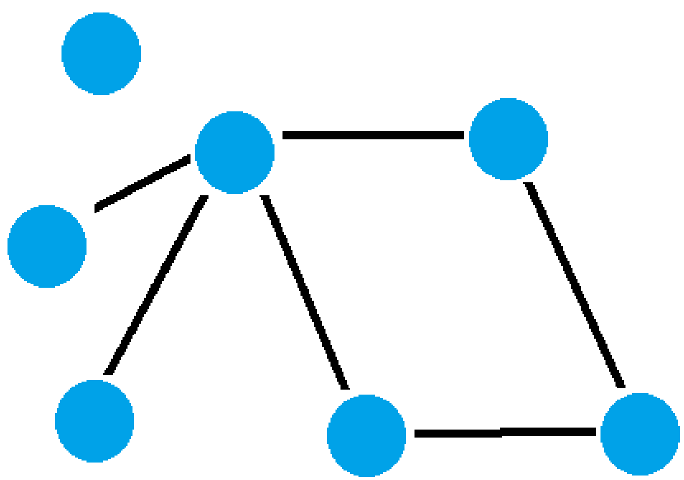
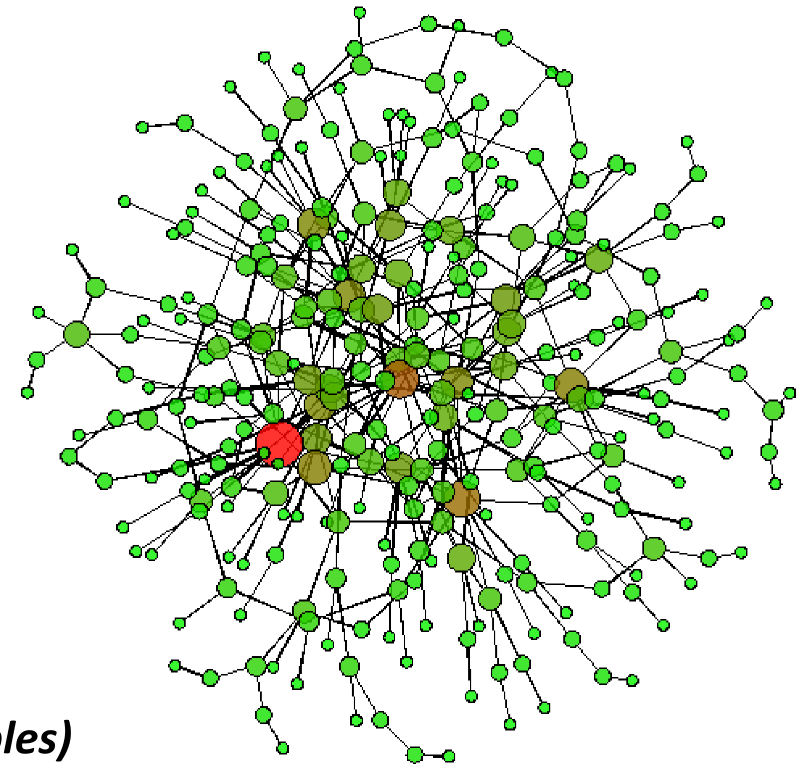
Imputing Attitudes on a Network

Richard Vale, Nov 12 and Dec 9-10 2014

Network: a collection of nodes (points) with relationships (lines, undirected in this case) between them.





(Examples)

The Problem

- Each node in the network has an “attitude” score (between 0 and 10.)

- The attitude scores are known for only some of the nodes.

- Want to impute (=guess) the attitude scores for the other nodes.

Difficult because …

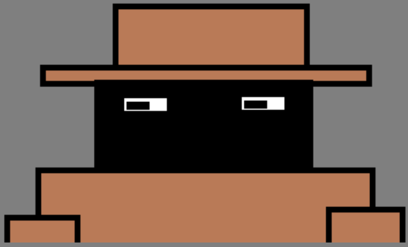
- What does the data represent? *Don’t know.*

- What does a typical data set look like? *Don’t know.*

- How many examples are available? *Only one, with n=395 nodes and 272 missing attitude scores.*

- What are we trying to achieve? *Don’t know.*

MYSTERY

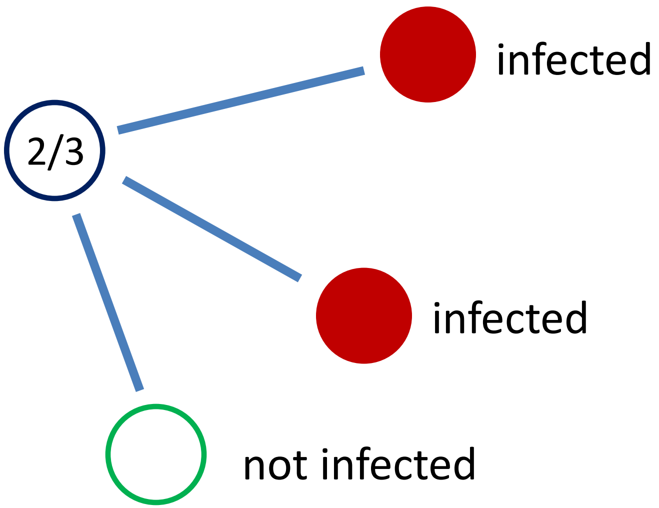


Approach: Find a model based on general theory and hope that it is useful

Idea: you will have a bad attitude if you are connected to people with a bad attitude (otherwise, why bother looking at networks at all?)

Epidemic model: Time dependent. Nodes are either infected or not infected.

Probability of infection at next step = proportion of adjacent nodes which are infected.



Can't quite use the epidemic model, but can use a variant of it:

Treat as an epidemic, with r_i = some measure of degree of infection with values in [0,1]. Given the unobserved r_i, suppose the observed attitude scores are given by

$$a_i = 10\,\frac{r_i + \beta \sum_{d(j,i)=1} r_j}{\deg(i) + 1}$$

$d(j,i)$ = distance between j and i

$\deg(i)$ = number of nodes adjacent to i

β = a tuning parameter (usually taken to be 1 in tests)

The n parameters r_i are linearly related to the observed attitude scores a_i via $a_i = Br_i$ for some $k \times n$ matrix B where k = number of known attitude scores.

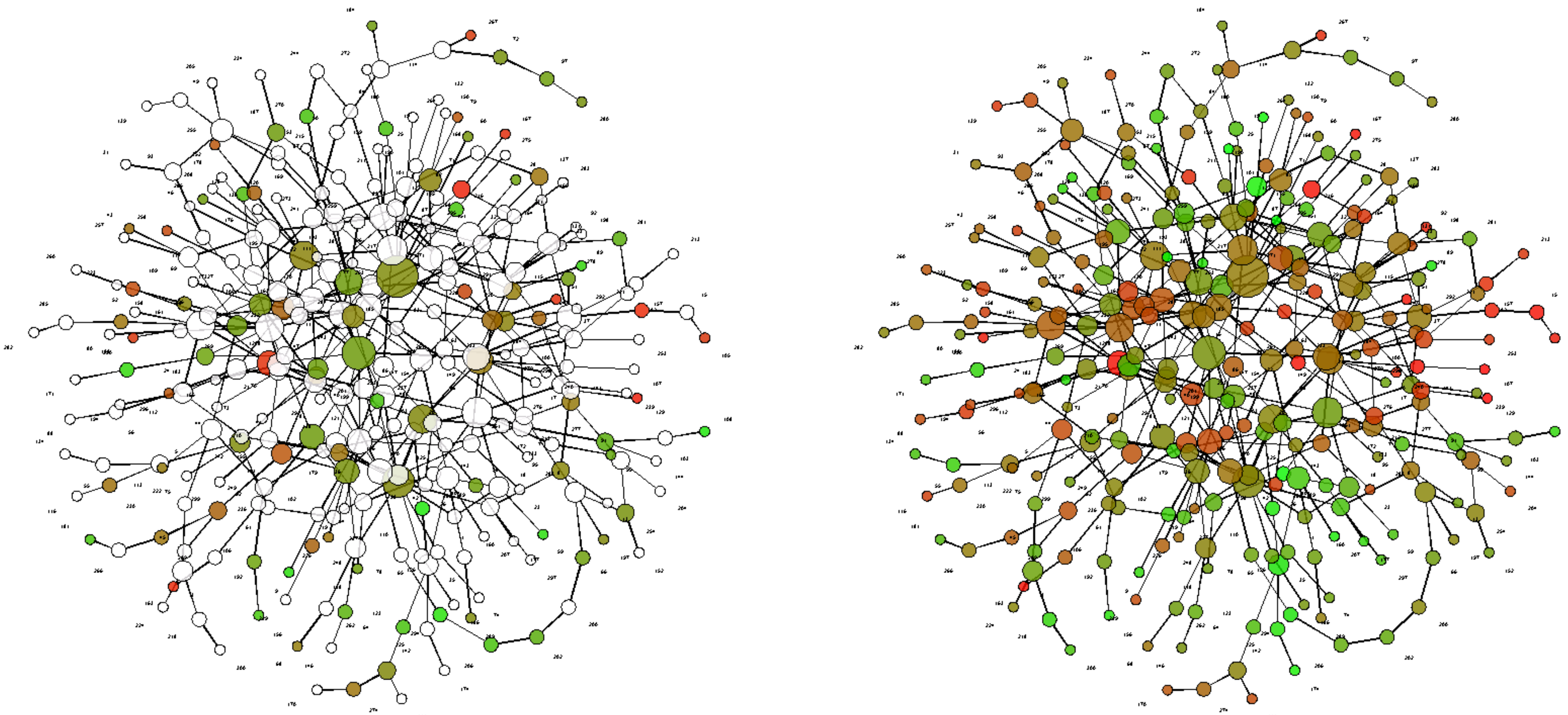
Strategy:

1. Find a vector (r_i) which gives a local minimum of

$$\|B(r_i) - (a_i)\|, \qquad 0 \leq r_i \leq 1.$$

2. Having made a choice of (r_i), impute the missing a_i using the equation $a_i = 10 \dfrac{r_i + \beta \sum_{d(j,i)=1} r_j}{\deg(i) + 1}$

3. (Bonus: do it several times with random starting points to get a feel for the uncertainty of the results.)



Evaluating the results

- Since we have a model with n parameters r_i and k data points, in-sample performance is not a useful measure of performance.

- It is not clear exactly how cross-validation could be performed (should you delete edges as well as nodes? If so, how?)

- It is not clear what measure should be used to evaluate the performance (sum of squared errors is useless. Sum of absolute errors is also problematic; does the end user care whether it’s a 6 or a 7?)