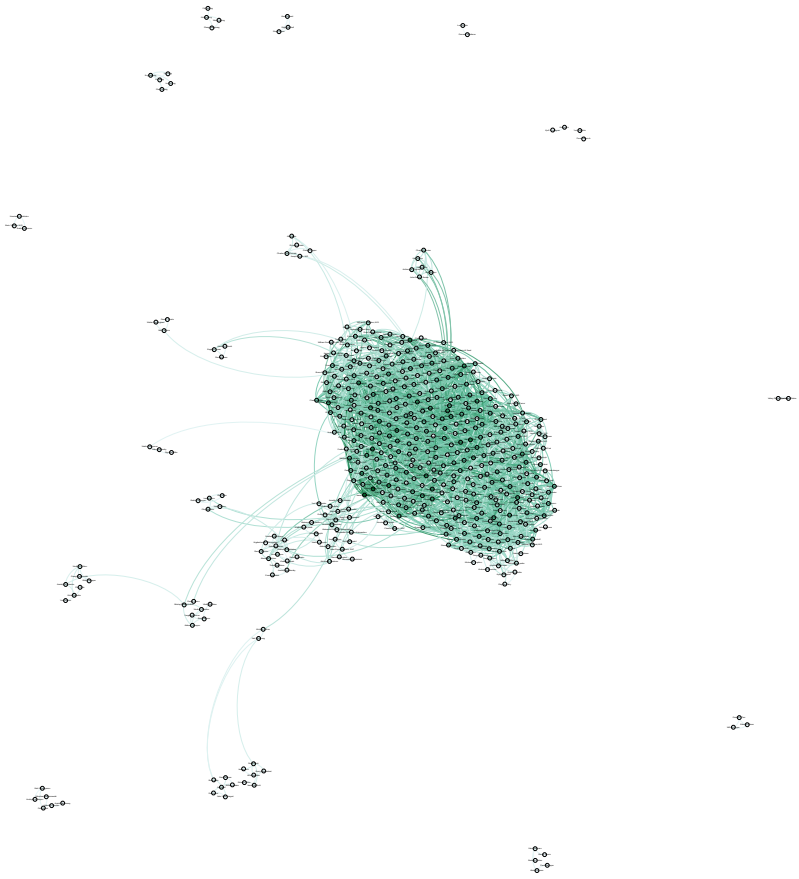

Figure 2: This graph shows the degree of each node. Darker green shows a higher degree, and a lighter green in a lower degree. The nodes are labeled with the identities of the astronauts.

e) Run a community detection algorithm on your network. How many communities did you find? What is the size distribution of the communities? Use the visualization of point d) and color code the communities. Can you interpret what you found?

The community detection algorithm[1] found 22 communities.

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, *Fast unfolding of communities in large networks*, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

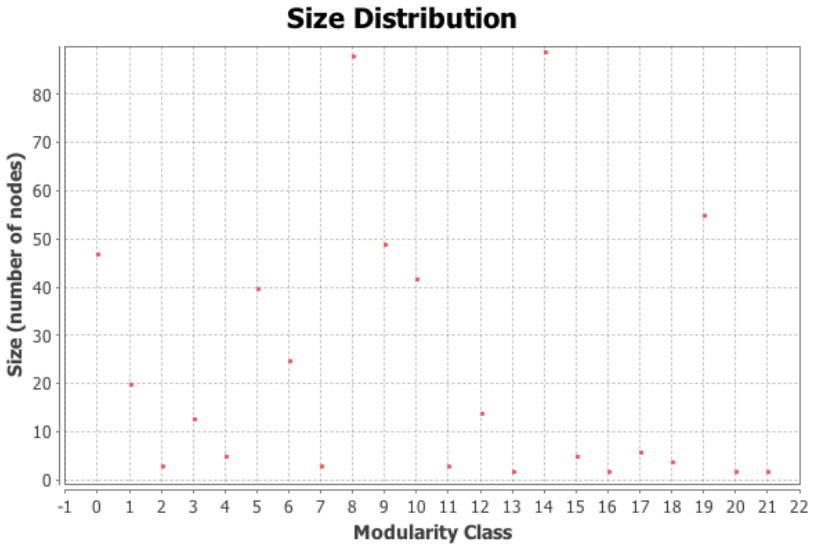

Figure 3: Distribution of community sizes.

The small communities show groups of (mostly early) astronauts that only flew one mission. As such, they form a small community with the astronauts that they flew their single mission with. The blue community that comes off the primary cluster shows several of the Apollo missions, and the pink and orange communities just above it show groups of early Russian astronauts. The large central hub formed due to collaboration between US and Russian space agencies in the past few decades. Parts of the hub can be further decomposed into early/middle/late shuttle flights.

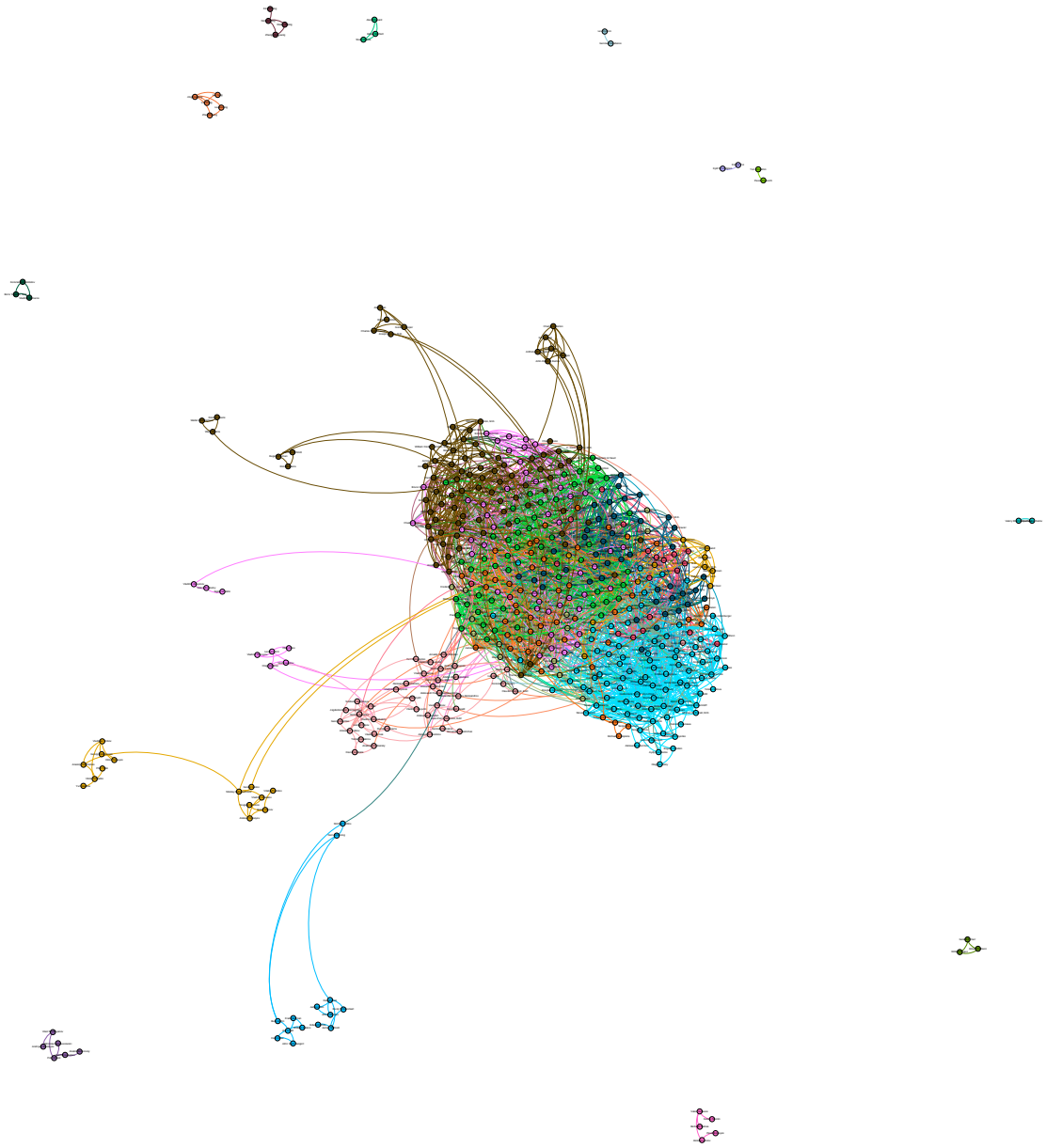

Figure 4: This graph shows the community structure of the network. Each community is colored differently, and each node is labeled with the astronaut's name.