20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Network Centrality and Super-Spreaders in Infectious Disease Epidemiology

A.H. Dekker

Graduate School of Information Technology & Mathematical Sciences of the University of Ballarat, Ballarat, Victoria, Australia
Email: dekker@acm.org

Abstract: So-called "super-spreaders," who are particularly effective in transmitting infectious diseases, are of concern to public health officials. In this paper, we study the phenomenon of "super spreaders" using the Susceptible-Infected-Recovered (SIR) model of infectious disease. In particular, we explore *network centrality measures* as potential predictors of the average number of other people who will be infected by a given node in a social network.

We consider six centrality measures: the *node degree d*, *closeness centrality* C_C, *valued centrality* C_V, *Jordan centrality* C_J, *betweenness* C_B, and *eigenvector centrality* C_E. These measures are correlated to varying extents, with a 0.97 correlation between closeness centrality and valued centrality, but only a 0.14 correlation between Jordan centrality and betweenness.

We report simulation experiments in which the duration of infection is one time-step, and infection begins with a solitary individual. Results are averaged over 1,000,000 simulated runs. We use a varied sample of 15 social networks, and vary the probability q that, given an infected person x and a susceptible person y connected by a link in the social network, the infection spreads from x to y.

In the highly infectious case with $q = 0.9$, the best predictor of the average number of other people who will be infected by a given node in a social network is the betweenness C_B, with an R^2 value of 81.4%. For q in the range 0.5 to 1.0, the product $0.74\ C_B^{0.49}\ d^{1.03}\ q^{0.24}$ has an R^2 value of 87.7%, and this leads to a method for targeted vaccination.

In the less infectious case with $q = 0.05$, the node degree d is the best predictor of "super-spreading." In contrast to Macdonald *et al*. (2012), eigenvector centrality C_E is not a good predictor. This is because the recursive definition of eigenvector centrality sometimes results in it simply highlighting one densely connected network subset, rather than acting as a true centrality measure.

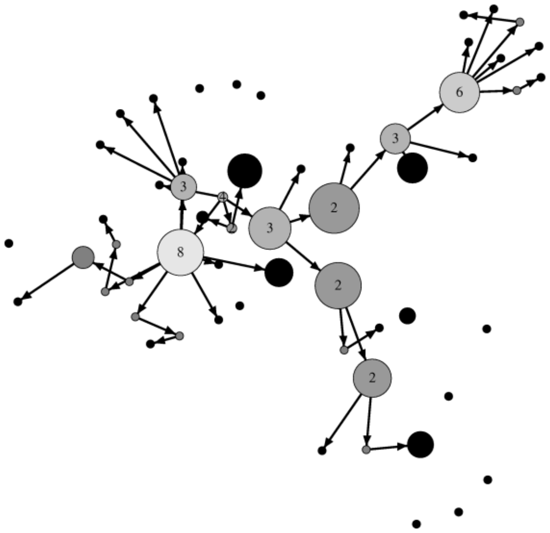

Figure (i). Spread of infection in one social network. Only links along which infection spreads are shown. Node area shows betweenness C_B, which is a good predictor of the number of other nodes infected, indicated with node colour and label. The two light-coloured "super-spreaders" have high betweenness scores.

Keywords: *Networks, epidemiology, SIR model, centrality, super-spreaders*

1. INTRODUCTION

Infectious diseases are a perennial concern for public health officials. As infections spread through social networks, one category of person of special concern is the "super-spreader," who is particularly effective in transmitting the disease (DMERI SARS Investigation Team, 2005).

Here we study the phenomenon of "super spreaders" using the well-known Susceptible-Infected-Recovered (SIR) model of infectious disease (Anderson and May, 1991; Skvortsov *et al.*, 2007; Dekker, 2008a). We assume that the duration of infection is one time-step, and that infection begins with a solitary individual. The only parameter is then the probability q that, given an infected person x and a susceptible person y connected by a link in the social network, the infection spreads from x to y.

Figure 1 shows an example of simulated disease spread, where the social network is a 57-node connected subset of a scientific coauthorship network (Newman, 2006). This diagram resembles the spread of real diseases (CDC, 2003), although the numbers infected are lower.

We explore a variety of network centrality measures to see which best predict the extent to which a node spreads the infection. That is, we wish to predict numbers like those in Figure 1, for a varied sample of 15 social networks, and averaged over 1,000,000 simulated runs.

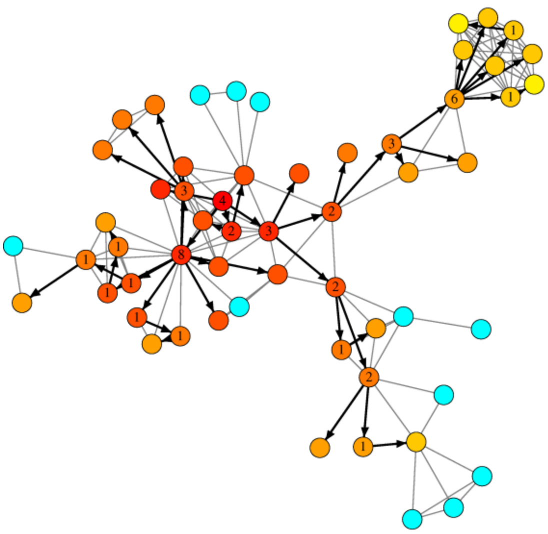

Figure 1. Spread of infection in a social network, starting with the red node labeled "4." The infection probability is q = 0.5. Blue nodes are never infected. Numbers in each node show how many other nodes it infects. Note the "super-spreader" infecting 8 other nodes. The grey links do not participate in spreading the infection.

2. CENTRALITY MEASURES

In this paper we consider six measures of network centrality: the *node degree d*, and five other measures. Four of these other measures were studied in Dekker (2008b).

Closeness centrality C_C is one of the most widely used centrality measures (Wasserman and Faust, 1994). However, it has the disadvantage of being always zero for disconnected networks. The closeness centrality of node x is defined as the reciprocal of the average of the distances $D(x, y)$:

$$C_C(x) = \frac{n-1}{\sum_{y \neq x} D(x,y)} = \frac{1}{\underset{y \neq x}{AVG}\, D(x,y)}$$

where n is the number of nodes in the network, and $D(x, y)$ is the shortest-path network distance between the nodes x and y: that is, the number of links in the shortest path between x and y.

Valued centrality C_V was introduced as an alternative to closeness centrality (Dekker, 2005). Although originally intended for valued networks, with ties of varying strength, it is equally applicable to ordinary

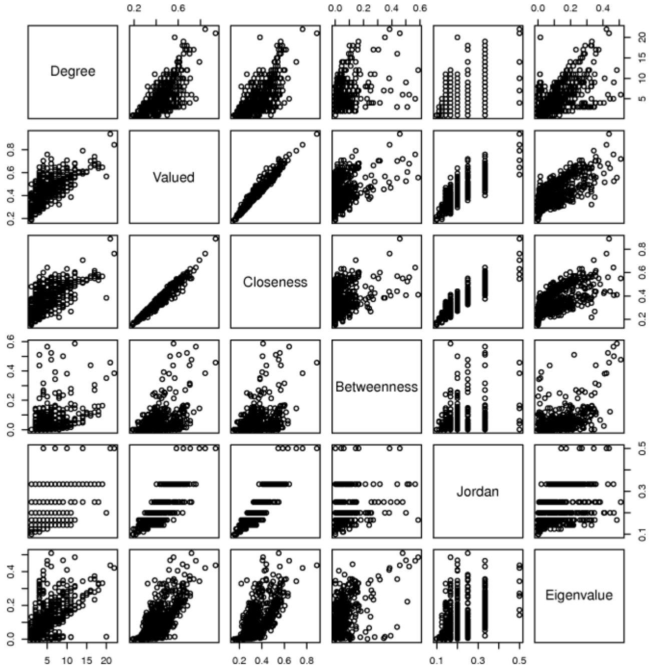

Figure 2. Correlations between six centrality measures.

networks. It has the advantage of being well-defined even for disconnected networks. It is defined similarly to closeness centrality, but is the average of the reciprocal of $D(x, y)$, rather than the reciprocal of the average:

$$C_V(x) = \frac{1}{n-1}\left(\sum_{y\neq x}\frac{1}{D(x,y)}\right) = \underset{y\neq x}{AVG}\left(\frac{1}{D(x,y)}\right)$$

Jordan centrality C_J was introduced implicitly by Hage and Harary (1995), and is derived from the "Jordan centre" of a network. It uses only the largest of the distances $D(x, y)$:

$$C_J(x) = \frac{1}{\underset{y\neq x}{MAX}\ D(x,y)}$$

Hage and Harary suggest that identifying the nodes with the highest C_J can offer useful insights into a network.

Betweenness C_B is based on counting the number of geodesics (shortest paths) g_{xy} between nodes x and y, and looking at the number $g_{xy}(z)$ which travel via node z:

$$C_B(z) = 0.001 + \frac{2}{(n-1)(n-2)}\sum_{x\neq z}\sum_{x<y\neq z}\left(\frac{g_{xy}(z)}{g_{xy}}\right)$$

Since betweenness scores are sometimes zero, we adjust them by adding 0.001, so that taking logarithms is possible. The number 0.001 is chosen to be just below the smallest nonzero non-adjusted score of 0.0014.

Eigenvector centrality C_E is defined by the unique all-positive eigenvector v satisfying $Av = \lambda v$, where A is the adjacency matrix. This will be the eigenvector corresponding to the largest eigenvalue λ. The value of C_E for the i^{th} node will be the i^{th} element of v, satisfying the recursive equation:

$$C_E(i) = \frac{1}{\lambda}\sum_j A_{ij}C_E(j)$$

Eigenvector centrality correlates with social attributes like prestige (a prestigious person is linked to by other prestigious people). Google's PageRank for web pages is a variation of eigenvector centrality.

Table 1 and Figure 2 show the correlations between these six measures. These correlations range from a very high 0.97 for valued centrality C_V and closeness centrality C_C (which measure essentially the same thing) to 0.14 for *betweenness* C_B and *Jordan centrality* C_J (which measure quite different things).

Table 1. Correlations between six network centrality measures (values ≥ 0.84 shown in blue).

	Degree *d*	Valued C_V	Closeness C_C	Betweenness C_B	Jordan C_J	Eigenvalue C_E
Degree *d*	**1**	0.70	0.64	0.43	0.51	0.58
Valued C_V	0.70	**1**	**0.97**	0.40	**0.84**	0.66
Closeness C_C	0.64	**0.97**	**1**	0.32	**0.91**	0.56
Betweenness C_B	0.43	0.40	0.32	**1**	0.14	0.50
Jordan C_J	0.51	**0.84**	**0.91**	0.14	**1**	0.36
Eigenvalue C_E	0.58	0.66	0.56	0.50	0.36	**1**

The 0.001 added to betweenness scores allows logarithms of the centrality measures to be taken, and this avoids problems due to skewness in betweenness scores. Correlations between the logarithms of the centrality measures are roughly similar to Table 1, ranging from 0.18 to 0.97. The correlation between degree and betweenness C_B increases from 0.43 to 0.62 when logarithms are taken. However, the correlation between degree and eigenvalue centrality C_E decreases from 0.58 to 0.47, since the use of logarithms reveals considerable scatter among nodes with low degree and low eigenvalue centrality.

3. NETWORK SAMPLE FOR EXPERIMENTATION

We explored disease-spreading in a varied sample of 15 social networks: 10 natural (with between 10 and 62 nodes) and 5 artificial (all with 60 nodes). The sample was chosen so that the real-world networks would not be outweighed by the artificial ones. Since our data points are individual nodes, this gives 635 data points: 335 from natural networks and 300 from artificial ones.

We are interested in predictors that can, if necessary, be applied to portions of a network, even when the full size of the network is unknown. Using a sample of networks of different sizes facilitates this by ensuring that there are no hidden dependencies on the network size. The networks used for our experiments were:

- the two island voyaging networks of Hage & Harary (1995) – one with $n = 12$, average degree 3.33, and average network distance $D_{\text{ave}} = 2.53$; and the other with $n = 10$, average degree 3.4, and $D_{\text{ave}} = 2.11$;
- the Florentine families network in Wasserman & Faust (1994) – with $n = 15$, average degree 2.67, and $D_{\text{ave}} = 2.49$, after deleting an isolate;
- two work communication networks – one with $n = 33$, average degree 6.97, and $D_{\text{ave}} = 2.11$; and the other with $n = 47$, average degree 6.64, and $D_{\text{ave}} = 2.37$;
- two Internet social networks – one from a newsgroup with $n = 40$, average degree 2.6, and $D_{\text{ave}} = 3.95$; and the other from a blogging network with $n = 25$, average degree 6, and $D_{\text{ave}} = 1.88$;
- an association network between dolphins in a community living off Doubtful Sound, New Zealand (Lusseau *et al*., 2003), with $n = 62$, average degree 5.13, and $D_{\text{ave}} = 3.36$;
- a connected subset of a scientific coauthorship network (Newman, 2006), with $n = 57$, average degree 5.23, and $D_{\text{ave}} = 3.66$;
- a social network from a karate club at a US university (Zachary, 1977), with $n = 34$, average degree 4.59, and $D_{\text{ave}} = 2.41$;
- two random (Erdős-Rényi) networks (Bollobás, 2001), each with $n = 60$ nodes, one with average degree 4 and $D_{\text{ave}} = 3.12$, and the other more dense, with average degree 8 and $D_{\text{ave}} = 2.16$;
- two scale-free (preferential-attachment) networks (Albert and Barabási, 2002; Barabási, 2002), each with $n = 60$ nodes, one with average degree 3 and $D_{\text{ave}} = 3.57$, and the other more dense, with average degree 6 and $D_{\text{ave}} = 2.36$; and
- a small-world network, generated by applying the Watts rewiring process (Watts and Strogatz, 1998; Watts, 2003) to 10% of the links in a 60-node antiprism (the resulting network has average degree 4 and $D_{\text{ave}} = 3.74$).

The first 10 networks in the sample were also used in Dekker (2008b) and Dekker (2010).

4. FIRST EXPERIMENT: HIGH INFECTION PROBABILITIES

In our first experiment, we simulated disease-spreading from a random starting point, with a probability of infection of 0.9 for each link in the network. For each node, we calculated the number of other nodes it infected, averaged over 1,000,000 runs. We then explored the ability of the six centrality measures to predict these numbers, using power-law regression. The second column of Table 2 shows the results. The best predictor, by a considerable margin, was the (adjusted) betweenness score C_B, which predicted the number of nodes infected with an R^2 value of 81.4% (i.e. a correlation of 0.90). Figure 3 illustrates this.

These results are explained by the fact that most nodes are being infected, but the "wave of infection" is most likely to travel via the nodes with high betweenness scores C_B. As Christakis and Fowler (2010) point out, more central individuals tend to become infected first. Our results contradict those of Kitsak *et al.* (2010), who find betweenness is not a good predictor. This is due to the use of lower probabilities q in their work.

Table 2. Results of power-law regression in predicting number of nodes infected (for $q = 0.9$ and $q = 0.05$), using R^2 as an indicator of prediction success. The highest entries in each column are shown in blue.

Measure	R^2 **($q = 0.9$)**	R^2 **($q = 0.05$)**
Betweenness C_B	**81.4%**	34.3%
Degree d	62.1%	**79.1%**
Valued centrality C_V	23.4%	73.9%
Eigenvalue centrality C_E	20.0%	30.3%
Closeness centrality C_C	15.0%	57.0%
Jordan centrality C_J	5.6%	38.5%

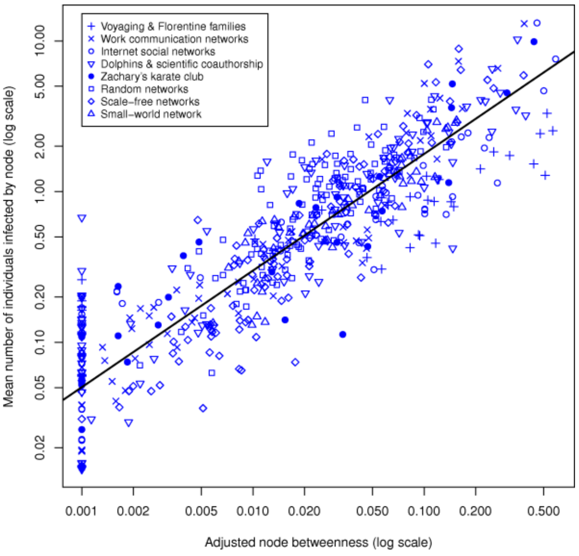

Figure 3. Betweenness C_B as a predictor of nodes infected. The R^2 value here is 81.4%.

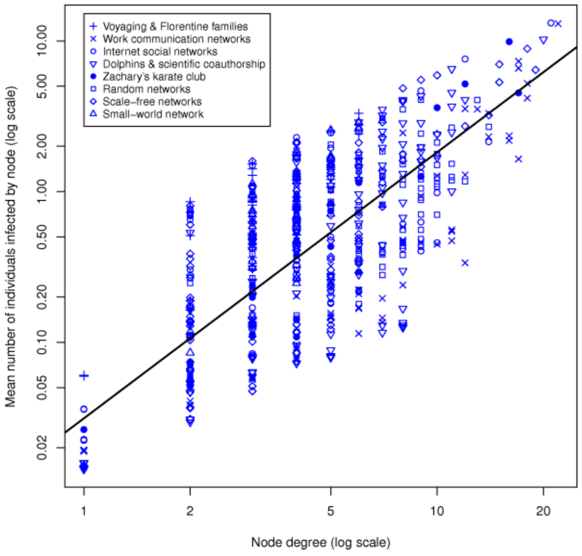

Figure 4. Node degree d as a predictor of nodes infected. The R^2 value here is 62.1%.

The second-best predictor is simply the node degree, with an R^2 value of 62.1% (see Figure 4). This is explained by the fact that nodes with high degree have more outgoing links and therefore are likely to infect more nodes (as well as being more likely to be infected early). In contrast to Macdonald *et al*. (2012), eigenvector centrality C_E is not a good predictor. Even a two-variable model using eigenvector centrality and node degree has an R^2 value of only 62.9%, which is little better than node degree on its own. Combining the two best predictors (betweenness C_B and node degree) does better, with an R^2 value of 89.7%. In fact, this dual predictor also handles variation in the infection probability q, to cover the range 0.5 to 1.0. The best three-variable predictor including q is the product 0.74 $C_B^{0.49}$ $d^{1.03}$ $q^{0.24}$ (i.e. approximately $0.74d\sqrt{C_B}\sqrt[4]{q}$), shown in Figure 5. This has an R^2 value of 87.7% (a correlation of 0.94).

These results suggest the use of the product $d\sqrt{C_B}$ as a measure for targeting nodes for vaccination. We test this with q = 0.75. For this infection probability, the denser networks almost always become totally infected. Using targeted vaccination (with the vaccine assumed to be 100% effective) the number of infected nodes drops significantly, for all but the denser of the two random networks, as shown in Figure 6. If just 10% of nodes are given targeted vaccination, then there is a drop from an average of 87.7% nodes infected to an average of 45.7%. For the dense random network (an unrealistic topology), the drop is from 99.7% to 89.4%.

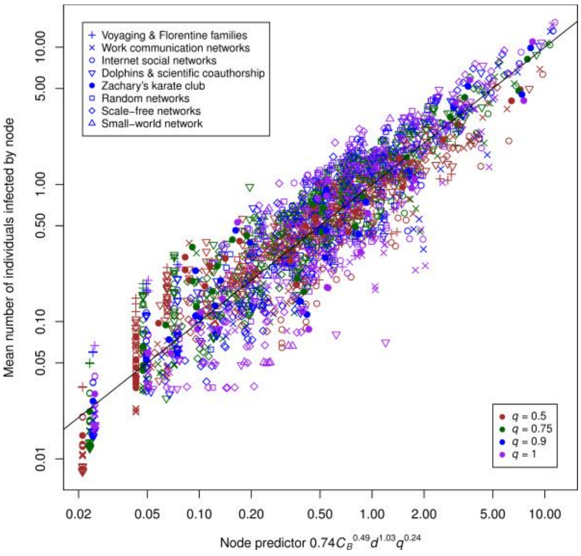

Figure 5. The term 0.74 $C_B^{0.49}$ $d^{1.03}$ $q^{0.24}$ as a predictor of nodes infected, for $0.5 \leq q \leq 1$. Here R^2 = 87.7%.

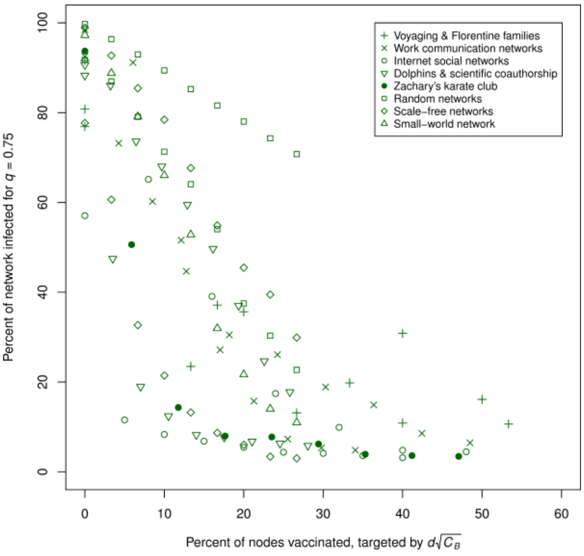

Figure 6. The result of using the product $d\sqrt{C_B}$ to target nodes for vaccination, with q = 0.75.

5. LOWER INFECTION PROBABILITIES

In our second experiment, we turn to diseases with lower infection probabilities, as in Kitsak *et al.* (2010). Specifically, we take $q = 0.05$.

As shown in the third column of Table 2, the results in this case are quite different. Betweenness C_B is now a poor predictor of the number of nodes infected. The node degree d is now the best predictor, with valued centrality C_V the next best. There is little benefit in two-variable models. For example, combining degree and betweenness gives $R^2 = 79.2\%$, almost identical to degree on its own. Figure 7 illustrates the performance of node degree as a predictor. Eigenvector centrality C_E, on the other hand, is a good predictor for some networks, but a very poor predictor for others, making it the worst predictor overall.

The poor performance of eigenvector centrality C_E, in contrast to the findings of Macdonald *et al.* (2012), results from our use of a varied sample of networks. Our sample contains some networks for which eigenvector centrality C_E is utterly useless. In particular, Figure 8 shows that for one network, a 57-node connected subset of a scientific coauthorship network (Newman, 2006), eigenvector centrality C_E does not act as a centrality measure at all, but simply highlights one densely connected subset. As a result of behaviour like this, eigenvector centrality C_E is not in general epidemiologically useful.

The reason that the node degree d is more useful with low infection probabilities is that the probability of a node being infected at all will depend on its degree.

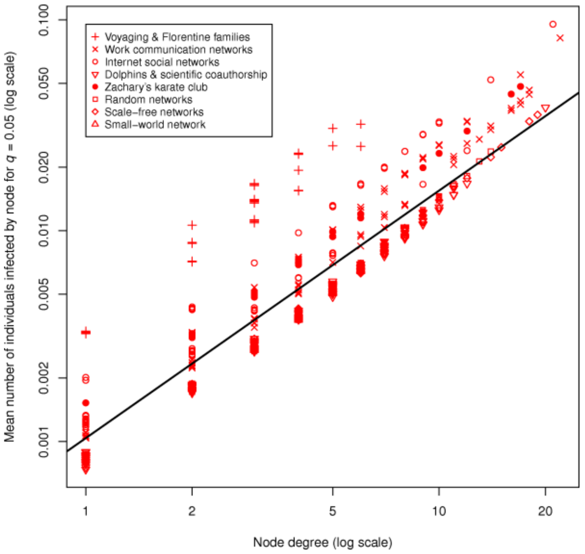

Figure 7. Node degree d as a predictor of nodes infected, for $q = 0.05$. The R^2 value here is 79.1%.

6. DISCUSSION AND CONCLUSIONS

Our two simulation experiments on a mixed sample of networks have explored the phenomenon of "super-spreaders" in communicating infectious disease. For highly infectious diseases, with the probability q of spreading the infection along a link ranging from 0.5 to 1.0, most individuals in the social network are eventually infected with the disease. In this case, the product $0.74\ C_B^{0.49}\ d^{1.03}\ q^{0.24}$ is the best predictor of the number of other people a given node will infect. This predictor has an R^2 value of 87.7% (a correlation of 0.94).

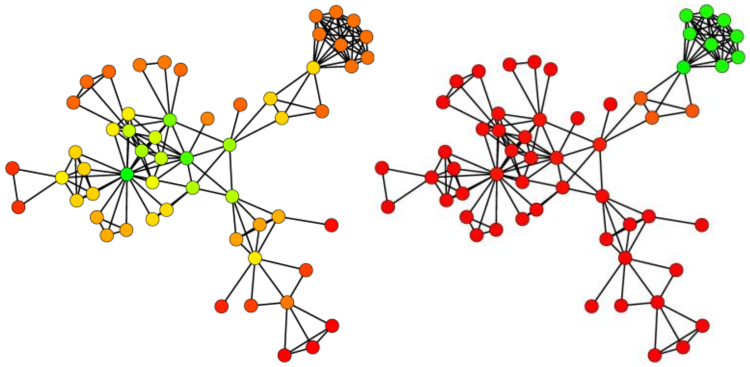

Figure 8. Comparison of valued centrality C_V (left) and eigenvector centrality C_E (right) for a 57-node connected subset of a scientific coauthorship network (Newman, 2006). Values are shown as a red-green spectrum. Here C_E simply highlights one subset at the top right.

This predictor is of more than academic interest, since the measure $d\sqrt{C_B}$ can be used to target nodes for vaccination. Such targeted vaccination can significantly reduce the total number of people who are infected by the disease. Even if we cannot directly calculate this product, we can target nodes of higher-than-average betweenness and degree using the strategy of randomly selecting people and then vaccinating one of their contacts (Cohen *et al.*, 2003). A third experiment showed that the probability of being selected this way has an 0.86 correlation with $d\sqrt{C_B}$.

For diseases which are far less infectious ($q = 0.05$), we found node degree to be the best predictor of the number of other people a given node will infect. In contrast to the findings of Macdonald *et al.* (2012), eigenvector centrality was a poor predictor in both experiments. This is due to the fact that the recursive definition of eigenvector centrality makes it essentially a measure of prestige, and sometimes simply highlights a densely connected subset of the network, rather than acting as a true centrality measure.

REFERENCES

Albert, R. and Barabási, A.-L. (2002). Statistical Mechanics of Complex Networks, *Reviews of Modern Physics*, **74**: 47–97.

Anderson, R.M. and May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press.

Barabási, A.-L. (2002). *Linked: The New Science of Networks*, Perseus Publishing, Cambridge, MA.

Bollobás, B. (2001). *Random Graphs*, 2nd edition, Cambridge University Press, Cambridge, UK.

CDC (2003). Severe Acute Respiratory Syndrome – Singapore, 2003, *Morbidity and Mortality Weekly Report*, **52** (18): 405–411, 9 May 2003, www.cdc.gov/MMWR/preview/mmwrhtml/mm5218a1.htm

Christakis, N.A. and Fowler, J.H. (2010). Social Network Sensors for Early Detection of Contagious Outbreaks, *PLoS One*, **5** (9): e12948, www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0012948

Cohen, R., Havlin, S., and ben-Avraham, D. (2003). Efficient Immunization Strategies for Computer Networks and Populations, *Phys. Rev. Lett.* **91**: 247901.

Dekker, A.H. (2005), Conceptual Distance in Social Network Analysis, *Journal of Social Structure*, **6** (3), www.cmu.edu/joss/content/articles/volume6/dekker/

Dekker, A. (2007). Studying Organisational Topology with Simple Computational Models, *Journal of Artificial Societies and Social Simulation*, **10** (4), article 6, jasss.soc.surrey.ac.uk/10/4/6.html

Dekker, A.H. (2008a). Network Effects in Epidemiology, *Proceedings of SimTecT 2008* (May 12–15, Melbourne), 39–44, Simulation Industry Association of Australia, ISBN: 0-9775257-4-0.

Dekker, A.H. (2008b). Centrality in Social Networks: Theoretical and Simulation Approaches, *Proceedings of SimTecT 2008* (May 12–15, Melbourne), 33–38, Simulation Industry Association of Australia, ISBN: 0-9775257-4-0.

Dekker, A.H. (2010). Average Distance as a Predictor of Synchronisability in Networks of Coupled Oscillators, *Proceedings of ACSC 2010, the 33rd Australasian Computer Science Conference*, January 18–21, Brisbane (*Conferences in Research and Practice in Information Technology*, **102**), 127–131, crpit.com/confpapers/CRPITV102Dekker.pdf

DMERI SARS Investigation Team (2005). Strategies Adopted and Lessons Learnt During the Severe Acute Respiratory Syndrome Crisis in Singapore, *Reviews in Medical Virology*, **15** (1): 57–70.

Gibbons, A. (1985). *Algorithmic Graph Theory*, Cambridge University Press.

Hage, P. and Harary, F. (1995). Eccentricity and Centrality in Networks, *Social Networks*, 17, 57–63.

Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., and Makse, H.A. (2010). Identification of Influential Spreaders in Complex Networks, *Nature Physics*, November, **6**: 888–893, www.nature.com/nphys/journal/v6/n11/abs/nphys1746.html

Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., and Dawson, S.M. (2003). The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-lasting Associations, *Behavioral Ecology and Sociobiology*, **54**: 396–405.

Macdonald, B., Shakarian, P., Howard, N., and Moores, G. (2012). Spreaders in the Network SIR Model: An Empirical Study, arxiv.org/abs/1208.4269

Newman, M.E.J. (2006). Finding Community Structure in Networks Using the Eigenvectors of Matrices, *Phys. Rev. E*, **74**: 036104.

Skvortsov, A.; Connell, R.; Dawson, P. & Gailis, R. (2007). Epidemic Modelling: Validation of Agent-based Simulation by Using Simple Mathematical Models, *MODSIM 2007 International Congress on Modelling and Simulation*, www.mssanz.org.au/MODSIM07/papers/13_s20/EpidemicModeling_s20_Skvortsov_.pdf

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press.

Watts, D.J. and Strogatz, S.H. (1998). Collective Dynamics of ‘Small World’ Networks, *Nature*, **393**: 440–442.

Watts, D.J. (2003). *Six Degrees: The Science of a Connected Age*, William Heinemann, London.

Zachary, W.W. (1977). An Information Flow Model for Conflict and Fission in Small Groups, *Journal of Anthropological Research*, **33**: 452–473.