

Spreading on networks: a topographic view

Geoffrey S Canright and Kenth Engo-Monsen

Telenor R&D

Abstract. We apply our previously developed method of “topographic” analysis of networks to the problem of epidemic spreading. We consider the simplest form of epidemic spreading, namely the “SI” model. We argue that the *eigenvector centrality* of a node is a good indicator of that node’s *spreading power*. From this we develop seven specific predictions. In particular, we predict that each *region* (as defined by our approach) will have its own S curve for cumulative adoption over time, and we describe the various phases of the S curve in terms of motion of the infection over the region. Our predictions are well supported by simulations. In particular, the significance of regions to epidemic spreading is clear. Finally, we develop a mathematical theory, giving partial support to our picture. The theory includes a precise quantitative definition of the spreading power of a node, and some approximate analytical results for epidemic spreading.

Short title: Spreading on networks

Keywords: epidemic spreading, eigenvector centrality, networks, regions, topography

1 Introduction

The general phenomenon of spreading over a network is ubiquitous. Examples include the spreading of a disease (in which case the network may be a social network, or a computing/communications network exposed to a virus); the spreading of gossip (social network; although here we will ignore the interesting property that gossip gets modified as it spreads); or the spreading of innovation (again a social network). In each of these cases, it is a common simplification to assign to each node only one of two possible states: ‘uninfected’ or ‘infected’. If you are uninfected (‘susceptible’), you are deemed liable to be infected by any infected neighbors. Correspondingly, if you are infected, you remain so for the duration of the experiment—and you remain capable of infecting any or all of your neighbors. Of course, on some appropriate time scale, nodes become ‘immune’ to the infection: a human develops antibodies, a machine gets antivirus software, the gossip becomes boring, or the innovation becomes outmoded. We focus on a shorter time scale here, so that we can ignore the state of acquired immunity. The process under study here may then be equally termed ‘infection’, ‘adoption’, or ‘epidemic spreading’; we will use all of these terms freely. The technical name for our model of spreading is ‘SI’, since the nodes have the two states Susceptible or Infected. .

Since spreading takes place over the links of a network, it is clear that the topology of the network can have a profound influence on the spreading process. In particular, we believe that the best understanding of spreading will come from a perspective which is based on a view of the whole network, and an understanding of that network’s structure. In earlier work [1], we have presented an approach to the analysis of network structure which is applicable to any network with symmetric (undirected) links. We also suggested that the analysis should be useful for the understanding of spreading over such a network. In this work, we elaborate on that suggestion, and then test it, in two ways. First, we report on a series of simulations, carried out on network structures which were obtained from empirically measured undirected networks. These simulations offer strong support to our qualitative picture of spreading. Secondly, we develop a mathematical definition of the spreading power of a node, as well as a mathematical description of the SI epidemic spreading process. While we cannot obtain exact answers to either of these formulations, we do find significant support for our qualitative arguments in the approximate results that we obtain.

Authors’ address: Telenor R&D, B6d, Snarøyveien 30, 1331 Fornebu, Norway

Phone: +47-91815638 (GSC), +47-41560883 (KEM)

Fax: +47-96211086

Email: geoffrey.canright@telenor.com, kenth.engo-monsen@telenor.com

Our approach departs from previous work in that we focus on both the *time* and *spatial progression* of the epidemic spreading. We take a spatial resolution which is not microscopic, but rather at the level of ‘neighborhoods’—connected subgraphs with roughly the same spreading power. More traditional approaches (reviewed in [6]) start from the ‘well-mixed’ approximation, that every node can infect every other with some probability, at all times. This approach may be said to have no network perspective; or, it may be said to postulate a graph with extremely good mixing—such as a random graph of high degree, or a complete graph. The review of Newman [6] also discusses more recent work, involving a network perspective. All such work is based on whole-graph properties, such as the node degree distribution; also, these approaches have focused on obtaining whole-graph results, either over time [7,8], or focusing especially on the infected fraction at very long times [9]. This latter question is of course only interesting for models more complex than the SI model; and indeed most work is directed towards the behavior of the SIS model (where nodes lose their infection after some time, and so become Susceptible again), or the SIR model (where nodes, after losing their infection, go through a refractory period).

Brauer [10] has examined the SI model for the case that the nodes (organisms, especially humans) are born and die. Because of the addition of these dynamic features, the steady infection rate is not necessarily 100%. This work uses the well-mixed approximation, which gives rise to coupled ordinary differential equations.

A work which is perhaps closest to the present work is that of Wang et al [11]. Their model is SIS, in that nodes can be “cured”; but it is based on a fully microscopic view of the network. In fact, their time evolution operator is the same as that we develop in Section 4.3.2, with two differences. One is their addition of the “curing” term. This term is simply a multiple of the unit matrix, and so does not change the dominant eigenvector—which remains that of the adjacency matrix A . Because their model is SIS, the long-time infection fraction is not obvious, and must be solved for. The second difference in the time evolution operator of Wang et al is that they neglect the cross terms—ie, those arising from multiple transmissions to an infected node. This approximation is valid for low infection fraction—while (as we discuss in detail below) it *may* also be good even as the infection fraction becomes large. Wang et al report simulations which offer some support for this statement.

We emphasize that our work, like that of Wang et al [11], uses the full adjacency matrix A in modelling the time evolution of the infection. Thus we start from a microscopic foundation. However, we will quickly appeal to a ‘*mesoscopic*’ picture, in which it is meaningful and useful to speak of neighborhoods and their properties. As far as we know, our work is unique in this regard.

We conclude this introduction with a brief review of the network analysis presented in [1]. Our strategy there was to choose eigenvector centrality as a useful measure of well-connectedness. Eigenvector centrality (EVC) has the desirable property that—since it depends on the properties of the *neighborhood* of a node, and not just of the node itself—it is rather ‘smooth’ over the graph (or network; we use these terms interchangeably). This is in contrast to the related quantity *degree centrality*, which simply counts the links leaving a node and so is completely local. The smoothness of the EVC allows one to think in terms of the ‘topography’ of the graph. That is, if a node has high EVC, its neighborhood (from smoothness) will also have a somewhat high EVC—so that one can imagine EVC as a smoothly varying ‘height’, with mountains, valleys, mountaintops, etc. We define precisely the notion of a *region* of the graph: all those nodes for which a steepest-ascent path terminates at the same local maximum of the EVC belong to the same region. That is, a region is a mountain, and it is defined by its peak. This definition gives a unique decomposition of any graph into one or more regions, with each node (almost always) belonging to a single region. Links connecting regions are also viewed to play a special role in the network; in the topographic analogy, they correspond to ‘saddles’.

In the next section we will give our reasons for believing that this method of analysis is useful for understanding spreading. We emphasize here that the term ‘region’ has, everywhere in this paper, a precise mathematical meaning; hence, to discuss subgraphs which lie together in some looser sense, we use terms such as ‘neighborhood’ or ‘area’.

2 Qualitative arguments

We wish to evaluate the nodes in a network in terms of their “spreading power”. That is, we know that some nodes play an important role in spreading, while others play a less important role. One need only imagine the extreme case of a star: the center of the star is absolutely crucial for spreading of infection over the star; while the leaf nodes are entirely unimportant, having only the one aspect (common to every node in any network) that they can be infected.

Clearly, the case of the star topology has an obvious answer to the question of which nodes have an important role in spreading (have high spreading power). The question is then, how to generate equally meaningful answers for general and complex topologies, for which the answer is not at all obvious? In this section we will propose and develop a qualitative answer to this question.

Our basic assumption (A) is simple, and may be expressed in a single sentence:

Eigenvector centrality (EVC) is a good measure of spreading power. (A)

We will test this idea via simulations (Section 3) and some theory (Section 4). First however we will explore the implications of this idea—coupling it to our earlier work, which uses EVC to define well-connected regions of any given network.

First we recall the definition of, and motivation for, EVC. A simple definition of centrality, which is certainly related to spreading power, is a node’s degree centrality, ie, its number of neighbors:

$$k_i = \sum_{j=nn(i)} 1, \quad (1)$$

where $nn(i)$ means “nearest neighbors of node i ”. This quantity is however not smooth at all—there is no necessary correlation between a given node’s degree and that of its neighbors.

Seeking a more smooth measure, we can give each node a centrality score which is simply the average of its neighbors’ scores:

$$x_i = \frac{1}{k_i} \sum_{j=nn(i)} x_j. \quad (2)$$

As shown in [1], this choice gives a node importance weight which is *too* smooth: normal solutions to (2) give equal scores x_i to every node i . Hence we discard this idea.

Eigenvector centrality involves one small change from (2): one defines the centrality of node i as being proportional to the sum of (but not the average of) i ’s neighbors’ centralities:

$$e_i = \frac{1}{\lambda} \sum_{j=nn(i)} e_j. \quad (3)$$

Eq. (3), in contrast to (1), makes a node’s centrality dependent on its neighbors’ centralities—hence giving a ‘smooth’ centrality measure—but it also gives nontrivial solutions [unlike (2)]—because, since λ is the same for all nodes, Eq. (3) does not completely cancel out the boost in centrality from having many neighbors [as (2) does]. Eq. (3) can be rewritten as

$$Ae = \lambda e, \quad (4)$$

where e is the vector of centrality scores, and A is the network’s adjacency matrix. Eq. (4) reveals the motivation for the name ‘eigenvector centrality’: the EVC of e_i of node i is the i ’th component of a chosen eigenvector e of the adjacency matrix A . To ensure that all centrality scores are positive, one takes the *principal* eigenvector of A —that is, the one corresponding to the largest eigenvalue λ_{\max} .

Thus we see that, because a node’s EVC depends on that of its neighbors, the EVC values over a network may be thought of as ‘smoothly varying’ over the network. That is, a node with very high EVC cannot be surrounded by nodes with very low EVC. Of course, it is true that EVC tends to be positively correlated with a simpler measure of centrality, namely node degree. In fact, one might say that the principal difference between the two measures is that EVC is constrained by its definition to

be smooth, while node degree centrality is not. This difference can however be nontrivial. For instance, a node with high degree, surrounded by many leaf nodes, and linked only tenuously to the bulk of a large and well-connected network, will have a low EVC, in spite of its high degree. The point is that EVC is sensitive to properties of neighborhoods, while node degree is not.

Thus, in short, there are no isolated nodes with high EVC. That is, a node with high EVC is embedded in a neighborhood with high EVC. (There can however be relatively isolated nodes with *low* EVC, as this situation is self-consistent.) Now if we take our basic assumption A to be true, then there are no isolated nodes with high spreading power. Instead, there are neighborhoods with high spreading power. (But there can be isolated nodes with low spreading power.)

We then suppose that an infection has reached a node with modest spreading power. Suppose further that this node is not a local maximum of EVC; instead, it will have a neighbor or neighbors of even higher spreading power. The same comment applies to these neighbors, until one reaches the local maximum of EVC/spreading power.

Now, given that there are neighborhoods, we can discuss spreading in terms of neighborhoods rather than in terms of single nodes. It follows from the meaning of spreading power that a neighborhood characterized by high spreading power will have more rapid spreading than one characterized by low spreading power. Furthermore, we note that these different types of neighborhoods (high and low) are smoothly joined by areas of intermediate spreading power (and speed).

It follows from all this that, if an infection starts in a neighbourhood of low spreading power, it will tend to spread to a neighbourhood of higher spreading power. That is: spreading is faster *towards* neighborhoods of high spreading power, because spreading is faster *in* such neighborhoods. Then, upon reaching neighbourhood of the nearest local maximum of spreading power, the infection rate will also reach a maximum (with respect to time). Finally, as the high neighborhood saturates, the infection moves back ‘downhill’, spreading out in all ‘directions’ from the nearly saturated high neighborhood, and saturating low neighborhoods.

We note that this discussion fits naturally with our earlier work, which describes network topology in topographic terms. That is, the smoothness property of EVC allows one to think of a smoothly varying ‘height’ for nodes—so that one can define mountains, their peaks, valleys, etc. Putting the previous paragraph in this language, then, we get the following: infection of a hillside will tend to move uphill, while the infection rate grows with height. The top of the mountain, once reached, is rapidly infected; and the infected top then efficiently infects all of the remaining adjoining hillsides. Finally, and at a lower rate, the foot of the mountain is saturated.

We see now that this qualitative picture addresses nicely the various stages of the classic S curve of innovation diffusion [12]. The early, flat part of the S is the early infection of a low area; during this period, the infection moves uphill, but slowly. The S curve begins to take off as the infection reaches the higher part of the mountain. Then there is a period of rapid growth while the top of the mountain is saturated, along with the neighboring hillsides. Finally, the infection rate slows down again, as the remaining uninfected low-lying areas become infected.

One might object that this picture is too simple, in the following sense. Our picture gives an S curve *for a single mountain*. Yet we know from earlier analyses that a network is often composed of several regions (mountains). The question is then, why should such multi-region networks exhibit a single S curve?

Our answer here is that such networks need not necessarily exhibit a single S curve. That is, our arguments predict that each region—defined around a local maximum of the EVC—will have a single S curve. Then—assuming that each node belongs to a single region, as occurs with our preferred rule for region membership—the cumulative adoption/infection curve for the whole network is simply the sum of the adoption curves for each region. These latter single-region curves will be S curves. Thus, depending on the relative timing of these various single-region curves, the network as a whole may, or may not, exhibit a single S curve. For example, if the initial infection is from a peripheral node which is close only to one region, then that region may take off well before neighboring regions. On the other hand, if the initial infection is in a valley which adjoins several mountains, then they may all exhibit

takeoff roughly simultaneously—with the result being a sum of roughly synchronized S curves, hence a single S curve.

Let us now summarize and enumerate the predictions we take from this qualitative picture.

- a. Each region has an S curve.
- b. The number of takeoff/plateau occurrences in the cumulative curve for the whole network may be more than one; but it will not be more than the number of regions in the network.
- c. For each region—assuming (which will be typical) that the initial infection is not a very central node—growth will at first be slow.
- d. For each region (same assumption) initial growth will be towards higher EVC.
- e. For each region, when the infection reaches the neighborhood of high centrality, growth “takes off”.
- f. An observable consequence of (e) is then that, for each region, the *most* central node will be infected at, or after, the takeoff—but not before.
- g. For each region, the final stage of growth (saturation) will be characterized by low centrality.

3 Simulations

We have run simple simulations to test our qualitative picture. As noted above, we have implemented an SI model. That is, each node is Susceptible or Infected. Once infected, it remains so, and retains the ability to infect its neighbors indefinitely. We have used a variety of sample networks, extracted from data obtained in previous studies. Results reported here will be taken from simulations performed on: (i) seven distinct snapshots [2] of the Gnutella peer-to-peer file-sharing network, taken in late 2001; (ii) a social network of students at the Høgskolen i Oslo, measured in 2004 in the course of a study [3] on the spreading of innovation; (iii) a snapshot of the collaboration graph of our own R&D department; (iv) a snapshot of the collaboration graph for the Santa Fe Institute [4]. We will denote these graphs as (i) g_1 — g_7 ; (ii) hio ; (iii) fou (Norwegian for R&D); and (iv) sfi .

Our procedure is as follows. We initially infect a single node. Each link ij in the graph is assumed to be symmetric (consistent with our use of EVC), and to have a constant probability p , per unit time, of transmitting the infection from node i to node j (or from j to i), whenever the former is infected and the latter is not. (We continue to use the terms ‘infection’ and ‘adoption’ interchangeably.) All simulations were run to the point of 100% saturation. Thus, the ultimate x and y coordinates of each cumulative infection curve give, respectively, the time needed to 100% infection, and the number of nodes in the graph.

3.1 Gnutella graphs

The Gnutella graphs, like many self-organized graphs, have a power-law node degree distribution, and are thus well connected. Consistent with this, our analysis returned either one or two regions for each of the seven snapshots. We discuss these two cases (one or two regions) in turn. Each snapshot is taken from [2], and has a number of nodes on the order of 900—1000.

3.1.1 Single region

The snapshot termed ‘ g_3 ’ was found by our analysis to consist of a single region.

Figure 1 shows a typical result for the graph g_3 , with link infection probability $p = 0.05$. The upper part of the figure shows the cumulative S curve of adoption, with the most central node becoming infected near the ‘knee’ of the S curve. The lower part of Figure 1 shows a quantity termed μ —that is, the average EVC value for newly infected nodes at each time. Our qualitative arguments say that this quantity should first grow, and subsequently fall off, and that the main peak of μ should coincide with the greatest growth in the S curve. We see, from comparing the two parts of Figure 1, that the most central nodes are infected roughly between time 2 and time 20—coinciding with the period of

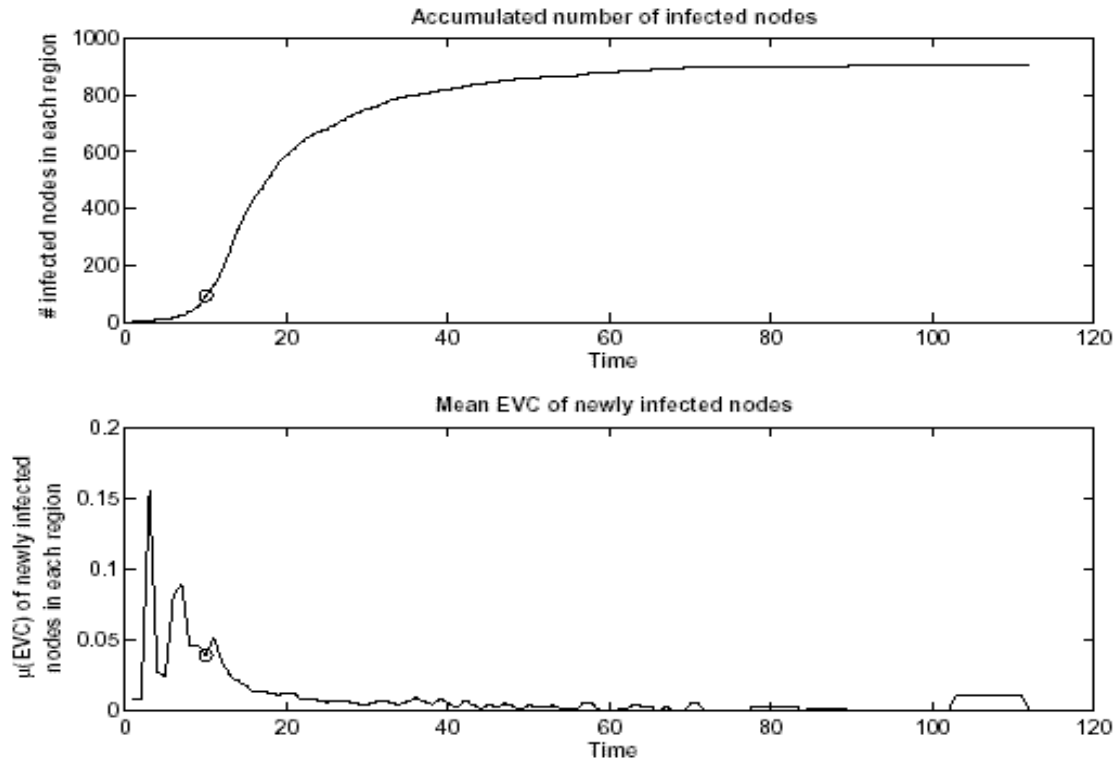


Figure 1. (top) Cumulative adoption for the Gnutella graph g_3 . The circle marks the time at which the most central node is infected. (bottom) Average EVC ('height') of newly infected nodes at each time step.

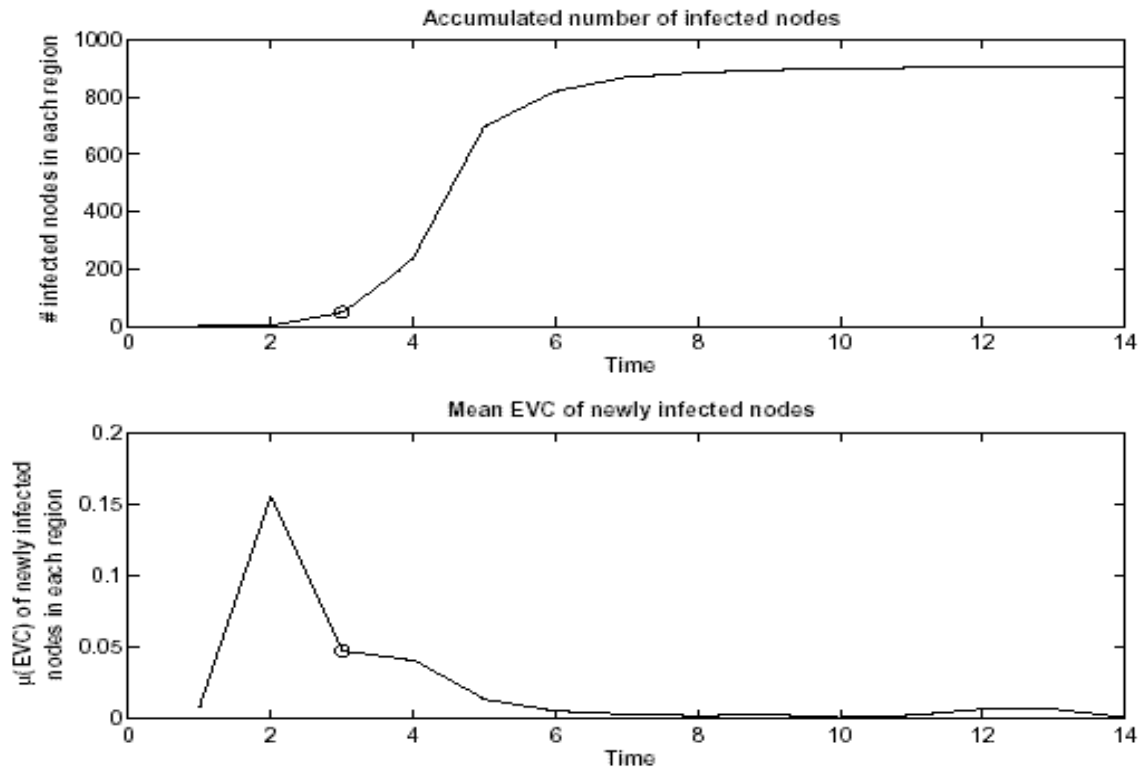


Figure 2. Same simulation as Figure 1, except for: (i) $p = 0.6$, and (ii) new random numbers determining the infection events.

maximum growth in the S curve. Thus, this figure supports all of our predictions a—g above, with the minor exception that there are some fluctuations superimposed on the growth and subsequent fall of μ over time.

It is interesting to note that this picture is rather insensitive to the probability parameter p . For example, Figure 2 shows results for the same graph and same initial node, with the probability now $p = 0.6$. We see that the main effect of this much higher p is simple, and offers few surprises: the time scale is of course much compressed, with the expected result that the cumulative curve is less smooth. We note that even the extreme case of setting $p = 1$ gives a picture very much like that of Figure 2.

3.1.2 Two regions

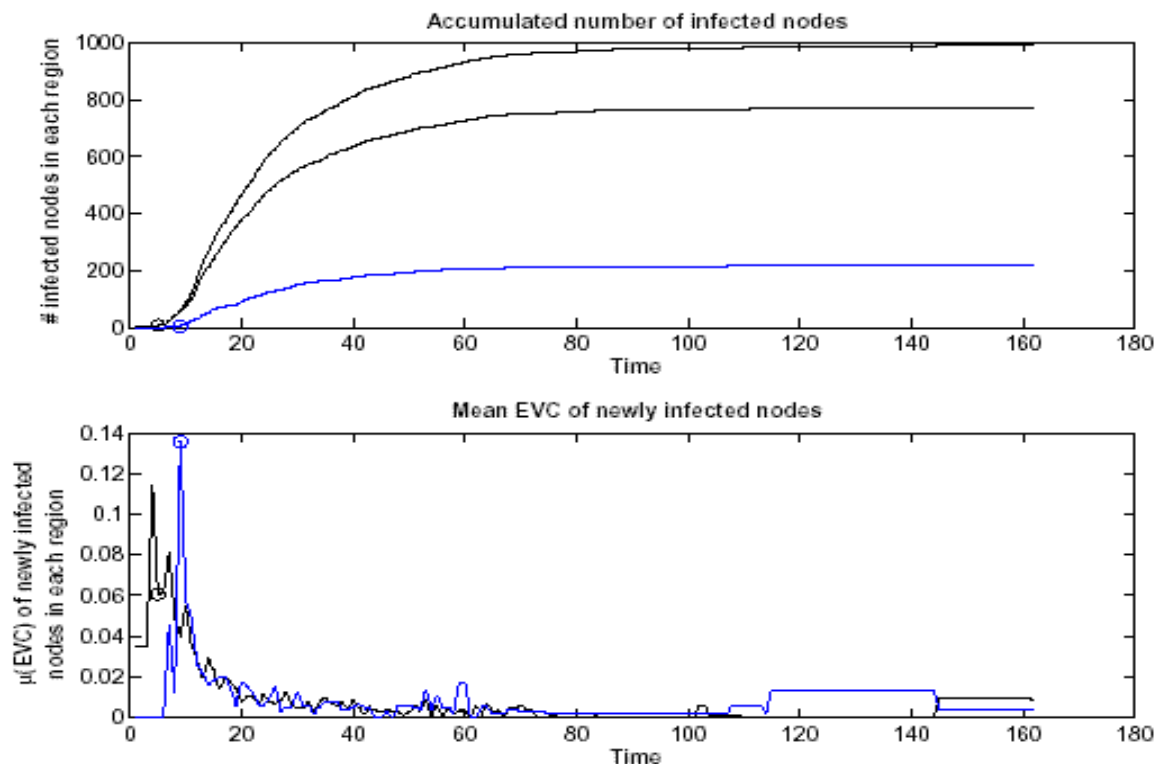


Figure 3. (top) Cumulative infection curves, and (bottom) μ curves, for a two-region graph g_1 , with $p = 0.04$. The upper plot shows results for each region (black and blue), plus their sum (also black). The lower plot gives μ curves for each region.

Figure 3 shows typical behavior for the graph g_1 , which consists of two regions. We see that the two regions go through takeoff roughly simultaneously. The result is that the sum of the regional infection curves is a single S curve. We also see that each region behaves essentially the same as did the single-region graph g_3 . For instance, the new centrality (μ) curves for each region first rise, and then fall, with their main peak (before time 20) coinciding with the period of most rapid growth for the respective regions (and for the whole graph). These results thus add further support to our predictions a—g.

3.2 Student network, HiO

For this graph we found a largest connected component (LCC) consisting of 249 nodes [3]. We call this LCC the ‘hio graph’. Our analysis found three regions for the hio graph. However, the spreading behavior on this graph is qualitatively like that for the g_1 graph: the three regions tend to take off roughly simultaneously. The effect is also the same: the cumulative infection curve for the entire graph is a clean S curve, composed of the sum of three smaller clean S curves. Also, each region behaves essentially as seen for the graph g_3 .

3.3 Telenor R&D

We have formed a collaboration graph for the researchers working at Telenor R&D. For this graph, we analyzed the largest connected component, consisting of 137 nodes. Our analysis gave a single region for this graph. Spreading behavior was much like that seen in Figure 1. One difference is that the S curve is less smooth—an effect of the small size of the fou graph. Another difference is that, for many simulations (but not all), the time period of the rise of the S is disproportionately large compared to the time needed for 100% of saturation. Also, in such cases, the time of infection of the most central node tends to fall rather late after the onset of the rise (the ‘knee’ of the S). We show an example of this behavior in Figure 4.

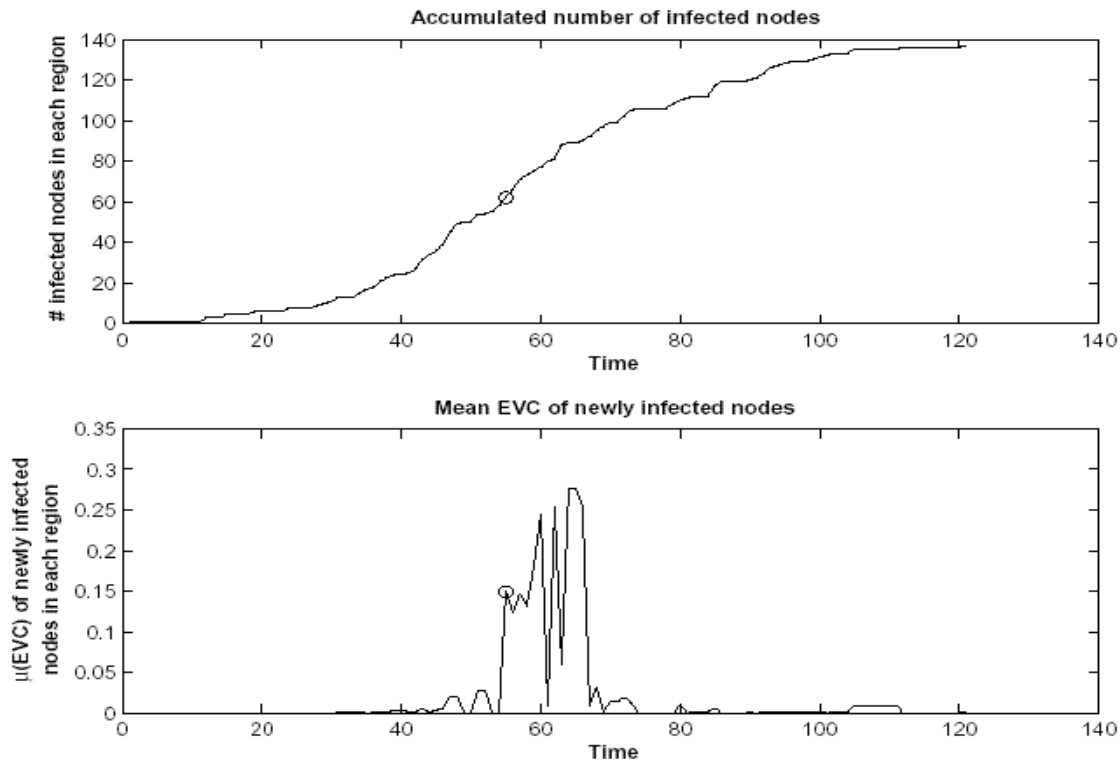


Figure 4. Spreading for the single-region fou graph, with $p = 0.04$. Note that the S curve is somewhat ‘flat’ on a time scale of 100% saturation.

The differences between Figures 1 and 4 are interesting; however both figures are fully consistent with our predictions a—g. We note also an interesting correlation here, which is not surprising: steeper S curves tend to be associated with earlier infection times (relative to the knee) for the most central node.

3.4 Santa Fe Institute

The sfi graph also gives three regions under our analysis. The spreading behavior varied considerably on this graph, depending both on the initially infected node, and on the stochastic outcomes for repeated trials with the same starting node.

Figure 5 shows an untypical case for this graph. The aspect that is untypical here is that the whole-graph cumulative infection curve resembles (somewhat) a single, smooth, S curve. This result was obtained however for the rather artificial starting condition that the first infected node was the most central node in the largest region—hence in the entire graph. The result is that this region takes off immediately. (Our prediction c thus does not hold; but its *assumption* has been violated by our infecting the most central node first.) The next largest region is however infected fairly soon thereafter, so that its takeoff is not clearly seen in the total curve. Finally, the third region takes off

considerably later. But, because it is small, and its takeoff occurs before the blue region is fully saturated, the takeoff of this third region is also not clear from inspection of the total infection curve.

From observing the μ curves, we see that that of the blue region is as predicted. The red region is similar but not visible on the scale of the figure. The largest (black) region's μ curve lacks the initial rise in centrality—but this is to be expected, as the infection began at the top.

Now we move to a more typical case for the sfi graph. Figure 6 shows the behavior for the same infection probability, but with a randomly chosen start node.

The interesting feature of this simulation is that the cumulative curve shows very clearly two takeoffs, and two plateaux. That is, it resembles strongly the sum of two S curves. And yet it is easy to see how this comes about, from our region decomposition: the blue and red curves take off roughly simultaneously, while the largest (black) region takes off only after the other regions are saturated.

The μ curves show roughly the expected behavior—the qualification being that they are rather noisy. Nevertheless the main peak of each μ curve corresponds to the main rise of the corresponding region's S curve.

We reiterate that the behavior seen in Figure 6 is much more typical for this graph than that seen in Figure 5.

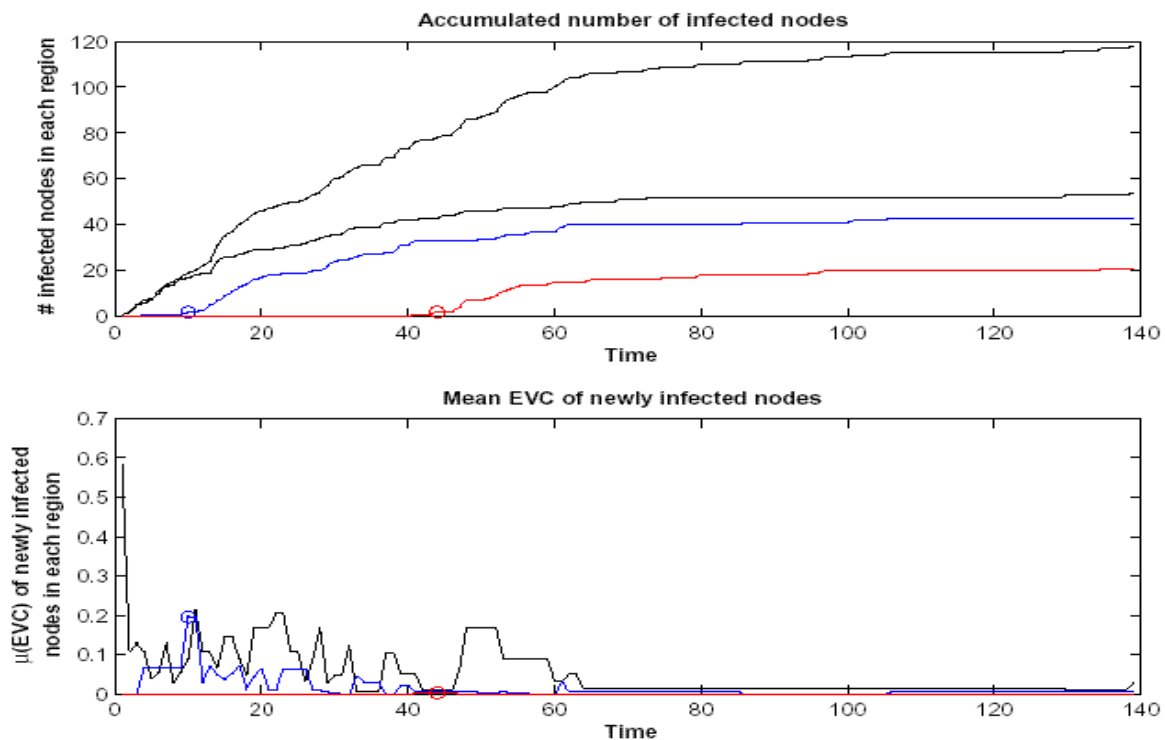


Figure 5. Spreading on the sfi graph, $p = 0.04$. The start node is the most central node in the largest of the three regions (black); hence its infection time is not marked.

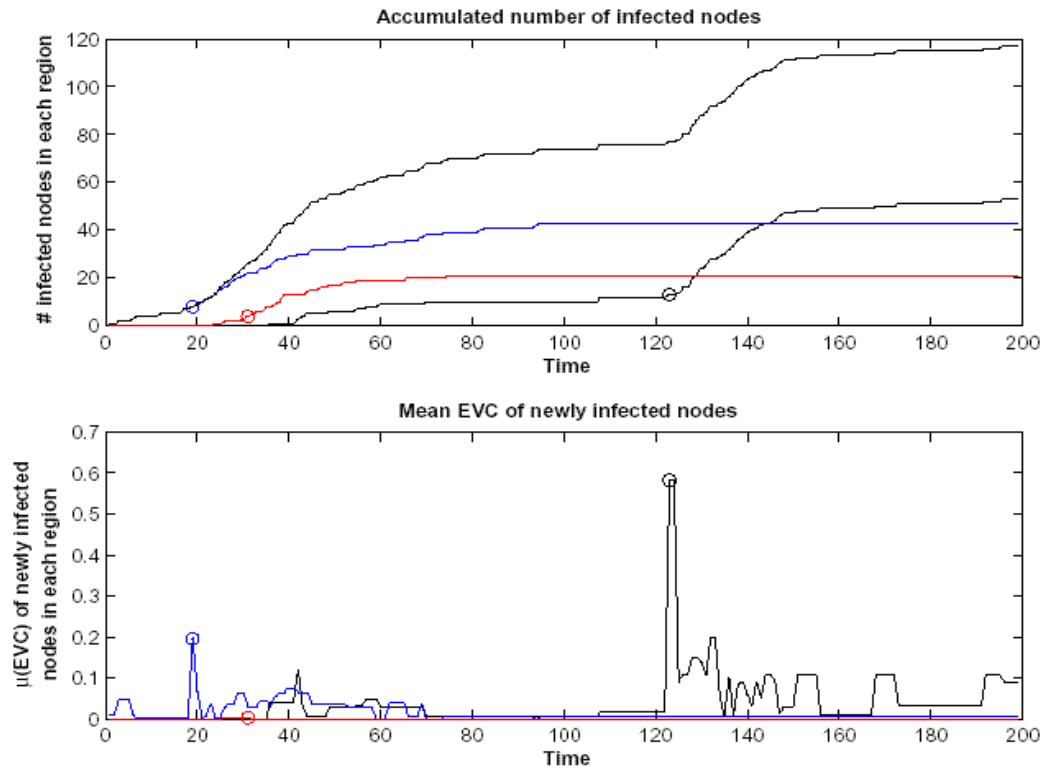


Figure 6. Same as Figure 5, except a randomly chosen start node.

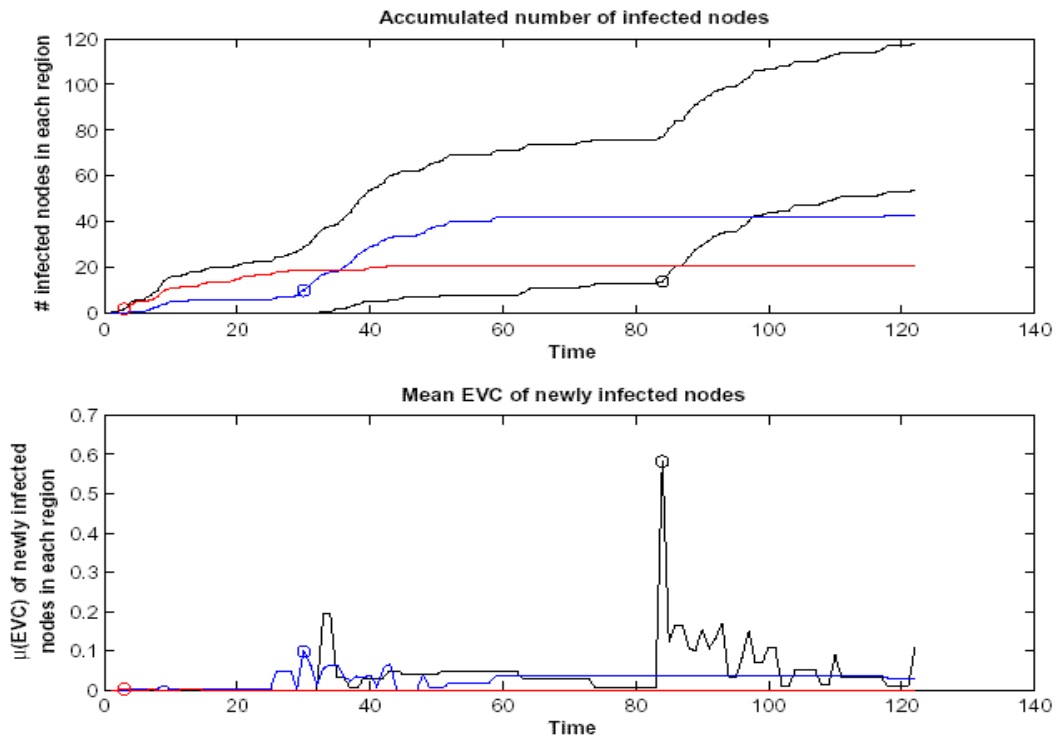


Figure 7. Spreading on the sfi graph, $p = 0.05$. In this case, the existence of three regions in the graph is clearly seen in the total cumulative infection curve.

We examine yet one more example from the sfi graph. Figure 7 shows a simulation with a different start node from Figures 6 or 5, and with $p = 5\%$. The message from Figure 7 is clear: the cumulative adoption curve shows clearly three distinct S curves—takeoff followed by plateau—one after the other. It is also clear from our regional adoption curves that each region is responsible for one of the S curves in the total adoption curve: the smallest (red) region takes off first, followed by the blue region, and finally the largest (black) region. In each case, the time of infection for the most central node of the region undergoing takeoff lies very close to the knee of the takeoff. And, in each case, the peak of the μ curve coincides roughly with the knee of the takeoff.

We note that the behavior seen in Figure 7 is neither very rare nor very common. The most common behavior, from over 50 simulations with this graph, is most like that seen in Figure 6; but we have seen behavior which is intermediate to that in Figures 5 and 6 (ie, neither clearly one nor clearly two takeoffs), and also behavior which is intermediate to that of Figures 6 and 7. In particular, multiple simulations with the same start node and probability as that for Figure 7 have yielded two clear takeoffs (as in Figure 6), three clear takeoffs (as in Figure 7), and intermediate behavior. It is interesting to note that the behavior for this start node, with $p = 1$ (hence with deterministic behavior), gives a cumulative adoption curve which is best described as showing between two and three takeoffs.

Thus the behavior of spreading on the sfi graph gives the strongest confirmation yet of our prediction b: that, for a graph with r regions, one may see up to r , but not more than r , distinct S curves in the total cumulative adoption curve. All of our observations, on the various graphs, are consistent with our predictions; but it is only in the sfi graph that we clearly see all of the (multiple) regions found by our analysis to be present in the graph.

One very coarse measure of the well-connectedness of a network is the number of regions in it, with a high degree of connectivity corresponding to a small number of regions, and with a large number of regions implying poor connectivity. By this very coarse criterion, the hio graph and the sfi graph are equally well connected. Based on our spreading simulations, reported here, we would say that in fact the sfi graph is less well connected than the hio graph. We base this statement on the observation that, for the hio graph, we never found cases where the different regions took off at widely different times—the three regions are better connected to one another than is the case for the sfi graph.

Examination of the sfi graph itself (Figure 8) renders this conclusion rather obvious: the three regions are in fact connected in a linear chain. Thus, it is not surprising that we can, in some simulation runs, clearly see the takeoff of each region, well separated in time from that of the other regions. The hio graph, in contrast, has many links between each pair of regions of the three. We note finally in this context that the Gnutella graphs are very well connected; those that resolve to two regions always have numerous links between the two.

Finally, we note that the sfi graph, while clearly conforming to our assertion b (we see up to three regions), does not wholly agree with assertion a (that each region will have a clear S curve). For instance, in Figure 6, the black region has a small ‘premature’ takeoff around time 40, giving a visible plateau before the ‘main’ takeoff after time 120. Similarly, both the blue and the black regions show such ‘premature’ takeoffs in Figure 7. These observations are not exceptional: the blue region in particular is prone to such behavior, showing two stages of growth in over half of the simulations we have run on the sfi graph. This weakens the support for assertion a; we would then say that a single S curve for a single region is a rule that is followed in most cases, but not all. However these two-stage takeoffs do not contradict the other predictions on our list. For instance, there is always a corresponding ‘premature hump’ in the μ curve for the region in question.

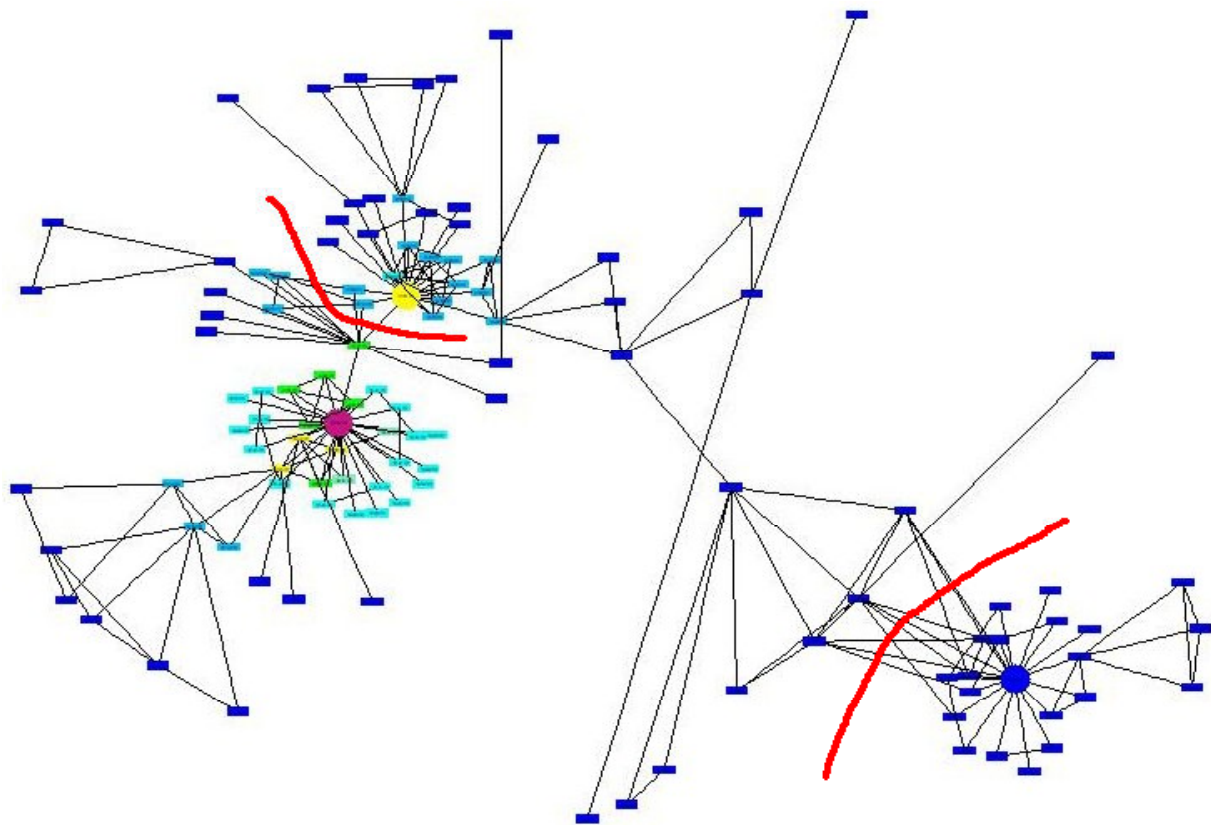


Figure 8. A visualization of the collaboration graph for the Santa Fe Institute. Colors of nodes code for eigenvector centrality, with warm colors implying higher EVC. Also, the most central node in each region is denoted with a round circle. Thick red lines show boundaries between regions. The leftmost region (with a magenta center) is the largest, and has black curves in Figures 5—7; the rightmost region is the smallest (red curves in Figures 5—7); and the middle region has blue curves in Figures 5—7. Hence, in the text, we refer to these three regions, respectively, as the ‘black region’, the ‘blue region’, and the ‘red region’.

It is clear from Figure 8 where the ‘premature takeoffs’ arise. The black region includes a subcluster which is easily identified visually in Figure 8 (close to the next region). This subcluster is connected to the remainder of the black region by a single link. These same statements hold true for the blue region—it has a subcluster, joined to it by a single link, but close to the rightmost region. Hence, in Figure 7, for example, the infection starts on the ‘far side’ of the red region. Then it spreads to the adjacent subcluster of the blue region, but does not reach the most central nodes of that region until after about 20 more units of time (about $1/p$). Soon after the center of the blue region is infected, the subcluster of the black region begins to be infected; but again there is a delay (about twice $1/p$) before reaching the main center of the black region.

We have explored other rules for assigning nodes to regions [1]. For instance, if we simply use shortest distance rather than steepest ascent as our criterion, then the two abovementioned subclusters move: the blue subcluster moves to the red region, and the black subcluster becomes a set of ‘border nodes’ equidistant from both centers. We have not used the shortest distance rule in this paper, because it tends to place too many nodes in general in border regions. Possibly, one could obtain cleaner S curves for the three sfi regions using the shortest distance rule. However, this rule for defining regions ignores the view that the EVC should be viewed as a height function; hence it is less topographically motivated, and so may also give poorer results in many cases.

4 Towards a quantitative theory

In this section, we will outline a quantitative theory expressing our ideas. We use the word “outline” as we do not obtain exact analytical results for either our definition of node spreading power, or for our dynamic equations for the SI spreading process. Hence we leave the further development of this theoretical outline to future work.

4.1 Definitions

We return to our basic question: how can we quantify the “spreading power” of nodes in a network? If we can do this, then we can test our basic assumption: that eigenvector centrality is a good measure of spreading power.

We propose first to examine nodes pairwise. Consider the pair ij . We seek an “infection coefficient” $C(i,j)$ which quantifies the ease with which an infection may be spread from i to j (or vice versa, since the links are symmetric). Clearly, $C(i,j)$ should be large if there are many short paths from i to j . That is: we do not wish only to consider shortest paths; as long as $p < 1$, infection can spread from i to j via any of the many paths from i to j . However, longer paths should receive a lower weight in $C(i,j)$.

Next we note that the (i,j) element of the matrix A^h contains an integer giving the number of paths of length h between nodes i and j (where A is the adjacency matrix). Hence, the collection of matrix elements $A^h(i,j)$ appear to offer the information we need to evaluate $C(i,j)$. There is however one difficulty with using the elements $A^h(i,j)$: they include paths which are self-retracing. Our aim in defining $C(i,j)$ is to include paths which might be travelled by an infection moving from i to j (or the reverse). We do not believe it makes sense to include self-retracing paths in computing $C(i,j)$: for one thing, such paths include paths in which j is infected from i , and then infected again. Furthermore, since $C(i,j)$ must be symmetric, we also exclude paths which revisit the start node before finally reaching the end node. Finally, we generalize this logic to claim that the definition of $C(i,j)$ should also exclude paths which revisit *any* intermediate node—hence *all* self-retracing paths should be excluded. (As we see below, self-retracing paths have zero effect on the evolution of the spreading.)

Thus, for each path length h , we do not want to use the matrix element $A^h(i,j)$ to count possible infecting paths. Instead, we want the quantity $NSR^h(i,j)$, which is simply a notation for the quantity $A^h(i,j)$ corrected to include only non-self-retracing paths, of length h , between i and j . Here it is important to clarify a question regarding our notation: for A^h , the ‘ h ’ is in fact an (integer) exponent; while for the matrix NSR^h , the ‘ h ’ is only a superscript—we do *not* claim that NSR^h is simply the h ’th power of NSR^1 .

In fact, we know of no general method for computing $NSR^h(i,j)$ for all h . Clearly, $NSR^1 = A$; and $NSR^2 = A^2 - \text{Diag}(A^2)$. However we know of no general expression for larger h .

Given an expression for NSR^h , we build $C(i,j)$ as follows. The sum

$$\sum_{h=1}^{\max} NSR^h(i,j)$$

is an integer, counting the total number of distinct, non-self-retracing paths between i and j . (Note that, given the restriction to non-self-retracing paths, there will be an upper limit “max” to the sum over number of hops h .) However, this sum gives equal weight to paths of all lengths. Instead we want a weight function $f(h)$ in the sum which gives diminishing weight to paths of increasing length h . At this point we do not claim to have a conclusive argument which can determine the choice of weight function $f(h)$ —other than that it be monotonically decreasing with h . The extreme case [$f(1) = 1$, with $f(h) = 0$ for $h > 1$] is as uninteresting as the equal-weight case, since it only gives weight to one-hop paths, and so gives the result that a node’s spreading power is simply its degree.

For simplicity, we set $f(h) = w^h$, where w is some positive weight which is less than one. One may argue that this choice best reflects the case in which each link has a probability $p < 1$ (per unit time) of transmitting the infection. This gives the following expression for the infection coefficient:

$$C(i, j) = \sum_{h=1}^{\max} w^h NSR^h(i, j) . \quad (5)$$

We take (5) as our working definition of infection coefficient. Now we wish to use this definition in order to obtain a definition of spreading power for a node. The idea is again simple: If a node has a high infection coefficient with respect to many other nodes, then it has a high spreading power. And in this case we see no need for a decreasing weight function; to find the spreading power $S(i)$ of node i , we simply will add up the infection coefficients involving node i . That is:

$$S(i) = \sum_j C(i, j) = \sum_j \sum_{h=1}^{\max} w^h NSR^h(i, j) . \quad (6)$$

Note that the sum over j can be unconstrained, since, by excluding self-retracing paths, we have forced $C(i, i) = 0$.

4.2 An approximation

Equation (6) gives an expression for the spreading power $S(i)$ of node i . However, lacking a general expression for NSR^h , we cannot test our assumption—namely, that EVC has a strong positive correlation with $S(i)$.

We can however say something about this connection, if we make a gross oversimplification of our expression for $S(i)$. That is, we include self-retracing paths, and so replace NSR^h with A^h . We will use $S^a(i)$ to denote this approximation. That is, we take

$$S^a(i) = \sum_j \sum_{h=1} w^h A^h(i, j) . \quad (7)$$

We leave the upper limit of the h sum unspecified. (It is unbounded if one considers all self-retracing paths.)

Now we gather together all the coefficients $S^a(i)$ into the single vector S^a . Also we define $\mathbf{1}$ to be a vector, of length equal to the number of nodes in the graph, with all entries equal to one. Then

$$\sum_j A^h(i, j) = (A^h \mathbf{1})_i .$$

Hence

$$S^a = \sum_{h=1} w^h A^h \mathbf{1} .$$

Next we decompose the vector $\mathbf{1}$, using as a basis the set of eigenvectors α of A :

$$\mathbf{1} = \sum_{\alpha} \sigma_{\alpha} \underline{\alpha} , \quad (8)$$

where σ_{α} is the scalar product of $\mathbf{1}$ and α . This gives, in turn,

$$S^a = \sum_{\alpha} \sum_{h=1} (w \lambda_{\alpha})^h \sigma_{\alpha} \underline{\alpha} = \sum_{\alpha} g_{\alpha} \underline{\alpha} . \quad (9)$$

The quantity g_{α} is simply the scalar product of α with the vector S^a . Now we note that the principal eigenvector of A is simply the vector of eigenvector centrality scores. Let us denote this vector as π . If our basic assumption is true (and if S^a were a good approximation to S), then the overlap g_{α} should be

positive, and large in absolute value compared to the other values g_α . That is, S^α should be composed “mostly” of π .

Now we note that

$$g_\alpha = \sum_h (w\lambda_\alpha)^h \sigma_\alpha.$$

This reveals two reasons why g_π should be the largest of all the overlaps g_α . (i) The eigenvalue λ_π is the largest of all the eigenvalues. (ii) The coefficient σ_α is simply the algebraic sum of the components of the vector α . However the components of π are *all positive*—while, for all other eigenvectors, they are of mixed sign. Hence we expect the overlap σ_π to be, in general, larger than the others. For instance, if we normalize the set of α with the L1 norm, then σ_π is equal to 1, while all of the other σ_α are less than one.

These arguments would be persuasive, had they been obtained using S rather than S^α . Instead, they must be viewed as suggestive at best. It is interesting to note that the value of the parameter w plays no role in the argument.

4.3 Modelling the spreading process

Now we develop a mathematical model of the spreading process. Not surprisingly, the adjacency matrix A plays a central role in our model; hence this mathematical model throws some light on our notions of the role of eigenvector centrality in spreading. We begin our modelling with the deterministic case, $p=1$.

4.3.1 $p=1$

The case of SI spreading with $p=1$ has much in common with the calculation of eigenvector centrality. To make them as similar as possible, we note that the result of an iterative EVC calculation, using the power method, is independent of start vector. Therefore, we take the start vector for the EVC calculation to be the same as that for spreading: one unit of weight at node J , and zero elsewhere. We call this start vector x^0 .

To calculate EVC by the power method, starting with x^0 , one simply multiplies by the adjacency matrix A in each iteration. The matrix A is thus the “time evolution operator” for this process: $x^{t+1} = Ax^t$. After many iterations, there is large weight at every node (assuming one does not rescale the weights). These weights all grow with each subsequent iteration; but the *relative* values of the weights converge to a set of constants, giving a weight vector growing in length but not changing in direction. This convergent vector (regardless of length) is the eigenvector e whose entry e_i is the EVC for node i . Also, given our start condition, all the weights will be nonnegative integers at all times, and positive integers after some finite time.

Now we consider SI spreading for the same start condition x^0 . The time evolution equation is as follows:

$$x^{t+1} = \text{sgn}[(1+A)x^t]. \quad (10)$$

That is, (i) the A operator of the EVC evolution equation is replaced with $(1+A)$; and then (ii) the result is truncated using the sgn operator—which maps any positive integer to 1, and 0 to 0.

The first difference (i) reflects the fact that, each time a node sends infection to its neighbors, it retains its infected state—whereas, with the EVC calculation, nodes send out all of their weight at time t , so that their weight at $t+1$ depends entirely on inputs from neighbors. This difference is fairly trivial: the dominant eigenvector of $(1+A)$ is identical to the dominant eigenvector of A . Hence, without the sgn operator, both processes would converge to the same distribution.

The second difference—the truncation operator—reflects the fact that a node, once infected, cannot become “more infected” as a result of repeated transmissions from its neighbors. This difference of

course has a dramatic effect—the convergent distribution for the spreading process is the saturated state: a vector of all 1's. We note, in this context, that a major source (but not the only one) of repeated transmissions to a given node j is infection paths that are self-retracing. For example, one time step after node j is infected, all of j 's neighbors will be infected; and after one more time step, these neighbors will in turn, under the action of the A part of $(1+A)$, retransmit to j . The sgn operator nullifies the effects of all transmissions to a given node after the first; hence, self-retracing infection paths have zero effect under the SI time evolution operator.

Finally, we note that Equation (10) calls for truncating after every time step. However, it is also true that

$$x^t = \text{sgn}[(1+A)^t x^0]. \quad (11)$$

That is: since all positive integers map to 1, one can find the distribution at any time t by applying the $(1+A)$ operator t times, without truncating, and then finally truncating only once, at the end of the run. Before this final truncation, for large t , the distribution will approach the same as that for eigenvector centrality—then the truncation operator throws away all this information, and simply places a 1 at every infected node.

This observation suggests that one might observe a “piling up” of weight at nodes of high EVC, if one were to modify the spreading process in such a way that it were possible to pile up weight. We then note that *probabilities* have this property—that is, repeated time steps, when each link has probability p for transmission, will steadily increase the probability that an uninfected node will be infected by its infected neighbors. Hence we are motivated to look at the case of spreading for $p < 1$.

4.3.2 $p < 1$

To discuss the probabilistic case, we consider a series of experiments, with fixed network topology, fixed p , and fixed start vector x^0 . For many such experiments, we can discuss the probability of a given node j being infected at time t . As noted above, this probability (denoted u_j^t , to avoid confusion with the parameter p) will grow with time. An exact theory of these experiments will hold the probability u_j^t less than 1 for every node, for all time. However any approximate theory—especially when the question of interest is the rate of growth of infection probability with time—must take care to avoid spurious effects due to the approximations used.

An exact theory may be stated as follows. Define r_j^t to be the probability that node j receives a transmission at time t . (The timing convention here is that, if node j is not infected at time t , and receives a transmission at time t , then it is infected at time $t+1$. That is, we can think of the event r_j^t of receiving a transmission as occurring slightly ‘after’ time t .) The probability r_j^t depends of course on the infection probabilities u_k^t for the neighbors k of node j . Specifically: neighbor k has probability $(u_k^t)p$ of transmitting to j at time t . Hence the probability of *no* neighbor transmitting to j is

$$\prod_{k=nn(j)} (1 - u_k^t p) ;$$

and so the probability of at least one neighbor transmitting to j is

$$r_j^t = 1 - \prod_{k=nn(j)} (1 - u_k^t p) . \quad (12)$$

Now we look at u_j^t . We note that node j will be infected at time $t+1$, unless two conditions are fulfilled: (i) it was not infected at time t , *and* (ii) it did not receive a transmission at time t . Hence the probability that j is not infected at $t+1$ is $(1 - u_j^t)(1 - r_j^t)$. Thus the probability that j is infected at $t+1$ is

$$u_j^{t+1} = 1 - (1 - u_j^t)(1 - r_j^t) = u_j^t + r_j^t - u_j^t r_j^t . \quad (13)$$

Equations (12) and (13), along with the starting condition $u_j^0 = x_j^0 = \delta_{j=J}$, give the dynamical rule for the growth of probability in a system with all links having transmission probability p per unit time.

These equations are of course easier to simulate than to solve! (Examples of such simulation were given in Section 3.) However we will seek to learn what we can from them.

First we note that the RHS of (13) is of the form (noninteracting case) + (corrections). The corrections come in, once again, because we cannot infect a node twice. For example, the term $u_j^t r_j^t$ is a correction of this sort: the naïve rule (ie, one ignoring problems of double-counting) would be that the infection probability at $t+1$ is simply the first two terms, ie, the probability of infection at t , plus the probability of transmission at t . Also, the naïve version of (12) would ignore all terms of higher order than p in the product, giving

$$r_j^t \approx \sum_{k=nn(j)} u_k^t p = (pAu^t)_j . \quad (14)$$

Combining the two naïve forms of (12) and (13) gives

$$u^{t+1} \approx u^t + pAu^t = (1 + pA)u^t . \quad (15)$$

Thus we see that our naïve time evolution operator (ignoring double counting) is simply $(1+pA)$. Thus the naïve version for $p = 1$ is simply $(1+A)$ —the same as is obtained from the time evolution operator from Section 4.3.1, if we ignore the sgn operator.

We note in this context that the ‘system matrix’ (time evolution operator) of Wang et al [11] is simply our ‘naïve’ time evolution operator, corrected for a uniform decay rate for the infected state. Thus their system operator, our naïve time evolution operator, and the adjacency matrix A all have the same dominant eigenvector, the components of which are the nodes’ EVC values.

Thus we find that the probability vector may be written as

$$u^t = (1 + pA)^t x^0 + (\text{corrections}) . \quad (16)$$

The operator $(1+pA)$ has the same dominant eigenvector as does A itself. Hence repeated multiplication by $(1+pA)$ will drive a vector towards a distribution consistent with eigenvector centrality: each node’s weight will be proportional to its EVC score. However, at large t , we know that the (corrections) must also be large; so we cannot draw any clear conclusion about the large- t case from Equation (16). We do however believe that Equation (16) supports assertion d. in our list of assertions a—g of Section 2. That is, for small t and small p —which thus characterizes the early, flat part of the S curve—we expect the corrections to be small.

We have carried out a perturbation expansion, in powers of p , for the coupled equations (12)—(13) above. We expanded to terms of $O(p^3)$. This allows us to look at the vector u^t up to time $t = 3$, and hence within a radius of three hops from the start node J . Our results are somewhat cluttered, and will not be given here. However, they conform (as they must) to the form of Equation (16); and furthermore, we see that, for each power of A , the leading (lowest order in p) term is simply $(pA)^t$ —the “naïve” term. That is, the correction terms, at each radius from J , are of higher order in p than the naïve term. Thus the naïve term dominates—and, as we pointed out above, the naïve term tends to drive the probability vector towards higher EVC. Thus, our perturbation expansion also supports our prediction d.

This observation, although gleaned from our perturbation expansion, in fact holds in general. That is, paths between i and j which are short have a higher probability of infecting j from i than longer paths—for instance, a path of length 4 acquires a factor p^4 , while a path of length 3 has a corresponding factor p^3 . Hence, even without extending our series expansion in p to higher order, we know that the leading term at a given radius from start node J comes from the “naïve” part of (16)—which is, in turn, the part which tends to build up weight in the same pattern as the eigenvector centrality [the dominant eigenvector of $(1+pA)$]. In other words: we are stating the obvious—that a node j is most likely to be infected from J over a shortest path between them—and then drawing a less obvious conclusion: that the (corrections) in (16), for the infection $J \rightarrow j$, are likely to be smaller than the leading term. This is certainly true for very small p ; and yet it is also true for the other extreme of $p = 1$ —where only shortest paths play a role. Hence we speculate that the naïve term in (16) dominates for any p . If this is true, then infection probabilities will approach a distribution proportional to eigenvector centrality, as long as those probabilities do not approach their upper limit

of 1 too closely. That is, we again find support for our claim that infection will move towards higher EVC. Finally, we note that our simulations show qualitatively the same behavior (for the same start node) over the whole range of p , from a few per cent to one. This observation supports our speculation that the dominant term in (16) is the same for all p .

5 Discussion and future work

We have applied our method of structural analysis for undirected graphs, developed in [1], to the problem of gossip-like spreading on a network. We believe that our topographic picture of the structure of a network, based on using eigenvector centrality (EVC) as a height function over the network, with mountains, peaks, slopes, and valleys, is an excellent starting point for an understanding of spreading. In this work, we have built from this starting point, and developed a set of qualitative arguments which yield seven specific predictions (Section 2). Our picture, in short, is that an initial infection on the side of a mountain will run ‘up’ the mountain, while the rate of infection of new nodes grows with height. This is a self-reinforcing process, so that infection rate takes off at some point high up on the mountain, and the whole top is saturated quickly; finally the remaining hillsides are saturated at an ever decreasing rate. These predictions are tested, and convincingly confirmed, in a series of simulations that we have run on various social networks in our possession.

To supplement these qualitative arguments and simulations, we have developed a mathematical theory of two things: the definition of the spreading power of a node, and the dynamics of simple SI spreading. In each case, exact solutions are not possible, due to the problem that double infections must not be counted. However, in each case, we have shown that ignoring the double-counting problem gives an approximation which supports our basic claim (spreading power may be approximated by EVC). In particular, we present arguments why the correction terms due to double counting are likely to be small compared to those which ignore double counting; and these latter terms support the claim that infection probability is positively correlated with EVC.

There is much work remaining to be done on the theoretical side. The connection between our definition of spreading power $S(i)$ for node i , and the eigenvector centrality EVC for the same node, needs to be tested further. To do this, one needs ways to calculate the number of non-self-retracing paths at distance h . This could be done numerically for the smaller graphs we have studied, and compared with EVC. We also have some (unpublished) analytical results for EVC on certain types of tree graphs; it would be interesting to compare these values to values for NSR^h .

It would be of interest to apply our topographic approach to understanding spreading to other cases. For one thing, our analysis suggest simple ways of *hindering* spreading: one could immunize (permanently) key nodes in the network, such as highly central nodes in one region, or nodes whose links provide bridges between regions. These strategies for ‘immunizing’ a network could be tested via further simulations. Also, one can study ways for modifying the topology of the network, towards the goal of hindering or helping spreading. In the former case one might seek to isolate regions even further; in the latter, one might look for small modifications that cause two or more regions to fuse into one. Again, such strategies may be tested by simulations. Other, perhaps more realistic, cases to be studied, in the light of our topographic picture, include the case in which nodes are infected from ‘outside’ the graph at some steady rate; or the case where nodes lose their infection status after some time (SIS), perhaps after a refractory period (SIR).

We note that the *sfi* graph offered the most extreme behavior, stemming from the weak coupling both between and within its three regions. Thus we find the *sfi* graph to be the least well connected, in terms of criteria derived from our structural analysis, and from our observations of spreading. We have argued in [1] that the property of being poorly connected is related to poor mixing, and thus to a small eigenvalue gap (difference between the dominant and second eigenvalues of A). It would be of interest to test these ideas with the set of graphs studied here. If the gap is small in poorly connected (by our definition) graphs, then many things will be relatively sensitive to small changes in topology: their EVC values (from the dominant eigenvector); their topography, as obtained by our analysis; and their spreading behavior. This too could be tested.

More generally, we wish to deepen and render more quantitative our notion of the well-connectedness of a graph. Clearly, our coarse starting point (number of regions) give useful information in itself; but much more remains to be done. Besides making a connection to the eigenvalue gap, one could seek to quantify the degree of inter-region connectedness. Furthermore, one should seek connections and correlations between these different measures.

Finally, we note that our method of network structure analysis suggests a method of graph visualization. Figure 8 is an example of this: nodes in each region are placed close together on the page, so that the regions (and the height function defining them as ‘mountains’) are visually clear. Figure 8 represents work in progress. We hope to be able to present improvements on, and refinements of, this visualization approach in future work. One very attractive goal is to display epidemic spreading simulations on a network visualized as in Figure 8—in the form of snapshots, or a movie. Then one can hope to see our predictions a—g, not in the form of static plots as in this paper, but in the form of time development of the infection over the topography (as displayed in 2D) of the graph.

Acknowledgments. We thank Michelle Girvan for sharing data on the SFI collaboration graph with us, and Mark Burgess for helpful discussions. This work was partially supported by the Future & Emerging Technologies unit of the European Commission through Project BISON (IST-2001-38923), and partially supported by the EU within the 6th Framework Programme under contract 001907 (DELIS).

References

- [1] Canright G, Engø-Monsen K: Roles in Networks. *Science of Computer Programming* 2004;53: 195—214.
- [2] Jovanovic MA, Annexstein FS, Berman KA: Scalability issues in large peer-to-peer networks - a case study of Gnutella. Technical Report, University of Cincinnati (2001).
- [3] Canright G, Engø-Monsen K, Weltzien Å, Pourbayat F: Diffusion in social networks and disruptive innovations. *IADIS e-Commerce 2004 proceedings*. Lisbon 2004.
- [4] Girvan M, Newman MEJ: Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 2002;99: 8271—8276.
- [5] Bonacich P: Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* 1972;2: 113—120.
- [6] Newman MEJ: The structure and function of complex networks. *SIAM Review* 2003;45: 167—256.
- [7] Pastor-Satorras R, Vespignani A: Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett* 2001;86: 3200—3203.
- [8] Pastor-Satorras R, Vespignani A: Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E* 2001;63: 066117.
- [9] Newman MEJ: Spread of epidemic disease on networks. *Phys. Rev. E* 2002;66: 016128.
- [10] Brauer F: A model for an SI disease in an age-structured population. *Discrete and Continuous Dynamical Systems* 2002;B2: 257—264.
- [11] Wang Y, Chakrabarti D, Wang C, Faloutsos C: Epidemic spreading in real networks: an eigenvalue viewpoint. *Proceedings, 22nd Symposium on Reliable Distributed Systems (SRDS 2003)*, 25—34.
- [12] Rogers EM: *Diffusion of Innovations*, ed 3. Free Press, New York, 1983.