On the radius of centrality in evolving communication networks

Article

Accepted Version

On the radius of centrality in evolving communication networks

Danica Vukadinović Greetham, Zhivko Stoyanov and Peter Grindrod

March 3, 2014

Abstract

In this article, we investigate how the choice of the attenuation factor in an extended version of Katz centrality influences the centrality of the nodes in evolving communication networks. For given snapshots of a network, observed over a period of time, recently developed communicability indices aim to identify the best broadcasters and listeners (receivers) in the network. Here we explore the attenuation factor constraint, in relation to the spectral radius (the largest eigenvalue) of the network at any point in time and its computation in the case of large networks. We compare three different communicability measures: standard, exponential, and relaxed (where the spectral radius bound on the attenuation factor is relaxed and the adjacency matrix is normalised, in order to maintain the convergence of the measure). Furthermore, using a vitality-based measure of both standard and relaxed communicability indices, we look at the ways of establishing the most important individuals for broadcasting and receiving of messages related to community bridging roles. We compare those measures with the scores produced by an iterative version of the PageRank algorithm and illustrate our findings with two examples of real-life evolving networks: the MIT reality mining data set, consisting of daily communications between 106 individuals over the period of one year, a UK Twitter mentions network, constructed from the direct *tweets* between 12.4k individuals during one week, and a subset the Enron email data set.

1 Introduction

Human communications in our age leave digital traces from which social networks can be created. These networks are large, multi-layered and dynamic, i.e. they evolve from moment to moment. Although in the last couple of decades we have seen a rapid development in the methods for analysing and modelling regular and arbitrary static networks, recent digital trends identified the need for existing methodologies to be extended and adapted to dynamic, evolving networks.

One of the very important and well-researched characteristics of an individual (a node) in a social network is its centrality score. Centrality measures the relative importance of a node and determines its involvement in a network. Although different centrality measures have been proposed, tested and compared on undirected, directed and weighted networks (for reviews see [4, 30, 16]), only relatively recently research has focused on centrality in evolving networks [25].

As the notion of centrality changes depending on the context, we are especially interested in the concept of the radius of centrality[2]. The radius of centrality corresponds to the size of magnifying glass, or the size of a window through which we observe a network. So, instead of looking at a simple ranking of vertices, we are interested in their neighbourhood of a given radius. This is motivated by structural holes [7] and the strength of weak links [22] theories that highlight the importance of "bridges" in networks.

1.1 Previous work

For a thorough review of different centrality measures in social networks see [4]. Centrality in weighted social networks is discussed in [30]. Regarding centrality which takes into account the number of possible paths from and to nodes, Katz centrality [28] computes the relative influence of a node within a network by measuring the number of its immediate neighbors and all the other nodes in the network that connect to the node under consideration through the immediate neighbors. Walks made to distant neighbors are penalised by an attenuation factor α. In [16] the matrix exponential[1] was used as a centrality measure where each path of length l is penalised by a factorial of l. The Katz centrality was recently revisited in [25]. A new measure, called communicability across evolving networks, based on the extension of Katz centrality to evolving networks is proposed, which takes into account the time-flow[2]. Communicability across evolving networks is already successfully implemented on the small scale, on mobile phone and email communication networks [25, 23] and C. Elegans brain networks [11]. However scaling it up to very large data sets involves handling large matrices. Communicability is extended further in [24] adding an additional attenuation factor for time, which discounts for age (older paths are penalised more heavily).

Another measure of centrality in static networks, Bonacich centrality [2], [3] introduces another parameter, similar to Katz centrality, but penalising direct and indirect links. Recently, this measure was revisited in [20] where the authors investigated ranking of nodes and communities' structure using a normalised variant of Bonacich centrality[3]. Note that in the original Bonacich centrality, the attenuation factor α is bounded by the spectral radius (the largest eigenvalue of an adjacency matrix of a graph) in order for the infinite sum to converge. The normalised version converges in any case, so no such bound is needed anymore. The authors motivate their work by the fact that the computation of a spectral radius ρ_A is difficult, "especially for large networks" (see pp. 2 of [20]). However, we note that in social network analysis settings where networks are mostly sparse, power iteration or similar methods could be used. Those methods would efficiently compute an approximated value of ρ even for very large sparse matrices, or else the Perron-Frobenius theorem (see e.g. [19]) provides simple but useful spectral radius bounds. However, we argue that a bound $\alpha < \frac{1}{\rho(A)}$ is limiting and should be relaxed for different reasons. Namely, when the spectral radius is greater than 2, it penalises too heavily not-so-long paths, and thus lowers significantly the centrality ranking of nodes that connect different communities. This on the other hand might have implications in social network analysis of large networks. The "structural holes" theory [7] refers to the absence of links between two parts of a network. Brokerage exploits structural holes – an individual is connected to two other individuals or communities not mutually connected. This position could be beneficial for such an individual (a broker) as she/he could control the flow of information between two communities, profit from two different sources of information and mediate trade between them. Also bridges between different communities in a social network are important when trying to identify communities in a large unknown network and to run a network-based intervention which depends on community structure to change behaviour of individuals in the network [33].

In the empirical analysis of several real-world and artificial generic models of networks, Ja-

[1]The matrix exponential is defined for $A \in \mathbb{C}^{n\times n}$ by $e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots$.[26].

[2]If A sends a message to B today and B sends a message to C tomorrow, then A could possibly reach C with its message via B, but the opposite path from C to A is impossible, as it would go against the time-flow.

[3]The Bonacich $\alpha-$centrality matrix is given by $C(\alpha,\beta,n) = \beta A + \beta\alpha_1 A^2 + \cdots + \beta\prod_{k=1}^{n}\alpha_k A^{n+1}$, where A is adjacency matrix. The normalised centrality matrix is then defined as $NC(\alpha,\beta,n\to\infty) = \frac{C(\alpha,\beta,n\to\infty)}{\sum_{ij} C_{ij}(\alpha,\beta,n\to\infty)}$

maković et al. [27] looked at the different upper bounds of spectral radius from the simple bound given by the graph's maximal degree (see e.g. [34]) to more complex bounds featuring local information (average neighbours degree [12]) or global information such as diameter [10]. They found that for three real-world networks they investigated, a bound given by [12] was the closest to the observed values, while for the Internet autonomous systems topology that bound was over-estimating the real value significantly. When using artificial generic networks (random[15, 21], small-world[38], preferential attachment [1] networks) with the same number of nodes and edges as in the real-world networks, the spectral radii of all three types of networks were much smaller than the real-world one in the Internet AS topology case. This finding is important because the spectral radius is also found to be connected to epidemic spreading in networks (see [37, 8, 35]).

In the following section we discuss how communicability indices across evolving networks are related to spectral radius and compare a relaxed centrality measure presented in [36] with standard communicability [25], an extension of Estrada-Hatano communicability [16] for evolving networks, and iterative PageRank[31]. Our measure relaxes attenuation factor constraints previously imposed by the spectral radius in order for the communicability indices to be well-defined. We then compare different versions of vitality measure[4] based on centrality indices which may enable us to detect the individuals whose lack of existence would result in the biggest changes in centrality in evolving networks. We apply our findings to two real-life networks, comparing the original [25] versus relaxed communicability measures, their vitality versions, and benchmark them against the PageRank centrality [31] on the aggregated versions of those two evolving networks. We conclude with a discussion of results and give some ideas for future work.

2 Centrality measures for evolving networks

An evolving network is a family of graphs $G_i = (V, E_i)$, where the vertex set V is given in advance and is fixed throughout time, and E_i is a set of edges on V at time i. We assume that the time is discrete and finite, i.e. $i = 0, 1, \cdots, n$. The corresponding adjacency matrices at time i are denoted by A_i and are such that

$$A_i(k,l) = \begin{cases} 1, & \text{if there is an edge from } k \text{ to } l \\ 0, & \text{otherwise.} \end{cases}$$

2.1 Communicability and PageRank

For a static network, the Estrada-Hatano communicability matrix [18, 16] is obtained from the adjacency matrix, A, associated with the network by the formula

$$Q_A = e^A, \tag{1}$$

where the matrix exponential of A is defined as

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k. \tag{2}$$

[4]For a real-valued function f on G, a vitality measure based on f quantifies the difference between the value $f(G \setminus v)$ and $f(G)$ for each $v \in G$. See [5]

The (i,j)-th entry of Q_A is a weighted sum of all walks starting at vertex i and ending at j, where a walk of length k is penalised by $\frac{1}{k!}$. Therefore, the row sum $\sum_j Q_A(i,j)$ can be defined as the broadcast score of vertex i, and the column sum $\sum_i Q_A(i,j)$ as the receive score of vertex j. This can be extended to evolving networks by

$$Q = \prod_{i=0}^{n} e^{A_i}, \tag{3}$$

where A_i is the adjacency matrix at time i. Note that in general $A_iA_j \neq A_jA_i$, so $Q \neq e^{\sum_{i=0}^{n} A_i}$. The product (3) can be computed directly for small-scale networks. This sums up all walks starting at vertex i and ending in j (the hops might be in different time steps, but because of general noncommutativness of A_i and A_j, only walks continuing in future count) and the walks of k hops are penalised by $\frac{1}{k!}$. Given an evolving network, another version of a communicability matrix that is closer to the Katz definition of centrality (see [25] for more details) is given by:

$$Q = \prod_{i=0}^{n} (I - \alpha A_i)^{-1} \tag{4}$$

where I is the identity matrix, the constant $\alpha < \frac{1}{\max(\rho(A_i))}$, and $\rho(A_i)$ is the spectral radius of A_i. Henceforth, we will refer to (3) and (4) as "*exponential communicability*" and "*standard communicability*," respectively. Analogously to the static case, the broadcast (resp. receive) indices are equal to the row (resp. column) sums of Q.

For a static network with adjacency matrix A, one may also consider the Google matrix [6, 29], G, associated with it:

$$G = A^T D_{\mathsf{out}}^{-1}, \tag{5}$$

where $D_{\mathsf{out}} = \mathsf{diag}(d_1, d_2, \ldots, d_N)$ is a diagonal matrix and d_i is the out-degree of vertex i,

$$d_i = \sum_{j=1}^{N} A_{ij}.$$

Here we assume that N is the number of vertices in the network and $d_i > 0$ for all $1 \leq i \leq N$. Then the vector of the PageRank scores of all vertices, $\mathbf{r}$, is defined by

$$\mathbf{r} \propto (I - pG)^{-1}\mathbf{1} \quad \text{and} \quad \sum_i r_i = 1, \tag{6}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^T$, and the constant p is called the *damping factor*, normally chosen equal to 0.85. In order to find $\mathbf{r}$, one may first compute $\mathbf{x} = (I - pG)^{-1}\mathbf{1}$ and then $\mathbf{r} = \mathbf{x}/\|\mathbf{x}\|_1$. From the Perron-Frobenius theorem, $\mathbf{r}$ exists and is unique, and its entries are all positive. Therefore, $\mathbf{r}$ may be thought of as a probability distribution over all vertices, and large values of r_i would mean that vertex i is very important (and vice versa). From (6) one can see that the PageRank of a vertex is the sum of the PageRanks of the vertices that link to it, each divided by its corresponding out-degree. Therefore, as in Katz centrality, each vertex is given an initial score equal to $1/N$, but unlike Katz centrality, in the PageRank algorithm each vertex distributes its score equally amongst its adjacent neighbours.

However, in an evolving network, if we already know the PageRank at time i, we may want to use that as an initial score for the computation of the PageRank at time $i+1$. Analogously to (6), for $i = 0, 1, \ldots, n-1$ we would have

$$\mathbf{x}^{[i]} = (I - pG_i)^{-1}\mathbf{r}^{[i-1]} \quad \text{and} \quad \mathbf{r}^{[i]} = \frac{\mathbf{x}^{[i]}}{\|\mathbf{x}^{[i]}\|_1},$$

where G_i is the Google matrix at time i and $\mathbf{r}^{[-1]} := \frac{1}{N}\mathbf{1}$. Thus, the PageRank in an evolving network will be given by

$$\mathbf{r}_{PR} \propto (I - pG_n)^{-1}(I - pG_{n-1})^{-1} \cdots (I - pG_0)^{-1}\mathbf{1} \quad \text{such that} \quad \sum_{i=1}^{N} r_{PRi} = 1.$$

This is very similar to the computation of the receive scores in the exponential and standard communicability [17],

$$\mathbf{r}_E \propto \exp(A_n)^T \exp(A_{n-1})^T \cdots \exp(A_0)^T \mathbf{1} \tag{7}$$

and

$$\mathbf{r}_C \propto (I - \alpha A_n)^{-T}(I - \alpha A_{n-1})^{-T} \cdots (I - \alpha A_0)^{-T}\mathbf{1}, \tag{8}$$

respectively. (Here we have used the notation $A^{-T} := (A^{-1})^T = (A^T)^{-1}$ for any invertible matrix A.) Therefore, one of our aims will be to compare $\mathbf{r}_{PR}$, $\mathbf{r}_E$, and $\mathbf{r}_C$.

Finally, if the adjacency matrix at time i, A_i, contains vertices with zero out-degree, we replace each of the corresponding rows in A_i with the vector $\mathbf{r}^{[i-1]^T}$ (i.e. the PageRank from the previous time-step), or $\frac{1}{N}\mathbf{1}$ if $i = 0$.

2.2 Spectral Radius Bound

If the attenuation factor, α, is such that

$$\alpha < \frac{1}{\rho(A_i)} \tag{9}$$

each of the terms $(I - \alpha A_i)^{-1}$ on the right hand side of (4) is the Katz centrality for the network at time i and can be expanded in a Taylor series,

$$(I - \alpha A_i)^{-1} = \sum_{k=0}^{\infty} \alpha^k A_i^k. \tag{10}$$

The requirement (9) is necessary in order for the right hand side of (10) to converge (in the standard matrix norm). Applying the same argument to each of the terms in (4), we obtain that α must satisfy $\alpha < \frac{1}{\max(\rho(A_i))}$. On the other hand, looking at each individual A_i, α can be interpreted as the probability that, once sent, a message will be successfully transmitted by any receiving node to any of its contacts. Then the expected length of a single transmission, sent from nodes in the network corresponding to A_i, is[5]

$$\sum_{k=0}^{\infty} k\alpha^k(1-\alpha) = \frac{\alpha}{(1-\alpha)}, \tag{11}$$

[5]Note that $\sum_{k=0}^{\infty} k\alpha^{k-1} = \frac{1}{(1-\alpha)^2}$.

for $|\alpha| < 1$. However, for matrices whose spectral radius is greater than 2, α must be chosen less than 1/2, because of (9). So in this case, the expected length of a transmission must be less than 1, due to (11). Similarly, $\rho(A_i) > 3$ implies expected transmission length less than 1/2, and so on.

It was shown in [27] that some real-world networks have spectral radii greater than 2, so their expected path lengths are less than 1. This means that paths between two communities, especially, are too heavily penalised. In order to mitigate the attenuation, we propose to normalise A and relax the condition on the attenuation which allows for longer paths and so rewards individuals that act as bridges between different communities appropriately.

2.3 Relaxed Communicability

For a large data set where the size of the matrix Q is prohibitive, and computing the inverse of such a large matrix represents a challenge, an approximation of Q can be computed using a Taylor series approximation ignoring summands of order higher than some n, depending on the application. In order to compute $(I - \alpha A)^{-1}\mathbf{1}$ without storing Q, the following method can be used where $\mathbf{b}$ is initialised to the all ones vector of length n.

$$(I - \alpha A)^{-1}\mathbf{b} = \mathbf{b} + \alpha A\mathbf{b} + \alpha^2 A^2\mathbf{b} + \cdots \tag{12}$$

We will use this representation to define new relaxed communicability indices. Instead of having $\alpha = \frac{1}{2\max(\rho(A_i))}$, for the expression to converge, it is enough for α to be less than 1, and that A is normalised. From the expression for the expected path length (11), ensuring that

$$\alpha < \frac{l}{1+l} \tag{13}$$

where $l \in \mathbb{N}$ is the expected path length, we have that α will always be less than 1 and we can set parameter l on a desired path length depending on a context, i.e. what kind of centrality we are interested in. Thus, to obtain relaxed communicability indices, one should choose a length of path l depending on the application, calculate α from (13) for the given l, initialise b to be all-ones vector and multiply it with α and the matrix A_i normalised with the 2-norm of A_i iteratively. Summing up all iterative factors up to the order n, which depends on how small one's approximation error needs to be, gives the result for A_i. Results need to be multiplied for all consecutive A_is. In the case of a small graph, Q can be obtained directly from (4) using the computed α and replacing A with $\frac{A}{\|A\|}$.

2.4 Vitality Measure

For a given real-valued function f on a network, sometimes the importance of individual vertices or edges is quantified by computing the difference of the value of f on a network with and without a vertex or an edge. This type of measure is called a vitality measure [5] and can be used for different functions f. We are motivated here by the need to rank vertices by their centrality, and having the additional challenge to do this dynamically. For a given evolving communication network, one can envisage the task of assigning centrality indices dynamically and trying to identify either vertices that experienced a significant change in ranking, or the time-period in which the most significant change happened or both.

In order to rank the nodes by importance during a time period we formulated a vitality-based measure by computing the corresponding centrality indices in the absence of one node at time.

For a series of adjacency matrices $A_{i_1}, \cdots, A_{i_2}$ we compute communicability indices using both standard and relaxed communicability indices. Furthermore, for each vertex k, we compute Q_k, which is obtained by deleting exactly the kth row and the kth column from $A_{i_1}, \cdots, A_{i_2}$, and then calculating both versions of communicability. Then we calculate the difference between Q_k indices and Q for each k, as a sum of least squares to check which nodes are responsible for the biggest changes in indices' values. We give a pseudo-code for our vitality measure (Algorithm 1 below), which is independent of the version of communicability used, standard or relaxed. We will discuss in the next section how results depend on the type of communicability used.

VITALITY MEASURE

```
compute indices (column and row sums of Q for A_1...A_n)
ls=0
for j=1:N
    remove A_k(j, .) and A_k(., j) for k= 1,..., n
    compute indices_j (column and row sums of Q_j)
end
for i=1:j-1
    ls(j,1)=ls(j,1)+(indices_j(i,1)-indices(i,1))^2;
    ls(j,2)=ls(j,2)+(indices_j(i,2)-indices(i,2))^2;
end;
for i=(j+1):n
    ls(j,1)=ls(j,1)+(indices_j(i-1,1)-indices(i,1))^2;
    ls(j,2)=ls(j,2)+(indices_j(i-1,2)-indices(i,2))^2;
end;
ls=sqrt(ls)
```

Algorithm 1: The vitality measure of each vertex is an entry in the variable `ls` ($N \times 2$ array): the first column of `ls` contains vitality measures of broadcast scores, and the second - of receive scores. The variable `indices` ($N \times 2$ array) contains the column- and row-sums of Q. When Q_j is computed, we have removed the j-th row and j-th column in each adjacency matrix, A_i. The variable `indices_j` ($(N-1) \times 2$ array) contains the column- and row-sums of Q_j.

3 Two case-studies

We used two real-world data sets. The first one is part of the MIT Reality Mining data set (c.f. [13, 14]) and consists of the mutual mobile phone communications over a year of 106 individuals, most of whom were either students, or faculty members at MIT. The second is the data set obtained from Twitter UK mentions network collected on our behalf by Datasift, Twitter's certified partner. The network was created from public messages that users located in UK sent to each other on Twitter using the @ sign during 1 week in Dec 2011. In both cases we aggregated the data on a daily basis.

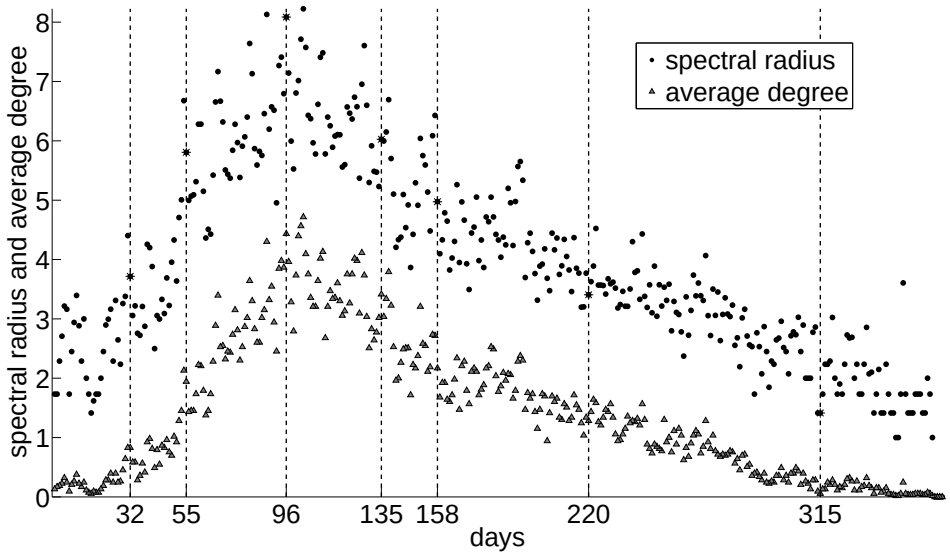


Figure 1: MIT data: Spectral radii of A_1 to A_{365} matrices.

3.1 Case-study 1: MIT Reality Mining Data

The data is aggregated on a daily basis, and contains 365 binary adjacency matrices, from the 20th July 2004 onwards, denoted with A_1 to A_{365}. An entry (i, j) of A_k is equal to 1 if there was at least one phone call between individuals i and j on day k. Fig. 1 shows the spectral radii and the average degrees of all 365 matrices. One can observe how the communication activity (average degree) changes through the year: e.g. we have very little communication during the summer vacation (this is, roughly speaking, the first 30–40 days), followed by a significant increase, which coincides with the start of the new academic year (between the end of September and the beginning of October) until about Christmas and New Year, etc.

In Fig. 2 we show an example of a daily communication network. The vertices with labels 11, 46, 60, and 72 (highlighted) have relatively small degrees, but they connect different communities and therefore are important. From all 365 daily networks, we chose a subset of seven, on the days 32, 55, 96, 135, 158, 220, and 315, as being representative of the variation in the volume of communication throughout the year, in terms of the spectral radii (and the average degrees) of all adjacency matrices (see Fig. 1). We computed broadcast and receive scores using the standard, relaxed, and exponential versions of communicability, as well as PageRank. Our aim here is twofold: on the one hand, we would like to see which of these four methods captures better the importance of community bridges; on the other hand, we want to compare the receive scores computed by all four methods (incl. PageRank) over the seven-day network. We do this in three steps: firstly, we consider day 32 (Fig. 2) on its own, as a static network, and compare (pairwise) the scores produced by all methods, and the relaxed vitality measure. The idea is to see, graphically, how each method ranks community bridges. The results are plotted in Figures 4, 5, and 6. Secondly, we compare how standard, relaxed, and exponential broadcast scores, computed over all seven days, rank community bridges. The results are in Table 1, and Figures 8 and 9. Finally, we consider the receive scores produced by all methods (Figures 10, 11, and 12) and discuss the differences between

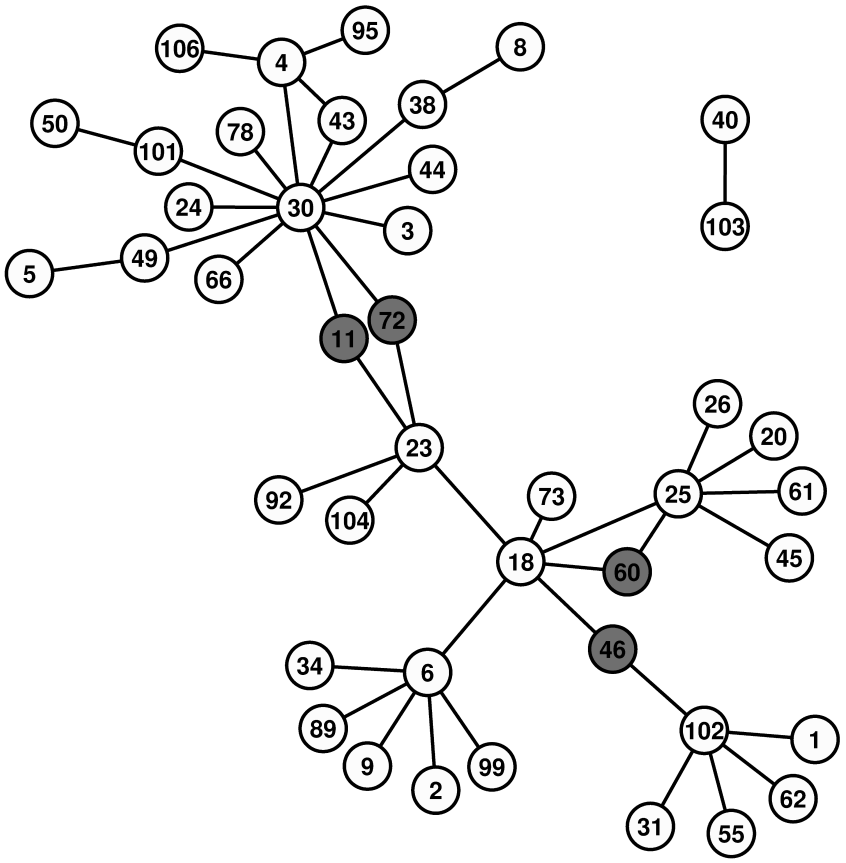
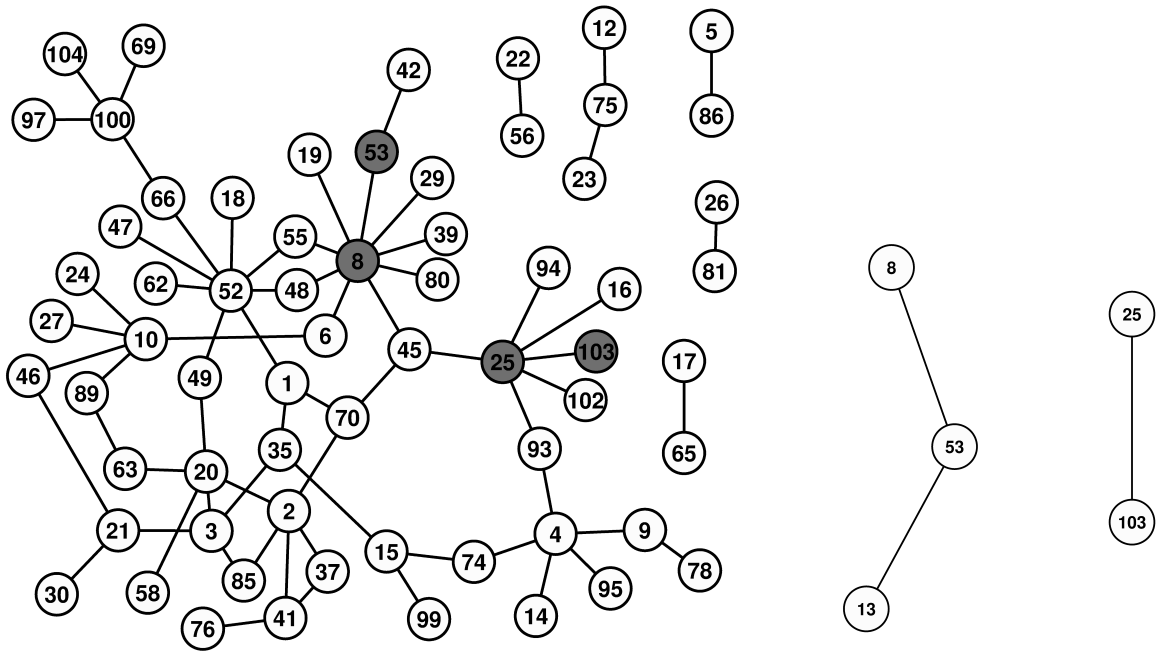


Figure 2: MIT data, the network on day 32.



(a) Day 220 (b) Day 315

Figure 3: MIT data, the network on days 220 and 315. The latter consists of nodes 8, 13, 25, 53, and 103 only.

them. In order to do that, and since the computation of receive scores resembles that of broadcast scores with time reversed, we also give the communication networks on days 220 and 315 (Fig. 3).

We start by considering day 32 as a static network. The results in Fig. 4 and 5 suggest that there is a high correlation between Katz centrality (the basis of standard communicability) and Estrada-Hatano communicability (the static equivalent of exponential communicability). Also, in this particular case we have established that there is a relatively strong correlation between relaxed communicability (RC) and Katz centrality (KC). Further, we notice that both KC and Estrada-Hatano communicability (EH) rank community bridges higher than does PageRank (PR). This is because community bridges are usually adjacent to hubs - if not directly, then a short walk away from them. Thus, in KC and in EH community bridges benefit fully from the high score of the hubs, while in PageRank that latter score is distributed equally amongst all vertices adjacent to the hub. Also, the scores of vertices 46 and 60 are lower than those of 11 and 72, because the latter are closer to a bigger community. We can also see that PR tends to rank vertex 102 higher than KC and EH. This can be explained by noting that, even though 102 is a hub, it is relatively far from the rest of the network. Therefore, its importance will be scaled down by KC and EH. On the other hand, it is known that, for an undirected network, PR is similar to degree centrality [32]. The same applies to vertex 6, and to a lesser extent to 25.

From Fig. 6 we can see that, in terms of ranking vertices 11, 46, 60, and 72, there is a relatively strong correlation between the relaxed communicability scores and the relaxed vitality measure (both with $l = 4$). Specifically, vertices 11 and 72 have ranks 13 and 14 according to the vitality measure, and ranks 7 and 8 according to relaxed communicability; vertices 46 and 60 have ranks 15

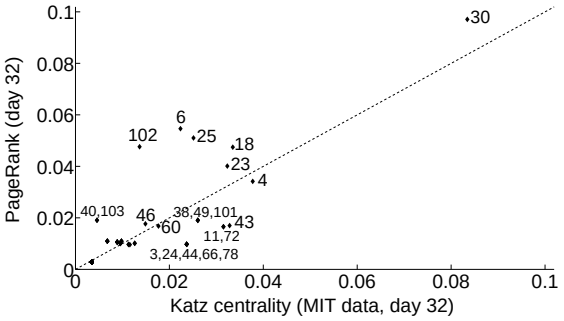


Figure 4: MIT data, day 32: PageRank vs. Katz centrality.

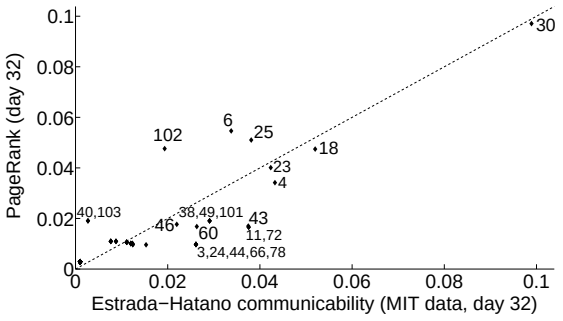


Figure 5: MIT data, day 32: PageRank vs. Estrada-Hatano communicability.

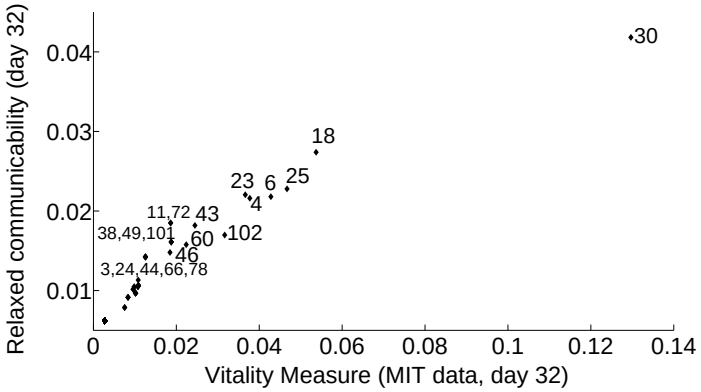


Figure 6: MIT data, day 32: Relaxed communicability ($l = 4$) vs. Relaxed vitality measure ($l = 4$).

and 9 (respectively) according to the vitality measure, and ranks 15 and 14 (respectively) according to relaxed communicability. The overall rank correlation (Kendall's τ coefficient) between the vitality measure and relaxed communicability scores is also (relatively) high in this case, $\tau = 0.8$.

Table 1 presents a comparison of the rankings of the highlighted vertices, in descending order (1 top to 106 bottom), when different methods were used. The table shows that, in terms of broadcast score, relaxed communicability (with expected pathlengths 4 and 7) ranks bridge vertices much higher than does standard communicability. Also, as expected, we can see that the importance of community bridges (as broadcasters) increases when we increase the expected pathlength in the relaxed communicability. Similarly, by comparing the coefficients scaling A^k in (2) and (10), we can see why exponential communicability has ranked bridge nodes considerably higher than standard or relaxed communicability have (see Fig. 7). For comparison, in this particular computation of standard communicability, we have found $\alpha = 0.9/\max_i(\rho(A_i)) = 0.11$, and in the case of relaxed communicability ($l = 7$) we have $\alpha = \frac{l}{l+1} = 0.875$ and $\|A_i\|_2 \approx 3.7, 5.8, 8.1, 6.0, 5.0, 3.4, 1.4$ in the chosen seven days (see Fig. 1).

In Fig. 8 and 9 we show scatter plots of standard vs. relaxed broadcast indices and standard vs. Estrada-Hatano (exponential) communicability indices. The upper left diagonals of the figures represent nodes that have higher rankings in relaxed (resp. exponential) than in standard indices. Note that in both standard and relaxed communicability more weight for broadcast indices lies on

Rank / Vert. no.	St. br.	Rel. br. ($l = 4$)	Rel. br. ($l = 7$)	Exp. br.	Vit. m.
11	27	17	13	8	23
46	36	25	24	19	22
60	60	34	25	18	32
72	52	24	18	10	28

Table 1: Comparison of the ranking of the broadcast scores of vertices 11, 46, 60 and 72 (see Fig. 2) over all seven days. Here the ranking is in descending order (i.e. vertex ranked 1 has the highest broadcast score, and vertex ranked 106 has the lowest). Different columns correspond to different methods for computing centrality: Standard broadcast, Relaxed broadcast with $l = 4$, Relaxed broadcast with $l = 7$, Exponential broadcast, and Relaxed vitality measure.

the first matrix in the sequence[6], that is, the adjacency matrix of day 32 in this case, while for receive indices it is the last matrix that carries most of the weight. While Estrada-Hatano indices do not correlate with standard or relaxed indices, they still rank higher most of the community bridges. In particular, vertices 11 and 72 are ranked much higher in terms of their exponential broadcast scores than they are according to their standard and relaxed ones. Thus, Estrada-Hatano (exponential) broadcast indices could be used when the expected transmission length, l, is not known. However, if l is known, relaxed communicability will highlight more relevant nodes.

With respect to the above, it is worth mentioning that, generally speaking, whether a method is better, or not, is not determined by how high it ranks community bridges, although this feature is important. The similarity between exponential and relaxed communicability is that they put much higher weights on shorter paths, but while exponential communicability still allows for infinite paths, relaxed communicability only considers walks whose length doesn't exceed a given number. Which of these two methods is better can only be determined by the particular application.

We now turn our attention to comparing receive scores over the seven-day network from all the methods. From all definitions of communicability (see also (7) and (8)), standard, relaxed and exponential, it is clear that if the adjacency matrices A_i are symmetric, computing receive scores is the same as computing broadcast scores with time reversed. Therefore, the networks on days 220 and 315 will greatly influence the final receive scores. This is why in Fig. 3 we have shown those two networks. From Figures 10 and 11 we can see that relaxed and exponential receive scores are highly correlated. Also, it is clear that standard receive scores fail to recognise vertices 13, 53, 103, and to some extent also 25. The reason for this can be seen in the way the constant α is chosen, $\alpha = 0.9/\max_i \rho(A_i)$, and the fact that matrices A_{220} and A_{315} have the smallest 2-norms of all seven matrices. Therefore, standard communicability significantly scales down the impact of days 220 and 315 to the overall computation of the receive scores. The degrees of vertices 13, 53 and 103 over the seven days are: $[0, 1, 7, 4, 2, 0, 1]$, $[0, 1, 4, 2, 0, 2, 2]$ and $[1, 3, 2, 2, 0, 1, 1]$, respectively. This means that vertices 53 and 103 have low, or no activity during the days that count the most in standard communicability.

On Fig. 12 it is worth noting that iterative PageRank scores correlate well with relaxed receive scores, apart from vertex 103. When $l = 4$, relaxed communicability places all the weight on dynamic walks whose length doesn't exceed 4 edges. As can be seen in Fig. 3(a), these walks are much less in number for vertex 103, than they are for 53. We have established numerically that this

[6]For another extension of standard communicability that adds a penalty with regard to time see [24]

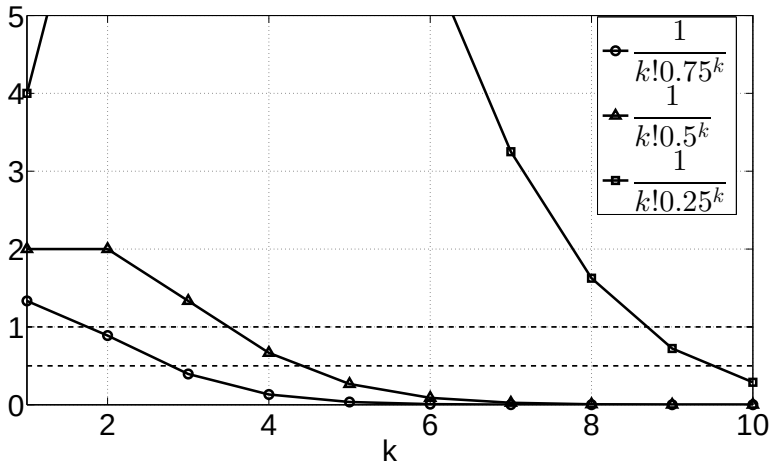


Figure 7: The ratio between the coefficients scaling A^k (i.e. walks of length k) in the exponential, and in the standard (and relaxed) communicability: $\frac{1}{k!}$, and α^k, respectively. Here $\alpha = 0.25$, 0.5, and 0.75; $k = 1, 2, \ldots, 10$.

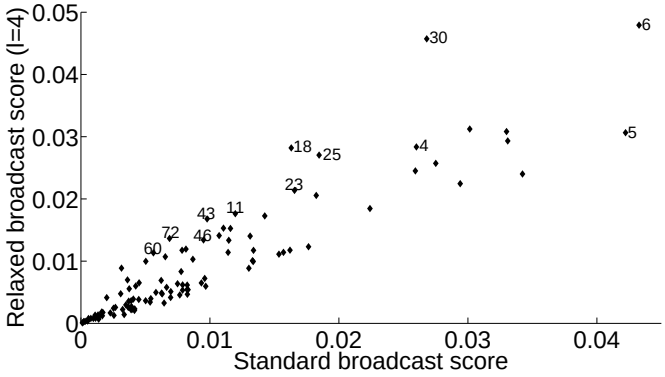


Figure 8: Relaxed ($l = 4$) vs. standard broadcast scores over all seven days.

"discrepancy" between the two methods is reduced when one increases l, the expected pathlength.

There are, undoubtedly, many efficient methods of implementing PageRank numerically. It is beyond the scope of this paper to discuss this matter. Therefore, we only note that in the implementation of PageRank for this paper we have used a numerical solver, which we repeat many times for the iterative version of PageRank. As far as we are aware, in PageRank, if one has to deal with very large matrices, the Power Method can be used instead. Compared to this, a step of the relaxed communicability method only involves $\mathcal{O}(l)$ matrix-vector multiplications and $\mathcal{O}(l)$ additions. This is repeated for each time step (snapshot) of the algorithm. Therefore, the implementation of relaxed communicability for this paper is more efficient than that of the iterative PageRank.

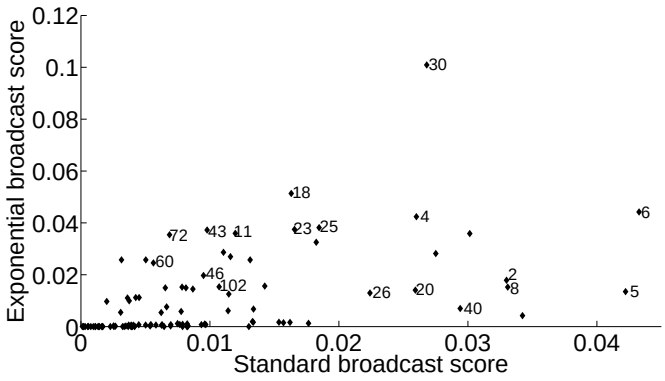


Figure 9: Exponential vs. standard broadcast scores over all seven days.

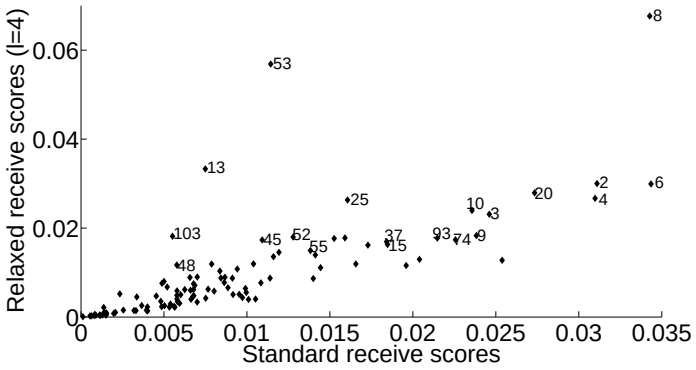


Figure 10: Relaxed receive scores ($l = 4$) vs. standard receive scores, computed over all seven days.

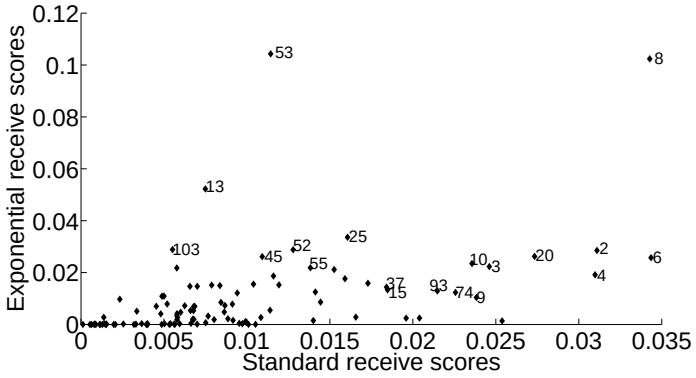


Figure 11: Exponential receive scores vs. standard receive scores, computed over all seven days.

3.2 Case-study 2: Twitter Mentions Network Data set

The data set comprised around a million tweets between UK users that contained mention of another UK user (via the @ sign). The nodes represented the users, and if user A's tweet contained

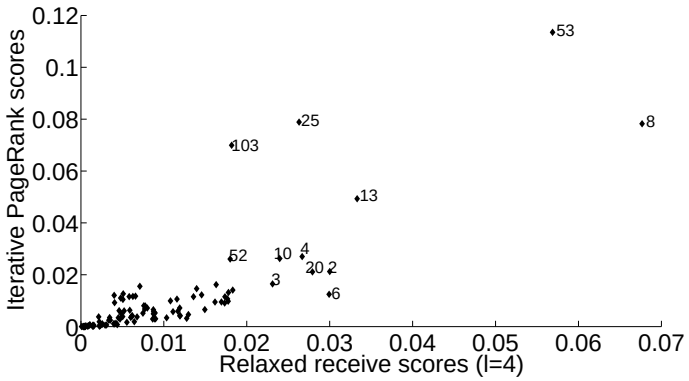


Figure 12: Iterative PageRank scores vs. relaxed receive scores ($l = 4$), computed over all seven days.

"@B", an edge between A and B was created. Only reciprocated edges were kept and multi-edges were ignored. All daily tweets were aggregated into a daily network, so we finished with 7 daily undirected graphs with 12408 nodes and around 2.7k edges on average. We computed both standard and relaxed communicability indices, both using rank obtained from communicability, and rank obtained from vitality based measure (deleting each node and computing the sum of differences for all the other nodes as described earlier).

3.3 Results

Although the computation of vitality measure is quite demanding (one needs to recompute communicability matrices for each node once) this is feasible as the daily networks are quite sparse.

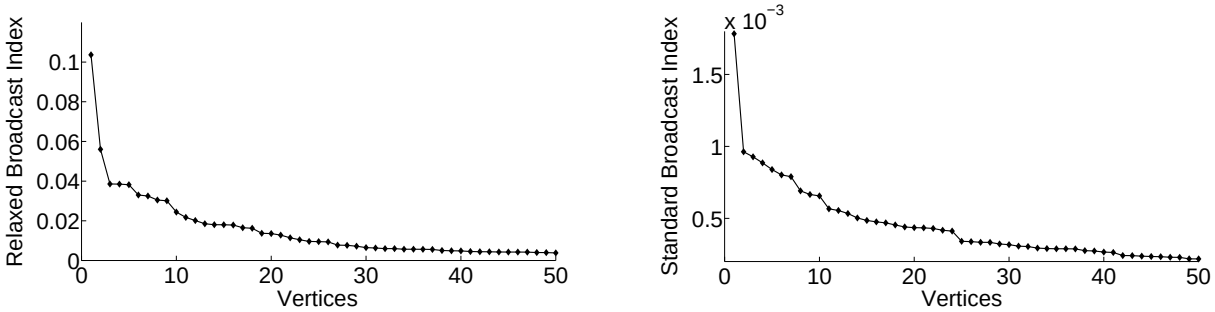


Figure 13: Twitter, top 50 vertices according to the ranking based on standard (left) and relaxed (right) vitality.

At 12 408 vertices and 7 time-steps, this collection contains relatively big, but not large networks. Their broadcast indices decrease quickly so we ranked the indices from largest to smallest with respect to broadcast and looked in more detail at the first fifty indices. On Fig 13 one can see the difference in ranking between the two methods. Several vertices that are ranked much higher in relaxed than in standard broadcast index correspond again to vertices with relatively small degrees and were picked up as they connect different communities (e.g. vertex ranked 39 in relaxed is ranked 278 in standard and has a degrees equal to $(2, 0, 0, 3, 1, 0, 0)$ respectively in 7 daily networks.)

3.4 Comparison with a recent study of the Enron email data

Here we consider the Enron email data set[7], from which we have extracted a subset consisting of emails sent between 151 Enron employees over a period of 1138 days, starting on 11 May 1999. In [9], a similar subset of the Enron data set was analysed as a static, aggregate network. Here we shall initially represent the data as an evolving network, in which each employee is a vertex and each day is a (separate) time step. We shall assume that there is a (directed) $i \mapsto j$ link at a certain time step, if i has sent at least one email to j on that day. We shall also remove all self-loops, that is, we shall assume that there are no links connecting a vertex with itself.

In [25], the same evolving network was used as a demonstration of standard communicability. The differences between in-degree and receive scores (resp. between out-degree and source scores) were noted, and it was emphasised that the direction of the flow of information over the (evolving) network was mainly caused by time. This was shown by comparing broadcast and receive scores of the symmetrised network with the corresponding scores for the directed network (see [25], §4.3 for further details). Because of this similarity between the directed and the symmetrised networks (see [25]), and also for simplicity, here we consider only the symmetrised network. In other words, in the directed evolving network that we defined above, for each $i \mapsto j$ link, we shall add the "reverse", $j \mapsto i$ link, if the latter is not already present. Our aim is to demonstrate some of the differences between standard and relaxed communicability.

In Fig 14(a) and Fig 14(b) we have plotted standard receive scores versus relaxed receive scores for $l = 2$ and $l = 7$. We can see from both figures that the correlation[8] between standard and relaxed scores is quite small: specifically, when $l = 2$ corr $= 0.08$, and when $l = 7$ corr $= 0.03$. Furthermore, when $l = 2$, amongst the top ten nodes (ranked according to standard and relaxed communicability) there are 4 nodes in common, while for $l = 7$ there are 3 nodes that overlap. However, there are 19 nodes in common amongst the top twenty nodes in both standard and relaxed communicability ($l = 2$ and $l = 7$) implying that both techniques are not entirely different from each other.

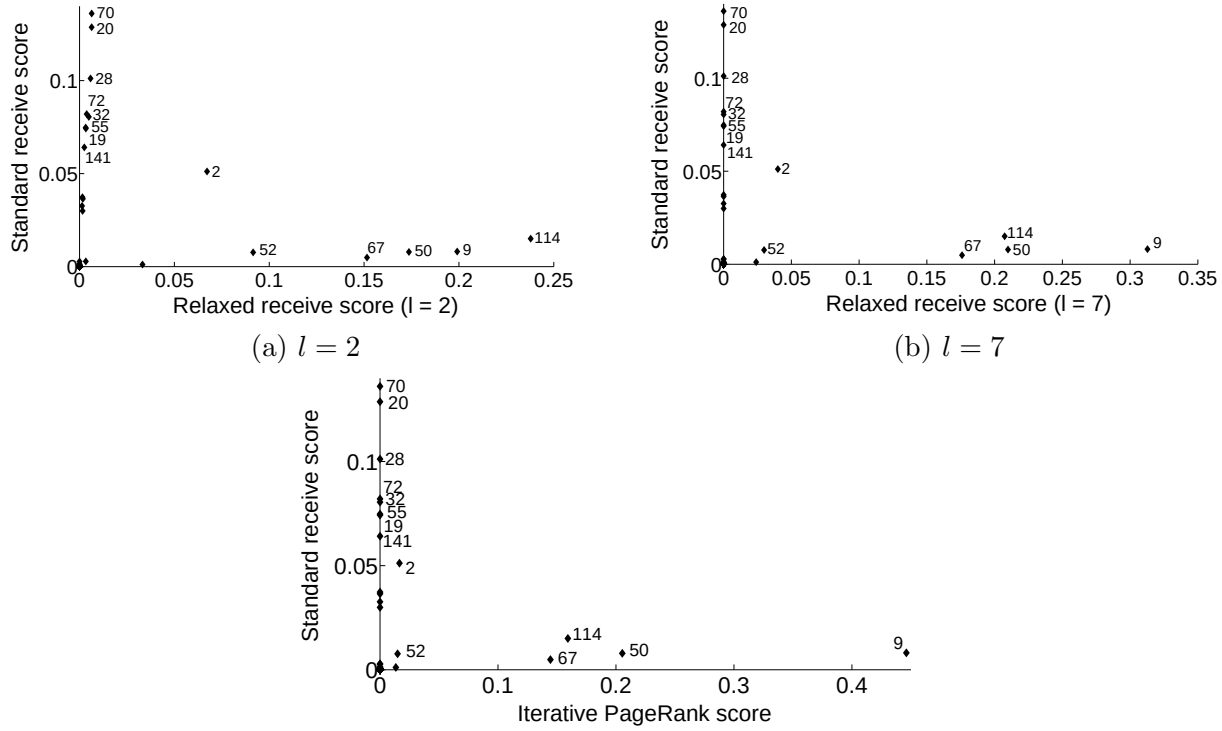
The small correlation between standard and relaxed receive scores for this data set is mainly due to the fact that the last time steps (days) have very low activity (see Fig 15). As we know, as a rule of thumb, receive scores tend to increase if nodes have been active during the last time steps. But, as we can see in this example, if the overall activity during a time step is low, compared to the maximum activity for the network, the influence of that time step is scaled down by standard communicability. This is why standard communicability ranks vertices 9, 50, 67 and 114 relatively low, while it ranks vertices $70, 20, 28, \ldots$ very high. From the left plot of Fig 15, we can see that vertex 70 is active only in days in which the overall activity is high, and inactive during (the last) days with low or no activity. (The adjacency matrix on day 1003 has the highest norm over all 1138 adjacency matrices.) Relaxed communicability, on the other hand, doesn't scale down so severely the communication on day 1138 (see the left and right plot on Fig 15) and this is why vertices 9, 50, 67 and 114 have very high ranks - they are the only active vertices on day 1138. Furthermore, these vertices' receive scores increase with l, especially that of vertex 9, whose degree on day 1138 is the highest (see Fig 14(b)).

Similarly to relaxed receive scores, iterative PageRank also scales each time step according to its own activity. Therefore, it is not unreasonable to expect "similar" results from these two methods. This can be confirmed by comparing Fig 14(a) and Fig 14(b) with Fig 14(c). The

[7]For more details, see `https://www.cs.cmu.edu/~enron/`

[8]In this subsection, by "correlation" (also denoted by corr) we mean the Pearson's correlation coefficient.

similarity between these figures is also reflected in the corresponding correlation between standard receive and iterative PageRank scores, corr = 0.02, and the overlapping sets of vertices ranked in the top ten and in the top twenty by both methods, 3 and 19, correspondingly. Moreover, the correlation between relaxed receive scores ($l = 7$) and iterative PageRank scores is 0.97 and their top ten and top twenty vertices match exactly. The difference between the two methods is that PageRank scales the transimissions sent by each active vertex according to the volume of activity of that vertex, while relaxed communicability scales all transmissions in a given time step uniformly, proportionally to the volume of activity in that time step. Another important feature of PageRank is that it accounts for "random" communication between any pair of vertices in the network, even if that communication is not given as a link in the adjacency matrix. In the context of email data (resp. mobile phone data) this could mean that we also take into account transmissions other than emails (resp. phone calls) that may have occured between the vertices.



(a) $l = 2$

(b) $l = 7$

(c) Iterative PageRank

Figure 14: *Top plots*: Enron data, standard receive scores vs. relaxed receive scores for different values of l. *Bottom plot*: Enron data, standard receive scores vs. Iterative PageRank scores

4 Conclusions

In this paper we investigated different versions of computing centrality rankings on an evolving network. We have presented a brief overview of three known approaches to computing centrality: standard [25], exponential [16] and relaxed [36] communicability. We have compared these methods on real-life evolving networks, obtained from mobile phone, Twitter and email communications. Apart from pointing out some of the main advantages of using, in certain situations, relaxed

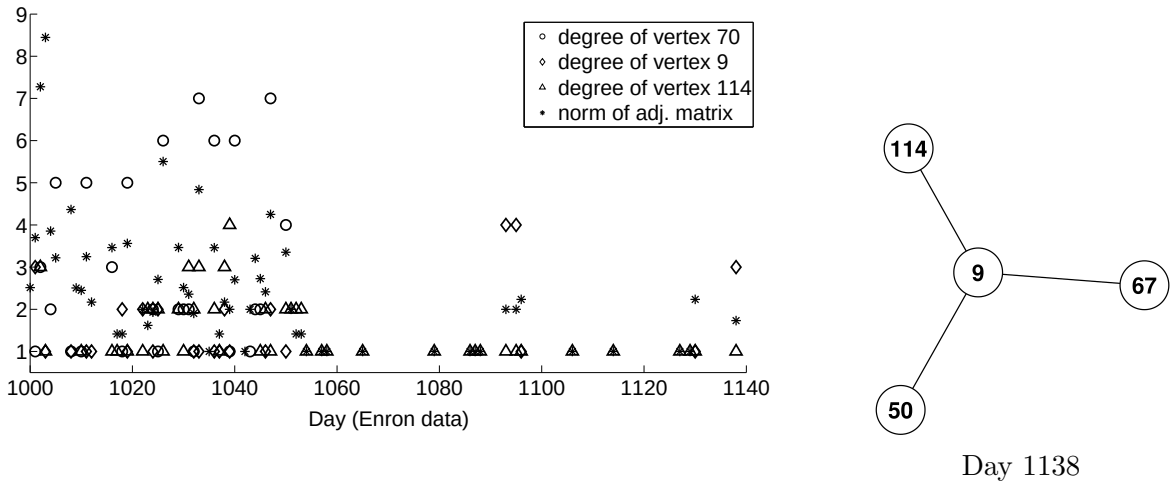


Figure 15: *Left plot*: Enron data, degrees of vertices 9 ($\Diamond$), 70 ($\bigcirc$) and 114 ($\triangle$) between days 1000 and 1138, compared with the 2-norm (sp. radius) of the corresponding adjacency matrix ($*$). *Right plot*: Enron data, the network on day 1138.

and exponential communicability over standard communicability, we also compared these methods with PageRank, and specifically, with an iterative (or dynamic) version of PageRank. We have also explored a *vitality measure*, which is a way of ranking vertices depending on the amount of change their communication abstinence would bring to the rest of the evolving network.

We think that relaxed communicability is a significant improvement of the "standard" version of communicability [25]. While the computation of communicability for small data sets is relatively simple and fast, when data sets are large (and most real data sets *are* large), an issue of handling very large matrices arises. Hence, to make such computations tractable (e.g. to avoid inverting large matrices) one can use a truncated Taylor expansion as an approximation. In the light of this remark, and also because in practice the path length of information flow through a network within a single snapshot is finite, we think that relaxed communicability models the cascades of communication within a single time step more realistically than standard communicability. Furthermore, in our opinion, relaxed communicability successfully deals with some important issues related to the choice of the attenuation factor, α, in standard communicability. Specifically, we found that choosing $\alpha = 1/\max_i \rho(A_i)$, where A_i is the adjacency matrix for snapshot i, tends to scale down the influence of snapshots with fewer edges (as the norm of A_i in that case would be smaller). Therefore, in relaxed communicability we update the attenuation factor at each time step (by scaling each adjacency matrix by its spectral radius) and are thus able to pick up nodes that are active in snapshots with (relatively) low activity. Moreover, (11) and (13) demonstrate a dependance from which one could determine the attenuation factor if the expected transmission length is known in advance. The other method, in which adjacency matrices are scaled independently from each other, is the iterative version of PageRank, but the two methods differ. We have shown in §3.1, in the discussion about Fig. 12, that the finite path length in relaxed communicability leads to significant differences in terms of the rankings produced by relaxed communicability and the iterative version of PageRank. In fact, we think that it is more realistic to define concepts such as *brokers* and/or *community bridges* in the context of the (expected) transmission length, if it is known.

We hope that a parametric approach that can be optimised according to a particular application

will be a useful addition to a standard evolving social network analysis toolbox, especially when the expected length of message transmission plays an important role, i.e. it is either given or can be approximated. An improvement to this approach could be when one allows the expected path length to vary over the time steps. We have shown that our method is better than Page Rank in identifying community bridges in a sense of giving them higher centrality ranking. Thus, for a given radius of centrality our method can be used to help identify important bridges in an evolving network.

Acknowledgments.

This work is funded by the RCUK Digital Economy programme via EPSRC grant EP/G065802/1 'The Horizon Hub'. We would like to thank Datasift for providing us with the Twitter data set, and Colin Singleton (CountingLab) and Des Higham (University of Strathclyde) for providing us with the Enron email data set.

References

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[2] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92:1170–1182, 1987.

[3] P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191 – 201, 2001.

[4] S. P. Borgatti and M. G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466 – 484, 2006.

[5] U. Brandes and T. Erlebach. *Network Analysis: Methodological Foundations (Lecture Notes in Computer Science)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and {ISDN} Systems*, 30(17):107 – 117, 1998.

[7] R. S. Burt. Brokerage and closure: An introduction to social capital. *Eur Sociol Rev*, 23(5):666–667, 2007.

[8] C. Castellano and R. Pastor-Satorras. Thresholds for epidemic spreading in networks. *Physical Review Letters*, 105:218701, 2010.

[9] A. Chapanond, M. Krishnamoorthy, and B. Yener. Graph theoretic and spectral analysis of enron email data. *Computational & Mathematical Organization Theory*, (11):265–281, 2005.

[10] S. Cioaba, D. Gregory, and V. Nikiforov. Extreme eigenvalues of nonregular graphs. *Journal of Combinatorial Theory, Series B*, 97(3):483 – 486, 2007.

[11] J. J. Crofts and D. J. Higham. Googling the brain: Discovering hierarchical and asymmetric network structures, with applications in neuroscience. *Internet Mathematics (Special Issue on Biological Networks)*, 2011.

[12] K. C. Das and P. Kumar. Some new bounds on the spectral radius of graphs. *Discrete Mathematics*, 281(1-3):149 – 161, 2004.

[13] N. Eagle and A. S. Pentland. Reality mining: sensing complex social systems. *Journal of Personal and Ubiquitous Computing*, 10:255–268, March 2006.

[14] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106:15274–15278, 2009.

[15] P. Erdös and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[16] E. Estrada and N. Hatano. Communicability in complex networks. *Physical Review E*, 77, 2008.

[17] E. Estrada and D. Higham. Network properties revealed through matrix functions. *SIAM Review*, (52):696–714, 2010.

[18] E. Estrada and J. A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Phys. Rev. E*, 71:056103, May 2005.

[19] F. Gantmacher. *The Theory of Matrices, Volume 2*. AMS Chelsea Publishing, 2000.

[20] R. Ghosh and K. Lerman. Parameterized centrality metric for network analysis. *Physical Review E*, 83(6):066118+, June 2011.

[21] E. Gilbert. Random graphs. *Ann. Math. Statist*, 30(4):1141–1144, 1959.

[22] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[23] P. Grindrod and D. J. Higham. Models for evolving networks: with applications in telecommunication and online activities. *IMA Journal of Management Mathematics*, 23:1–15, 2012.

[24] P. Grindrod and D. J. Higham. A matrix iteration for summarizing dynamic networks. *SIAM Review*, 55:118–128, 2013.

[25] P. Grindrod, M. C. Parsons, D. J. Higham, and E. Estrada. Communicability across evolving networks. *Phys. Rev. E*, 83:046120, Apr 2011.

[26] N. J. Higham. *Functions of Matrices*. SIAM, 2008.

[27] A. Jamaković, R. Kooij, P. Van Mieghem, and E. van Dam. Robustness of networks against viruses: the role of the spectral radius. In *Symposium on Communications and Vehicular Technology, 2006*, pages 35 –38, November 2006.

[28] L. Katz. A new index derived from sociometric data analysis. *Psychometrika*, 18:39–43, 1953.

[29] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[30] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245, 2010.

[31] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[32] N. Perra and S. Fortunato. Spectral centrality measures in complex networks. *Phys. Rev. E*, 78:036107, Sep 2008.

[33] T. Valente. Network interventions. *Science*, 337(6090):49–53, 2012.

[34] P. Van Mieghem. *Graph spectra for complex networks*. Cambridge University Press, 2011.

[35] P. Van Mieghem, J. C. Omic, and R. E. Kooij. Virus spread in networks. *IEEE/ACM Transactions on Networking*, 17(1):1–14, 2009.

[36] D. Vukadinović Greetham, Z. Stoyanov, and P. Grindrod. Centrality and spectral radius in dynamic communication networks. In D.-Z. Du and G. Zhang, editors, *Computing and Combinatorics*, volume 7936 of *Lecture Notes in Computer Science*, pages 791–800. Springer Berlin Heidelberg, 2013.

[37] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *In SRDS*, pages 25–34, 2003.

[38] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.