HW1

HW 1: Exploring network data

These questions come from Chapter 2-4 of the SANDr book.

1. First, download the Banerjee et al 2013 data we've discussed several times in class and load it into R. Use this data for the questions below (except Q6, where you can either simulate or use these data).
2. For this exercise, let's compare graph structure across the 75 villages. Within each village, compute some (2-3) individual level graph statistics (e.g. degree) and plot the distribution for each village. To do this, I recommend a boxplot where every box represents the village and in each box are individual level statistics individuals in each graph.
3. For the individual level graph statistics above, identify individuals who are extreme (either on the high or low end). Use covariates to provide aggregate summaries of the characteristics of these individuals.
4. Now, let's look at some graph level statistics. Make a histogram of 2-3 village level statistics. Use covariates to describe villages that are high or low on either end.
5. Use regression models to explore the association between village-level covariates (you can aggregate individuals ones if you'd like) and graph statistics.
6. Replicate Figure 2.2. Describe a data setting where you would expect to see each of the four types of graphs.
7. For one of the villages, construct a visualization of the graph. Label the nodes based on covariates. Do this for a couple of covariates.
8. Run the samecaste regression we talked about in class. For this, you can assume that $\delta = 0$ (that is, there are no other informative covariates). Describe why standard inferential techniques are or aren't valid here.