Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Christine Klymko Date

Centrality and Communicability Measures in Complex Networks:
Analysis and Algorithms

by

Christine Klymko
Doctor of Philosophy

Mathematics and Computer Science

Michele Benzi, Ph.D.
Advisor

Ronald Gould, Ph.D.
Committee Member

James Nagy, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Centrality and Communicability Measures in Complex Networks: Analysis and Algorithms

By

Christine Klymko
M.S. in Computer Science, Emory University, 2013
M.S. in Mathematics, Emory University, 2012
B.S. in Mathematics, Xavier University, 2008

Advisor: Michele Benzi, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Mathematics and Computer Science
2013

Abstract

Centrality and Communicability Measures in Complex Networks: Analysis and Algorithms
By Christine Klymko

Complex systems are ubiquitous throughout the world, both in nature and within man-made structures. Over the past decade, large amounts of network data have become available and, correspondingly, the analysis of complex networks has become increasingly important. One of the fundamental questions in this analysis is to determine the "most important" elements in a given network. Measures of node importance are usually referred to as *node centrality* and measures of how well two nodes are able to communicate with each other are referred to as the *communicability* between pairs of nodes. Many measures of node centrality and communicability have been proposed over the years. Here, we focus on the analysis and computation of centrality and communicability measures based on matrix functions.

First, we examine a node centrality measure based on the notion of *total communicability*, defined in terms of the row sums of the exponential of the adjacency matrix of the network. We argue that this is a natural metric for ranking nodes in a network, and we point out that it can be computed very rapidly even in the case of large networks. Furthermore, we propose a measure of the *total network communicability*, based on the total sum of node communicabilities, as a useful measure of the connectivity of the network as a whole.

Next, we compare various parameterized centrality rankings based on the matrix exponential and matrix resolvent with degree and eigenvector centrality. The centrality measures we consider are *exponential and resolvent subgraph centrality* (defined in terms of the diagonal entries of the matrix exponential and matrix resolvent, respectively), total communicability, and *Katz centrality* (defined in terms of the row sums of the matrix resolvent). We demonstrate an analytical relationship between these rankings and the degree and subgraph centrality rankings which helps to explain the observed robustness of these rankings on many real world networks, even though the scores produced by the centrality measures are not stable.

Finally, we propose an extension of subgraph centrality to directed networks, and we apply this extension to the problem of ranking hubs and authorities. The extension is achieved by *bipartization*, i.e., the directed network is mapped onto a bipartite undirected network with twice as many nodes in order to obtain a network with a symmetric adjacency matrix. We explicitly determine the exponential of this adjacency matrix in terms of the adjacency matrix of the original, directed network, and we give an interpretation of centrality and communicability in this new context, leading to a technique for ranking hubs and authorities. This method is compared to the well-known *HITS* algorithm as well as to several other ranking algorithms.

Centrality and Communicability Measures in Complex Networks: Analysis and Algorithms

By

Christine Klymko
M.S. in Computer Science, Emory University, 2013
M.S. in Mathematics, Emory University, 2012
B.S. in Mathematics, Xavier University, 2008

Advisor: Michele Benzi, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Mathematics and Computer Science
2013

Acknowledgements

There are many people without whom this dissertation would never have been completed. Foremost among them is my advisor, Dr. **Michele Benzi**. Throughout this process, you were a constant source of help and guidance. You have worked not only to help me complete my dissertation, but also to help establish me as a young researcher in the applied mathematics community and for this I am extremely grateful.

I would also like to thank my committee members, Dr. **Ronald Gould** and Dr. **James Nagy**. I truly appreciate the time and energy you have taken to help make my thesis the best possible.

Throughout the past five years, the Emory Mathematics and Computer Science Department has been my home base. I would like to thank all the faculty, staff, and other graduate students for making the department what it was.

Additionally, I would like to thank Prof. **Ernesto Estrada** (University of Strathclyde) for providing research ideas and datasets used throughout this thesis.

I would be remiss not thank to Mr. **Yu Wang** (Emory University) for performing the calculations with the Wikipedia graph in Chapter 2, and Dr. **Tamara Kolda** (Sandia National Laboratories) and Dr. **David Gleich** (Purdue University) for providing valuable suggestions for the work in Chapter 4.

In addition, I wish to thank Dr. **Tamara Kolda** and Dr. **Blair Sullivan** (NC State University) for providing me with opportunities to work at, respectively, Sandia and Oak Ridge National Labs during my summers. These internships have helped me grow as a researcher and introduced me to exciting new projects.

Finally, I would like to express my deepest gratitude to all my friends and family who supported me throughout this process. A special thanks goes out to my parents, Drs. **Paul and Nancy Klymko**. You both gave me more support than anyone has a right to expect.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and motivation

1.1.1 Introduction to complex networks

Complex systems are ubiquitous throughout the world, both in nature and within man-made structures. Over the past decade, large amounts of network data have become available and, correspondingly, the analysis of complex networks has become increasingly important. In response, researchers in diverse areas have focused their attention on such problems and the study of complex systems has begun to develop into a discipline in its own right [13, 21, 24, 40, 41, 84, 85]. Mathematicians, physicists, computer scientists, biologists, and social scientists (among others) have approached these problems from various angles[2, 28, 42, 43, 67, 88]. This large amount of research within diverse scientific communities has produced many interesting results, but there are still many fundamental questions that are not fully answered. Additionally, as available datasets become larger and larger, efficient data structures and computational methods become paramount.

One of the fundamental questions in network analysis is to determine the “most important" elements in a given network. The interpretation of what is meant by “important" can change from application to application. Measures of node importance are usually referred to as *node centralities*, and many centrality measures have been proposed, starting with the simplest of all, the node degree. This crude metric has the drawback of being too “local", as it does not take into effect the connectivity of the im-

mediate neighbors of the node under consideration. A number of more sophisticated centrality measures have been introduced that take into account the global connectivity properties of the network. These include various types of eigenvector centrality for both directed and undirected networks, betweenness centrality, and many others. Overviews of various centrality measures can be found in [13, 16, 21, 40, 72, 84, 85]. The centrality scores can be used to provide *rankings* of the nodes in the network. There are many different ranking methods in use (most of which depend on centrality measures), and many algorithms have been developed to compute these rankings. Information about the many different ranking schemes can be found in [8, 40, 67, 68, 72, 73, 74, 75].

1.1.2 Outline of thesis

This thesis investigates the use of matrix functions in the analysis of node centrality. In the remainder of this chapter, we will provide background information from various areas of mathematics. Sections 1.2 and 1.3 present basic concepts from graph theory and linear algebra that will be used in this thesis. Section 1.4 discusses properties found in many complex networks. Section 1.5 provides an overview of commonly used centrality and communicability measures, including HITS and PageRank while Section 1.6 discusses the comparison of rankings produced by different centrality measures. Section 1.7 provides details on iterative methods that are used in this thesis to approximate certain types of centrality scores.

Chapter 2 examines the use of a node centrality measure based on the notion of *total communicability*, defined in terms of the row sums of the exponential of the adjacency matrix of the network. We argue that this is a natural metric for ranking nodes in a network, and we point out that it can be computed very rapidly even in the case of large networks. Furthermore, we propose the total sum of node communicabilities as a useful measure of network connectivity. Extensive numerical studies are conducted in order to compare this centrality measure with the closely related ones of subgraph

centrality and Katz centrality on both synthetic and real-world networks.

In Chapter 3, we demonstrate an analytical relationship between rankings produced by parameterized centrality measures based on the matrix exponential and resolvent with those produced by degree and eigenvector centrality. The centrality measures considered are exponential and resolvent subgraph centrality (defined in terms of the diagonal entries of the matrix exponential and matrix resolvent, respectively), total communicability, and Katz centrality (defined in terms of the row sums of the matrix exponential and resolvent). This analysis helps to explain the observed robustness of these rankings on many real world networks, even though the scores produced by the centrality measures are not stable. A number of numerical experiments on both directed and undirected networks are conducted to analyze how this relationship affects node rankings.

In Chapter 4, we propose an extension of the measures of subgraph centrality and communicability to directed networks, and we apply them to the problem of ranking hubs and authorities. The extension is achieved by *bipartization*, i.e., the directed network is mapped onto a bipartite undirected network with twice as many nodes in order to obtain a network with a symmetric adjacency matrix. We explicitly determine the exponential of this adjacency matrix in terms of the adjacency matrix of the original, directed network, and we give an interpretation of centrality and communicability in this new context, leading to a technique for ranking hubs and authorities. The matrix exponential method for computing hubs and authorities is compared to the well known HITS algorithm, both on small artificial examples and on more realistic real-world networks. A few other ranking algorithms are also discussed and compared with our technique.

Finally, Chapter 5 contains concluding remarks and guidelines for future work.

The work presented in Chapters 2 and 3 is based on [10] and [8], respectively.

1.2 Basic concepts from graph theory

In this section, we provide an overview of the basic concepts, definitions, and terminology from graph theory that will be used throughout the thesis. A more comprehensive presentation can be found in [15, 32].

A *graph* $G = (V, E)$ is formed by a set of nodes (vertices) V and edges (links) E formed by unordered pairs of vertices. Every network is naturally associated with a graph $G = (V, E)$ where $|V| = n$ is the number of objects in the network and E ($|E| = m$) is the collection of connections between objects, $E = \{(i, j) \,|$ there is an edge between node i and node $j\}$. Nodes represent the elements that make up the network and edges represent various types of interactions among network elements. The *degree* d_i of a vertex i is the number of edges incident to i. A graph is *k-regular* if every node has degree $d_i = k$.

A directed graph, or *digraph*, $G = (V, E)$ is formed by a set of vertices V and edges E formed by ordered pairs of vertices. That is, $(i, j) \in E \not\Rightarrow (j, i) \in E$. In the case of digraphs, which model directed networks, there are two types of degree. The *in-degree* of node i, d_i^{in}, is given by the number of edges which point to i. The *out-degree*, d_i^{out}, is given by the number of edges pointing out from i.

The first result in graph theory, which was also the start of graph theory as a proper mathematical discipline, related the number of nodes, n, in an undirected graph to the number of edges, m. The following result, called the *handshaking lemma*, is from Leonhard Euler [49] and has since been extended to directed graphs.

Lemma 1.2.1 Let G be a graph with n nodes and m edges. Then,

(i) if G is undirected, then

$$\sum_{i=1}^{n} d_i = 2m,$$

(ii) if G is directed, then

$$\sum_{i=1}^{n} d_i^{in} = \sum_{i=1}^{n} d_i^{out} = m.$$

A *subgraph* of G is a graph $H = (V', E')$ such that $V' \subseteq V$ and $E' \subseteq \{(i,j) \in E \,|\, i,j \in V'\}$. The subgraph H is said to be an *induced subgraph* if $E' = \{(i,j) \in E \,|\, i,j \in V'\}$, that is if all edges possible are present. A subgraph H is said to be *maximal* in regards to a property if adding any additional nodes or edges to H would result in that property no longer holding.

A *walk* of length k in G is a set of nodes $i_1, i_2, \ldots i_k, i_{k+1}$ such that for all $1 \leq l \leq k$, there is an edge between i_l and i_{l+1} (or a directed edge from i_l to i_{l+1} in the case of a digraph). A *closed walk* is a walk where $i_1 = i_{k+1}$. A *path* is a walk with no repeated nodes. A *cycle* is a closed path. A graph G is *walk-regular* if the number of closed walks of length $k \geq 0$ starting at a given node is the same for every node in the graph. A graph is *simple* if it has no *loops* (edges from a node i to itself), no multiple edges, and unweighted edges.

The *distance*, $d(i,j)$, between nodes i and j in a graph is given by the length of the shortest path between them. If there is no path between two nodes, they are considered to be at an infinite distance. The *diameter* of a graph, d_G is given by the maximum shortest path length in the graph: $d_G = \max_{i \neq j} d(i,j)$. In an undirected graph, the distance function forms a *metric*. That is, the following lemma holds.

Lemma 1.2.2 Given an undirected graph G, the *distance function* $d : V \times V \to \mathbb{R}$ given by $d(i,j) = \{$the length of the shortest path between i and $j\}$ satisfies the following properties and is therefore a metric: for all nodes i, j, and k,

(i) $d(i,j) \geq 0$ and $d(i,j) = 0 \Leftrightarrow i = j$,

(ii) $d(i,j) = d(j,i)$,

(iii) $d(i,j) \leq d(i,k) + d(k,j)$.

In a directed graph, there is no guarantee that distances are symmetric and, thus, the distance function does not define a metric.

An undirected graph is *connected* if there exists a path between every pair of nodes. A digraph is *weakly connected* if the underlying undirected graph is connected. It is *strongly connected* if, for every pair of vertices i and j, there is both a directed walk from i to j and one from j to i. If an undirected graph G is not connected, then G can be partitioned into its *connected components*. The number of connected components is the number of maximal connected subgraphs. If G is directed, then we can distinguish between the *strongly connected components* and the *weakly connected components*. When a graph is partitioned into its (weakly) connected components, every node and edge is in exactly one component. When a digraph is partitioned into its strongly connected components, every node is in exactly one component and every edge is in at least one component. Often, in the analysis of complex networks we are concerned with the largest connected component of an undirected network or the largest strongly connected component of a directed network.

Often, we consider graphs with unweighted edges, but depending on the type of interaction edges can be weighted or unweighted. Weights on edges can be used to measure the amount of traffic between two nodes or other properties of data, such as the vulnerability of the connection to attack or disruption or a physical distance between two nodes. In the case of unweighted networks, we consider the weight of each edge to be equal to one.

Every graph is naturally associated with an *adjacency matrix*. The adjacency matrix of an unweighted graph G is given by the matrix $A \in \mathbb{R}^{|V|\times|V|}$ defined in the following way:

$$A = (a_{ij}); \quad a_{ij} = \begin{cases} 1, & \text{if } (i,j) \text{ is an edge in } G, \\ 0, & \text{else.} \end{cases}$$

If G is weighted, then:

$$A = (a_{ij}); \quad a_{ij} = \begin{cases} w_{ij}, & \text{if } (i,j) \text{ is an edge in } G, \\ 0, & \text{else} \end{cases}$$

where w_{ij} is the weight of edge (i,j). If G contains no loops, then A will have zeros along the diagonal. If G is undirected, A will be a symmetric matrix.

If A is the adjacency matrix of a network with unweighted edges, then $[A^k]_{ij}$ counts the number of walks of length k between nodes i and j. The diagonal entries, $[A^k]_{ii}$ count the number of closed walks of length k centered at node i, although some of these walks can be described as "illogical." For example, the walk $i \to j \to i \to j \to i$ is a closed walk of length 4 centered at node i (see [32, 47] for more details).

1.3 Basic concepts from linear algebra

This section contains an overview of some basic concepts from linear algebra. A more comprehensive guide can be found in many places, including [78, 90].

Given a matrix $A \in \mathbb{R}^{n \times n}$, let the *eigenvalues* of A be given by $\lambda_1, \lambda_2, \ldots, \lambda_n$ and the *eigenvectors* be given by $\mathbf{q}_1, \mathbf{q}_2, \ldots \mathbf{q}_n$ where $\mathbf{q}_i$ is the eigenvector associated with eigenvalue λ_i. If the eigenvalues of A are real, we label them in non-increasing order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. If A has complex eigenvalues, they are labeled with non-increasing modulus: $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$.

The *spectrum* of A is $\sigma(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and the *spectral radius* of A is given by $\rho(A) = \max_i(|\lambda_i|)$. The *spectral gap* of A is given by the difference between the two largest eigenvalues of A: $|\lambda_1 - \lambda_2|$. In the case of complex eigenvalues, it is given by the difference between the moduli of the two largest eigenvalues: $|\lambda_1| - |\lambda_2|$. The *range* of A is given by the span of the column space of A: $\mathcal{R}(A) = \{v \,|\, Ax = v \text{ for some } x \in \mathbb{R}^n\}$. The *null space* of A is given by $\mathcal{N}(A) = \{x \,|\, Ax = 0\}$.

The matrix A is said to be *positive* (written $A > 0$) if $a_{ij} > 0$ for all $1 \leq i, j \leq n$. It is said to be *non-negative* (written $A \geq 0$) if $a_{ij} \geq 0$ for all $1 \leq i, j \leq n$. The adjacency matrix of a graph G is non-negative if G is unweighted or if all the edge weights of G are positive.

Definition 1 A matrix A is said to be *reducible* if there exists a permutation matrix P such that

$$A = P^T AP = \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix}$$

where X and Z are both square [78, p. 671]. If A is not reducible, it is *irreducible*.

The adjacency matrix of an undirected graph G is irreducible if and only if G is connected. The adjacency matrix of a directed graph is irreducible if and only if G is strongly connected.

A non-negative, irreducible matrix A is said to be *primitive* if there is only one eigenvalue λ with $|\lambda| = \rho(A)$. Otherwise, it is *imprimitive*.

If A is a non-negative, irreducible matrix, the *Perron-Frobenius Theorem* gives useful information on the eigenvalues and eigenvectors of A:

Theorem 1.3.1 (Perron-Forbenius Theorem, [78, p. 673]) If $A \in \mathbb{R}^{n \times n}$ is non-negative and irreducible, then the following are true:

(i) $r = \rho(A) \in \sigma(A)$ and $r > 0$.

(ii) the algebraic multiplicity of r is 1.

(iii) There exists an eigenvector $\mathbf{x} > 0$ such that $A\mathbf{x} = r\mathbf{x}$.

(iv) The unique vector defined by $A\mathbf{p} = r\mathbf{p}$, $p > 0$, and $\|p\|_1 = 1$ is called the *Perron vector*. There are no non-negative eigenvectors of A except for positive multiples of $\mathbf{p}$, regardless of the eigenvalue.

If G is a connected, undirected graph with adjacency matrix A, then $\lambda_1 > \lambda_2$ by the Perron-Frobenius theorem. Since A is a symmetric, real-valued matrix, we can decompose A into $A = Q\Lambda Q^T$ where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ with $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$ and $Q = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]$ is orthogonal.

If G is a strongly connected digraph with adjacency matrix A, then, by the Perron-Frobenius theorem, $\lambda_1 = r$ is a simple eigenvalue of A and both the left and right eigenvectors of A associated with λ_1 are positive (as A^T is also irreducible). If G is also *diagonalizable*, then there exists an invertible matrix X such that $A = X\Lambda X^{-1}$ where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ with $\lambda_1 \geq |\lambda_i|$ for $2 \leq i \leq n$, $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, and $(X^{-1})^* = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$. The left eigenvector associated with λ_i is $\mathbf{y}_i$ and the right eigenvector associated with λ_i is $\mathbf{x}_i$. In the case where G is not diagonalizable, A can be decomposed using its Jordan Canonical form:

$$A = XJX^{-1} = X \begin{pmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \hat{J} \end{pmatrix} X^{-1}$$

where J is the Jordan matrix of A, except that we place the 1×1 block corresponding to λ_1 first. More information about the eigenvalues and eigenvectors associated with graphs can be found in [30].

An *M-matrix* is a real, non-singular matrix $A = (a_{ij})$ such that $a_{ij} \leq 0$ for all $i \neq j$ and $A^{-1} \geq 0$. Many properties of non-singular M-matrices are known (see [86] for a more comprehensive overview).

Theorem 1.3.2 Let $A \in \mathbb{R}^{n \times n}$ have positive diagonal entries and non-positive off-diagonal entries. Then, the following are equivalent [86]:

(i) A is a non-singular M-matrix.

(ii) there exists a matrix $B \geq 0$ and a real number $r > \rho(B)$ such that $A = rI - B$.

(iii) the real part of λ is positive for all $\lambda \in \sigma(A)$.

(iv) A^{-1} exists and is non-negative.

We now turn to functions of a matrix.

Definition 2 Given a scalar function f, f is said to be defined on the spectrum of A if the values

$$f^{(j)}(\lambda_i), \quad j = 0 : n_i - 1, \quad i = 1 : s$$

exist, where s is the number of distinct eigenvalues of A and n_i is the geometric multiplicity of λ_i. These are called the values of the function f on the spectrum of A ([65, p. 3]).

Suppose f is defined on the spectrum of A. If A is symmetric or diagonalizable, then we define $f(A) = Qf(\Lambda)Q^T$ or $f(A) = Xf(\Lambda)X^{-1}$, where

$$f(\Lambda) = \mathrm{diag}(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)).$$

If A is not diagonalizable, then

$$f(A) = \sum_{i=1}^{s} \sum_{j=0}^{n_i - 1} \frac{f^{(j)}(\lambda_i)}{j!} (A - \lambda_i I)^j G_i$$

where G_i is the oblique projector onto $\mathcal{N}((A-\lambda_i I)^{n_i})$ along $\mathcal{R}((A-\lambda_i I)^{n_i})$ [78, p. 603].

In this thesis, we are primarily concerned with two matrix functions: the *matrix exponential* and the *matrix resolvent*. The (parameterized) matrix exponential is given by $e^{\beta A}$, $\beta > 0$. In many applications, the parameter β is set to one, corresponding to the unscaled matrix exponential. The eigenvalues of $e^{\beta A}$ are given by $e^{\beta\lambda_1}, e^{\beta\lambda_2}, \ldots, e^{\beta\lambda_n}$. The power series expansion of $e^{\beta A}$ is given by:

$$e^{\beta A} = I + \beta A + \frac{\beta^2 A^2}{2!} + \cdots + \frac{\beta^k A^k}{k!} + \cdots = \sum_{k=0}^{\infty} \frac{\beta^k A^k}{k!}. \tag{1.1}$$

As $[A^k]_{ij} = [A^k]_{ji}$ counts the number of walks of length k between nodes i and j, $[e^A]_{ij}$, counts the total number of walks from node i to node j, penalizing longer walks by scaling walks of length k by the factor $\frac{\beta^k}{k!}$.

The matrix resolvent is given by $(I - \alpha A)^{-1}$, $0 < \alpha < \frac{1}{\lambda_1}$. It has eigenvalues $\frac{1}{1-\alpha\lambda_i}$ where $\lambda_i \in \sigma(A)$. Similarly to the matrix exponential, the entries of the matrix resolvent count the number of walks in the network, penalizing longer walks. This can be seen by considering the power series expansion of $(I - \alpha A)^{-1}$:

$$(I - \alpha A)^{-1} = I + \alpha A + \alpha^2 A^2 + \cdots + \alpha^k A^k + \cdots = \sum_{k=0}^{\infty} \alpha^k A^k. \tag{1.2}$$

Here, $[(I-\alpha A)^{-1}]_{ij}$, counts the total number of walks from node i to node j, weighting walks of length k by α^k. The bounds on α ($0 < \alpha < \frac{1}{\lambda_1}$) ensure that the matrix $I - \alpha A$ is invertible and that the power series in (1.2) converges to its inverse. The bounds on α also force $(I - \alpha A)^{-1}$ to be nonnegative, as $I - \alpha A$ is a nonsingular M-matrix.

A matrix A is *Hermitian* if it is equal to its own conjugate transpose, i.e $A^T = A^*$. Every real, symmetric matrix is Hermitian. A matrix is *upper Hessenberg* if all the entries below the first subdiagonal are zero.

1.4 Properties of complex networks

Networks can be used to describe and analyze many different types of interactions, from those between people (social networks), to the flow of goods across an area (transportation networks), to links between websites (the WWW graph), to thermodynamic interaction between particles (statistical mechanics), and so forth. Every network is associated with both a graph and an adjacency matrix, allowing the use of tools from graph theory and linear algebra to be used in their analysis. Although there is no precise definition of a complex network, there are a number of properties that complex networks tend to have in common. More information about complex networks can be

found in many places, including [40, 83, 84]

Complex networks are fundamentally different from random networks generated according to the Erdös-Rényi model [35, 36]. There are two closely related variants of this model: in the $G(n, m)$ model, a graph is chosen uniformly at random from the set of all graphs with n nodes and m edges. In the $G(n, p)$ model, a graph with n nodes is constructed by randomly connecting each pair of nodes with probability p. If p is set to $p = \binom{n}{2}m$, the two models behave similarly as n tends to infinity. Networks produced by Erdös-Rényi random models tend to have a Poisson degree-distribution. That is, the majority of nodes tend to have a degree close to the average degree, with relatively few outliers.

However, this is not the case for complex networks. It has been shown that many real-world networks follow a power law degree distribution, i.e. that the fraction of nodes of degree k is approximately $P(k) \approx k^{-\gamma}$ where often $2 < \gamma < 3$ [5, 51]. Others have a mixed degree distribution which follows an exponential degree distribution for low-degree nodes and changes into a power law distribution once the node degree increases sufficiently [80, 99]. Networks with these types of degree-distributions are often called *scale-free* networks. Power law degree-distributions have been observed in many types of networks including biological networks (such as protein-protein interaction and gene-regulation networks) [11, 42, 43], the internet and world-wide-web [3, 5, 51], and social networks [26]. More information on scale-free networks can be found in [24, 59].

The *clustering coefficient* of a network measures the propensity of the nodes in the network to form triangles. The clustering coefficient of a node i is given by

$$C_i = \frac{2|C_3(i)|}{d_i(d_i - 1)},$$

where $|C_3(i)|$ is the number of triangles in which node i is involved. This measures the

percentage of the paths of length two (wedges) centered at node i which have closed into triangles [95]. The clustering coefficient of the network as a whole is given by

$$\bar{C} = \frac{1}{n} = \sum_{i=1}^{n} C_i,$$

which gives the average node clustering coefficient across the network. A variation on the clustering coefficient, known as the *global clustering coefficient* or *network transitivity* was introduced in [82]:

$$C = \frac{3|C_3|}{|P_2|},$$

where $|C_3|$ is the number of triangles in the network and $|P_2|$ is the number of paths of length two in the network. Networks formed using the Erdös-Rényi random models have very low clustering coefficients while many real world networks tend to exhibit high clustering coefficients.

One property common to both Erdös-Rényi random networks and many real-world networks is the *small-world property* [95]. Informally, the existence of the small-world property in a network means that all the nodes in the network are only a relatively short distance from each other, creating a "small world" for the nodes to exist inside. More formally, it states that the *average path length*,

$$\bar{l} = \frac{1}{n(n-1)} \sum_{i,j \in V} d(i,j),$$

also known as the *mean shortest distance* between pairs of nodes, is small.

A variety of alternative generative models for complex networks have been proposed to address the deficiencies of the Erdös-Rényi random models. These include the preferential attachment model and variants [5, 71], the small-world model [95], and the Block Two-Level Erdös-Renyi (BTER) model [69].

1.5 Measures of centrality and communicability

Many measures of node centrality have been developed and used over the years. The higher the measure of node centrality, the more "important" a given node is considered to be in the network. In this section, we provide a brief overview of some of the most commonly used centrality measures. More information about various centrality measures can be found in [19, 40, 84, 94]

One of the simplest centrality measures is *degree centrality* [57]. In an undirected network, the degree centrality of node i is given by $C_d(i) = d_i = [A\mathbf{1}]_i$, where $\mathbf{1}$ is the vector of all ones. In a directed network, it is necessary to distinguish between the in-degree centrality, $C_d^{in}(i) = d_i^{in} = [\mathbf{1}^T A]_i$, and the out-degree centrality, $C_d^{out}(i) = d_i^{out} = [A\mathbf{1}]_i$. Degree centrality is a local measure of node importance, counting the number of nodes in the immediate neighborhood of a given node.

The *eigenvector centrality* of node i in a directed network is given by $C_{ev}(i) = \mathbf{q}_1(i)$ [16]. It gives the limit as k goes to infinity of the percentage of walks of length k which start at node i [30]. Thus, in contrast to degree-centrality, eigenvector centrality measures the global influence of nodes in the network. Variations on eigenvector centrality have been proposed to measure slightly different aspects of node importance [17]. In a directed network, the eigenvector centrality can be defined by both the dominant left eigenvector and the dominant right eigenvector. These can be used to rank nodes as authorities and hubs, respectively (see section 1.5.1 for definitions of hubs and authorities).

The *closeness centrality* [57] of node i is given by

$$CC(i) = \frac{n-1}{\sum_{j \neq i} d(i,j)}.$$

It measures the overall closeness of node i to the rest of the network.

Another commonly used centrality measure is *betweenness centrality* [56]. The be-

tweenness centrality of node i is given by

$$BC(i) = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

where σ_{jk} is the total number of shortest paths between nodes j and k and $\sigma_{jk}(i)$ is the number of those shortest paths which pass through node i. Thus, the betweenness centrality of node i measures how important i is for fast communication across the network as a whole. Variations of betweenness centrality have also been proposed [22].

The matrix resolvent $(I - \alpha A)^{-1}$, $0 < \alpha < \frac{1}{\lambda_1}$, was first used to rank nodes in a network in the early 1950s, when Katz used the column sums to calculate node importance [67]. The Katz centrality score of node i is given by $K_i(\alpha) = [(I-\alpha A)^{-1}\mathbf{1}]_i$. This counts the total number of walks between node i and every other node in the network, weighting walks of length k by a penalization factor of α^k. Closely related to Katz centrality is *resolvent subgraph centrality* [47, 48]. The resolvent subgraph centrality of node i is given by $RC_i(\alpha) = [(I - \alpha A)^{-1}]_{ii}$. This counts the number of closed walks centered at node i, again weighting walks of length k by α^k.

A more commonly used form of subgraph centrality is *exponential subgraph centrality* [47, 48]. The exponential subgraph centrality of node i is given by $SC_i(\beta) = [e^{\beta A}]_{ii}$. This counts the number of closed walks centered at node i, weighting walks of length k by $\frac{\beta^k}{k!}$. In Chapter 2, we introduce a new centrality measure based on exponential subgraph communicability [10, 45, 46].

Many centrality measures were originally developed for use on undirected networks and later needed to be extended for use on directed networks. In Sections 1.5.1 and 1.5.2, we describe HITS and PageRank, two rankings methods that were developed specifically for directed networks. Both HITS and PageRank were originally developed as web search algorithms. However, as they provide a ranking of the nodes in directed

networks, they can also be used as centrality measures in general digraphs.

The *communicability* between a pair of nodes i and j measures how well the two nodes can exchange information with each other. A (relatively) large communicability between a pair of nodes i and j indicates that information flows more easily between those two nodes than between pairs of nodes with lower communicability. In other words, a low communicability indicates that the two nodes cannot easily exchange information. Network communicability can also be interpreted in terms of the correlations between different components of physical systems; see, e.g., [46].

Two of the most commonly used communicability measures are *(exponential) subgraph centrality* and *resolvent subgraph centrality* [45, 47]. The subgraph communicability between nodes i and j is given by $[e^{\beta A}]_{ij}$, $\beta > 0$ (note that in the case of an undirected network, A is symmetric and $[e^{\beta A}]_{ij} = [e^{\beta A}]_{ji}$). This counts the total number of walks between nodes i and j, weighting walks of length k by $\frac{1}{k!}$. The resolvent subgraph communicability between nodes i and j is given by $[(I-\alpha A)^{-1}]_{ij}$, $0 < \alpha < \frac{1}{\lambda_1}$. As in the case of exponential subgraph centrality, this counts the total number of walks between nodes i and j, but here walks of length k are weighted by α^k.

1.5.1 HITS

The classical *Hypertext Induced Topics Search* (HITS) algorithm, first introduced by J. Kleinberg in [68] is based on the idea that in the World Wide Web, and indeed in all document collections which can be represented by directed networks, there are two types of important nodes: *hubs* and *authorities*. Hubs are nodes which point to many nodes of the type considered important. Authorities are these important nodes. From this comes a circular definition: good hubs are those which point to many good authorities and good authorities are those pointed to by many good hubs.

Thus, the HITS ranking relies on an iterative method converging to a stationary solution. Each node i in the network is assigned two non-negative weights, an *authority*

weight x_i and a *hub weight* y_i. To begin with, each x_i and y_i is given an arbitrary nonzero value. Then, the weights are updated in the following ways:

$$x_i^{(k)} = \sum_{j:(j,i)\in E} y_j^{(k-1)} \quad \text{and} \quad y_i^{(k)} = \sum_{j:(i,j)\in E} x_j^{(k)} \quad \text{for} \quad k = 1, 2, 3... \tag{1.3}$$

The weights are then normalized so that $\sum_j (x_j^{(k)})^2 = 1$ and $\sum_j (y_j^{(k)})^2 = 1$.

The above iterations occur sequentially and it can be shown that, under mild conditions, both sequences of vectors $\{\mathbf{x}^{(k)}\}$ and $\{\mathbf{y}^{(k)}\}$ converge as $k \to \infty$. In practice, the iterative process is continued until there is no significant change between consecutive iterates.

This iteration sequence shows the natural dependence relationship between hubs and authorities: if a node i points to many nodes with large x-values, it receives a large y-value and, if it is pointed to by many nodes with large y-values, it receives a large x-value.

In terms of matrices, the equation (3.1) becomes: $\mathbf{x}^{(k)} = A^T\mathbf{y}^{(k-1)}$ and $\mathbf{y}^{(k)} = A\mathbf{x}^{(k)}$, followed by normalization in the 2-norm. This iterative process can be expressed as

$$\mathbf{x}^{(k)} = c_k A^T A\mathbf{x}^{(k-1)} \quad \text{and} \quad \mathbf{y}^{(k)} = c'_k AA^T\mathbf{y}^{(k-1)}, \tag{1.4}$$

where c_k and c'_k are normalization factors. A typical choice for the initizialization vectors $\mathbf{x}^{(0)}$, $\mathbf{y}^{(0)}$ would be the constant vector

$$\mathbf{x}^{(0)} = \mathbf{y}^{(0)} = [1/\sqrt{n},\, 1/\sqrt{n},\, \ldots, 1/\sqrt{n}],$$

see [52]. Hence, HITS is just an iterative power method to compute the dominant eigenvector for A^TA and for AA^T. The authority scores are determined by the entries of the dominat eigenvector of the matrix A^TA, which is called the *authority matrix* and the hub scores are determined by the entries of the dominant eigenvector of AA^T,

called the *hub matrix*. Recall that the eigenvalues of both A^TA and AA^T are the squares of the singular values of A. Also, the eigenvectors of A^TA are the right singular vectors of A, and the eigenvectors of AA^T are the left singular vectors of A. See Algorithm 1 in Appendix A for a pseudocode implementation of the HITS algorithm.

A number of variations on the HITS algorithm have been proposed, for example [8, 20, 33, 72]

1.5.2 PageRank

The *PageRank* algorithm [23] is perhaps the most famous (and lucrative) algorithm in the world. It was developed to rank authoritative webpages in search engine results. Although PageRank, and its variant Reverse PageRank, described below, were originally developed to rank webpages, they have since been used to rank nodes in many diverse types of networks [11, 12, 55, 70, 81, 96].

In the PageRank algorithm, the importance of a webpage is determined by "votes," i.e., the number of links to that webpage. The main idea is that votes from more important sites count more than votes from less important sites and the significance of any vote is weighed by how many votes the "voting" webpage makes. At its most basic, given a directed network, the PageRank of node i is given by

$$PR_i = \sum_{\{j \,|\, (j,i) \in E\}} \frac{PR_j}{d_j^{out}}.$$

As the PageRank of a node depends on the PageRank of its neighbors, the rankings are calculated in an iterative manner. To begin, each node is given an initial ranking. Often, $PR_i^{(0)} = \frac{1}{n}$. Then,

$$PR_i^{(k)} = \sum_{\{j \,|\, (j,i) \in E\}} \frac{PR_j^{(k-1)}}{d_j^{out}} \quad \text{for} \quad k = 1, 2, 3... \tag{1.5}$$

Setting $\boldsymbol{\pi}^{(k)T} = (PR_1^{(k)}, PR_2^{(k)}, \ldots, PR_n^{(k)})$, (1.5) becomes

$$\boldsymbol{\pi}^{(k)T} = \boldsymbol{\pi}^{(k-1)T} P \tag{1.6}$$

where $P = \begin{cases} \frac{1}{|d_i^{out}|}, \text{if } (i,j) \in E \\ 0, else. \end{cases}$

Equation (1.6) almost corresponds to the power method applied to a finite state space *Markov chain* with *transition probability matrix* P. A Markov chain is a stochastic process that satisfies the *Markov property* [78, p. 687]. That is, given a family of random variables $\{X_t\}$ in which every X_t has the same range $\{S_1, S_2, \ldots, S_n\}$ (called the *state space*), the following holds:

$$P(X_{t+1} = S_j \,|\, X_t = S_{i_t}, X_{t-1} = S_{i_{t-1}}, \ldots, X_0 = S_{i_0}) = P(X_{t+1} = S_j \,|\, X_t = S_{i_t}).$$

This implies that the probability of entering state S_j at step $t+1$ only depends on the state at time t and not on previous states. If the time parameter is discrete, the Markov chain is a *discrete-time Markov chain*.

The *transition probability* $p_{ij}(t) = P(X_{t+1} = S_j \,|\, X_t = S_{i_t})$ measures the probability of moving from state S_i to S_j at time t. The *transition probability matrix* of a discrete-time Markov chain is given by $P(t) \in \mathbb{R}^{n \times n}$ with $P(t) = (p_{ij}(t))$. If the transition probabilities are fixed, the Markov chain is said to be *homogenous* and $P(t) = P$. The transition probability matrix P is a *(row-)stochastic* matrix. That is, $P \geq 0$ and $\sum_{j=1}^{n} p_{ij} = 1$ for all $1 \leq i \leq n$.

A Markov chain is *irreducible* if the transition probability matrix is irreducible. Equivalently, a Markov chain is irreducible if every state can be reached from every other state. The following theorem presents results on the convergence of irreducible Markov chains [78, p. 693]:

Theorem 1.5.1 Let P be the transition matrix for an irreducible Markov chain on states

$\{S_1, S_2, \ldots, S_n\}$, i.e. $P \in \mathbb{R}^{n\times n}$ is an irreducible stochastic matrix. Let $\boldsymbol{\pi}^T$ denote the left-hand Perron vector of P. The following statements are true for every initial distribution $\boldsymbol{\pi}^{(0)T}$:

(i) The kth step transition matrix is P^k, i.e, $[P^k]_{ij}$ is the probability of moving from S_i to S_j in exactly k steps.

(ii) The kth step transition vector is given by $\boldsymbol{\pi}^{(k)T} = \boldsymbol{\pi}^{(0)T}P^k$.

(iii) If P is primitive, then

$$\lim_{k\to\infty} P^k = \mathbf{1}\boldsymbol{\pi}^T \quad \text{and} \quad \lim_{k\to\infty} \boldsymbol{\pi}^{(k)T} = \boldsymbol{\pi}^T.$$

(iv) If P is imprimitive, then

$$\lim_{k\to\infty} \frac{I + P + \ldots + P^{k-1}}{k} = \mathbf{1}\boldsymbol{\pi}^T \quad \text{and} \quad \lim_{k\to\infty} \frac{\boldsymbol{\pi}^{(0)T} + \boldsymbol{\pi}^{(1)T} + \ldots + \boldsymbol{\pi}^{(k-1)T}}{k} = \boldsymbol{\pi}^T.$$

(v) For both the primitive and imprimitive cases, π_j, the jth component of $\boldsymbol{\pi}$, represents the long-run fraction of time that the chain is in S_j.

(vi) $\boldsymbol{\pi}^T$ is often called the *stationary distribution vector* for the chain because it is the unique distribution vector satisfying $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$

Equation (1.6) is not quite the power method applied to a finite state space Markov because the matrix P used is not necessarily stochastic. If node i has no out-edges, then the ith row of P will be all zeros. This is addressed by replacing zero rows with $\frac{e^T}{n}$, forming the matrix $\bar{P}$. In terms of web surfing, this corresponds to jumping to a random web-page when a page with no out-links is reached. However, $\bar{P}$ is still not guaranteed to be an irreducible matrix and, thus, the results from Theorem 1.5.1 do not guarantee converge to a unique stationary vector. This is addressed in the PageRank algorithm by

forming a matrix $E = \frac{1}{n}\mathbf{1}\mathbf{1}^T$ and setting

$$\bar{\bar{P}} = \alpha\bar{P} + (1-\alpha)E \quad \text{for } 0 < \alpha < 1. \tag{1.7}$$

Often, the values $\alpha = 0.85$ is used. The matrix $\bar{\bar{P}}$, often referred to as the "Google" matrix, is stochastic and irreducible. Therefore, by Theorem 1.5.1, it has a unique stationary distribution vector $\boldsymbol{\pi}$. The PageRank centrality of node i is given by π_i, the ith entry of $\boldsymbol{\pi}$. See Algorithm 2 in Appendix A for a pseudocode implementation of the PageRank algorithm.

It was pointed out in [54] that applying PageRank to the digraph obtained by reversing the direction of the edges provides a natural way to rank the hubs; this is usually referred to as *Reverse PageRank*. In other words, authority rankings are obtained by applying PageRank to the "Google" matrix derived from A, and hub rankings are obtained by the same process applied to A^T.

A more comprehensive overview of the PageRank algorithm can be found in [73]. Many studies have been done to analyze various aspects of the PageRank algorithm [4, 55, 58, 66, 74, 79].

1.6 Comparing centrality measures

There are many way to compare two sets of ranked lists. See [74] for an overview. In this thesis, we use (Pearson) correlation coefficients and the intersection distance method (see [50] as well as [14, 27]) on both the full set of nodes of a network G and on partial lists of nodes to measure similarities between the rankings obtained with two methods. The correlation coefficients are computed using lists of nodes in rank order. The intersection distances are computed using the lists of centrality values.

Given two ranked lists x and y, the intersection distance between the two lists is computed in the following way: let x_k and y_k be the top k ranked items in x and y

respectively. Then the top k intersection distance (or *intersection similarity*) is given by

$$\mathrm{isim}_k(x,y) := \frac{1}{k}\sum_{i=1}^{k}\frac{|x_i \Delta y_i|}{2i} \tag{1.8}$$

where Δ is the symmetric difference operator between the two sets. If the lists are identical, then $\mathrm{isim}_k(x,y) = 0$ for all k. If the two sequences are disjoint, then $\mathrm{isim}_k = 1$. Thus, a small value of $\mathrm{isim}_k(x,y)$ indicates a strong similarity between the two rankings.

Throughout the thesis, we denote by cc the correlation coefficient between the two vector rankings, and by cc_p the correlation coefficient between the top p% of nodes under two ranking systems. We denote by isim the intersection distance between two complete sets of rankings and by $\mathrm{isim}_{p\%}$ the intersection distance between the top $p\%$ of nodes.

1.7 Approximations of centrality scores

In this section, we will provide an overview of two iterative methods used in this thesis to approximate centrality measures based on matrix functions. The matrix exponential and resolvent are computationally expensive and exact computations are not possible for large networks. Several approaches are available for computing the matrix exponential [65]. A commonly used scheme is the one based on Padé approximation combined with the scaling and squaring method [64, 65], implemented in Matlab by the `expm` function. For an $n \times n$ matrix, this method requires $O(n^2)$ storage and $O(n^3)$ arithmetic operations; current implementations are geared toward small, dense matrices. Evaluation of the matrix exponential based on diagonalization also requires$O(n^2)$ storage and $O(n^3)$ operations. Furthermore, these methods cannot be easily adapted to the case where only selected entries (e.g., the diagonal ones) of the matrix exponential are of interest.

1.7.1 Approximation of diagonal entries of matrix functions

To compute the exponential (resolvent) subgraph centralities of the nodes in a network, only the diagonal entries of the matrix exponential (resolvent) are needed. Efficient, accurate methods for estimating (or, in some cases, bounding) arbitrary entries in a matrix function $f(A)$ have been developed by Golub, Meurant and collaborators (see [60] and references therein) and first applied to problems of network analysis by Benzi and Boito in [7]; see also [18]. They have been implemented in the Matlab toolbox `mmq` [77]. Here we provide a brief description of these methods. Additional details can be found in [7] and [60].

Let A be a real, symmetric, $n \times n$ matrix and let f be a function defined on the spectrum of A. Consider the eigendecomposition $A = Q\Lambda Q^T$ and $f(A) = Qf(\Lambda)Q^T$, where $Q = [\mathbf{q}_1, \ldots, \mathbf{q}_n]$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. For notational simplicity, here we assume that the eigenvalues of A are ordered as $\lambda_1 \leq \ldots \leq \lambda_n$. For given vectors $\mathbf{u}$ and $\mathbf{v}$ we have

$$\mathbf{u}^T f(A)\mathbf{v} = \mathbf{u}^T Q f(\Lambda) Q^T \mathbf{v} = \mathbf{w}^T f(\Lambda)\mathbf{z} = \sum_{k=1}^{n} f(\lambda_k) w_k z_k, \tag{1.9}$$

where $\mathbf{w} = Q^T\mathbf{u} = (w_k)$ and $\mathbf{z} = Q^T\mathbf{v} = (z_k)$. In particular, for $f(A) = e^A$ we obtain

$$\mathbf{u}^T e^A \mathbf{v} = \sum_{k=1}^{n} e^{\lambda_k} w_k z_k. \tag{1.10}$$

Choosing $\mathbf{u} = \mathbf{v} = \mathbf{e}_i$ (the vector with the ith entry equal to 1 and all the remaining ones equal to 0) we obtain an expression for the subgraph centrality of node i:

$$SC(i) := \sum_{k=1}^{n} e^{\lambda_k} \mathbf{q}_k(i)^2 ,$$

where $\mathbf{q}_k(i)$ denotes the ith component of vector $\mathbf{q}_k$. Likewise, choosing $\mathbf{u} = \mathbf{e}_i$ and $\mathbf{v} = \mathbf{e}_j$ we obtain the following expression for the communicability between node i and

node j:

$$C(i,j) := \sum_{k=1}^{n} e^{\lambda_k} \mathbf{q}_k(i)\mathbf{q}_k(j).$$

Analogous expressions hold for other matrix functions, such as the resolvent. Hence, the problem is reduced to evaluating bilinear expressions of the form $\mathbf{u}^T f(A)\mathbf{v}$. Such bilinear forms can be thought of as Riemann- Stieltjes integrals with respect to a (signed) spectral measure:

$$\mathbf{u}^T f(A)\mathbf{v} = \int_a^b f(\lambda)d\mu(\lambda), \quad \mu(\lambda) = \begin{cases} 0, & \text{if } \lambda < a = \lambda_1, \\ \sum_{k=1}^{i} w_k z_k, & \text{if } \lambda_i \leq \lambda < \lambda_{i+1}, \\ \sum_{k=1}^{n} w_k z_k, & \text{if } b = \lambda_n \leq \lambda. \end{cases}$$

This integral can be approximated by means of a Gauss-type quadrature rule:

$$\int_a^b f(\lambda)d\mu(\lambda) = \sum_{j=1}^{p} c_j f(t_j) + \sum_{k=1}^{q} v_k f(\tau_k) + R[f], \tag{1.11}$$

where $R[f]$ denotes the error. Here the nodes $\{t_j\}_{j=1}^p$ and the weights $\{c_j\}_{j=1}^p$ are unknown, whereas the nodes $\{\tau_k\}_{k=1}^q$ are prescribed. We have

- $q = 0$ for the Gauss rule,
- $q = 1$, $\tau_1 = a$ or $\tau_1 = b$ for the Gauss–Radau rule,
- $q = 2$, $\tau_1 = a$ and $\tau_2 = b$ for the Gauss–Lobatto rule.

For certain matrix functions, including the exponential and the resolvent, these quadrature rules can be used to obtain lower and upper bounds on the quantities of interest; prescribing additional quadrature nodes leads to tighter and tighter bounds, which (in exact arithmetic) converge monotonically to the true values [60]. The evaluation of these quadrature rules is mathematically equivalent to the computation of orthogonal polynomials via a three-term recurrence, or, equivalently, to the computa-

tion of entries and spectral information of a certain tridiagonal matrix via the Lanczos algorithm (see Algorithm 3 in Appendix A). More information on the Lanczos algorithm can be found in [76]. Here we briefly recall how this can be done for the case of the Gauss quadrature rule, when we wish to estimate the ith diagonal entry of $f(A)$. It follows from (1.11) that the quantity of interest has the form $\sum_{j=1}^{p} c_j f(t_j)$. This can be computed from the relation [60, p. 28]:

$$\sum_{j=1}^{p} c_j f(t_j) = \mathbf{e}_1^T f(J_p)\mathbf{e}_1,$$

where

$$J_p = \begin{pmatrix} \omega_1 & \gamma_1 & & & \\ \gamma_1 & \omega_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{p-2} & \omega_{p-1} & \gamma_{p-1} \\ & & & \gamma_{p-1} & \omega_p \end{pmatrix}$$

is a tridiagonal matrix whose eigenvalues are the Gauss nodes, whereas the Gauss weights are given by the squares of the first entries of the normalized eigenvectors of J_p. The entries of J_p are computed using the Lanczos algorithm with starting vectors $\mathbf{x}_{-1} = \mathbf{0}$ and $\mathbf{x}_0 = \mathbf{e}_i$. Note that it is not required to compute all the components of the eigenvectors of J_p if one uses the Golub–Welsch QR algorithm; see [60].

For small p (i.e., for a small number of Lanczos steps), computing the $(1,1)$ entry of $f(J_p)$ is inexpensive. The main cost in estimating one entry of $f(A)$ with this approach is associated with the sparse matrix-vector multiplies in the Lanczos algorithm applied to the adjacency matrix A. If only a small, fixed number of iterations are performed for each diagonal element of $f(A)$, as is usually the case, the computational cost (per node) is at most $O(n)$ for a sparse graph, resulting in a total cost of $O(n^2)$ for computing the subgraph centrality of every node in the network. If only $k < n$ subgraph centralities

are wanted, with k independent of n, then the overall cost of the computation will be $O(n)$ provided that sparsity is carefully exploited in the Lanczos algorithm and that only a small number p of iterations (independent of n) is carried out. Note, however, that depending on the connectivity characteristics of the network under consideration, the prefactor in the $O(n)$ estimate could be large. The algorithm can be implemented so that the storage requirements are $O(n)$ for a sparse network—that is, a network in which the total number of links grows linearly in the number n of nodes.

Additionally, in most applications one is not so much interested in computing an exact ranking of *all* the nodes in a network, but only in identifying the top k ranked nodes, where the integer k is small compared to n (for example, $k = 10$ or $k = 20$). Research has been done to develop methods that are capable of quickly identifying the top k nodes without having to compute accurate subgraph centrality scores for each node [8, 53]. We return to this topic in Section 4.8.

1.7.2 Approximation of row sums of matrix functions

To compute Katz centrality or total communicability (which is introduced in Chapter 2), all that is needed is the row sums of the matrix resolvent or exponential: $(I-\alpha A)^{-1}\mathbf{1}$ or $e^{\beta A}\mathbf{1}$. For digraphs, A^T may be used in place of A. In recent years, efficient algorithms have been developed for computing the *action* of a matrix function on a vector, that is, for computing the vector $f(A)\mathbf{b}$ for a given matrix A (usually large and sparse), vector $\mathbf{b}$, and function f. A particularly important case is that of the matrix exponential, since this provides a solution method for initial value problems for first-order systems of linear ordinary differential equations. These algorithms, based on variants of the Lanczos, Arnoldi or other Krylov subspace method, access the matrix A only in the form of (sparse) matrix-vector products and have $O(n)$ storage cost for a sparse $n \times n$ matrix A [65, Chapter 13]. When $\mathbf{b} = \mathbf{1}$, the vector with all its entries equal to 1, the

ith entry of the resulting vector $f(A)\mathbf{1}$ contains the ith row sum of $f(A)$:

$$[f(A)\mathbf{1}]_i = \sum_{j=1}^{n} [f(A)]_{ij}\,, \quad 1 \le i \le n\,.$$

This quantity can be computed much faster than the diagonal entries of a matrix function using current computational techniques. Indeed, computing individual entries of matrix functions $f(A)$ is generally costly for large A even with the best available algorithms [7, 46]. Of course, the same is true if the vector $\mathbf{1}$ is replaced by some other vector—typically, an "external importance vector" which can be used to take into account intrinsic, not network-related contributions to the centrality of each node [84, pp. 174–175].

An efficient algorithm for evaluating $f(A)\mathbf{v}$ using a restarted Krylov method has been presented in [1, 34]. In this approach, the basic operation is represented by matrix-vector products with A. This method has been implemented in the Matlab toolbox `funm_kryl` by Stefan Güttel [63].

Given $A \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, and f analytic on a neighborhood containing $\sigma(A)$, the evaluation of

$$f(A)\mathbf{b} \tag{1.12}$$

using Krylov subspace methods is based on an Arnoldi-like decomposition of A into

$$AV_m = V_{m+1}\tilde{H}_m = V_m H_m + h_{(m+1,m)}\mathbf{v}_{m+1}\mathbf{e}_m^T \tag{1.13}$$

where $\tilde{H} = h_{ij}$ is an $(m+1)\times m$ upper Hessenberg matrix, $H_m = [I_m\ \mathbf{0}]\tilde{H}_m$, and $\mathbf{e}_m \in \mathbb{R}^m$ is the mth unit vector. The columns of V_m form a basis of the Krylov subspace:

$$\mathcal{K}_m(A,\mathbf{b}) = \operatorname{span}\{\mathbf{b}, A\mathbf{b}, \ldots, A^{m-1}\mathbf{b}\}. \tag{1.14}$$

If A is Hermitian, then H_m in (1.13) is Hermitian tridiagonal and the basis $\{\mathbf{v}_m\}$ of $\mathcal{K}_m(A, \mathbf{b})$ is orthogonal. In this case, the Arnoldi approximation of (1.13) is given by:

$$f(A)\mathbf{b} \approx V_m f(H_m) V_m^* \mathbf{b} = V_m f(H_m)\mathbf{e}_1. \tag{1.15}$$

The evaluation of (1.15) requires that all m basis vectors of $\mathcal{K}_m(A, \mathbf{b})$ be available, which can be computationally infeasible for large A and/or m. A restarted Krylov subspace algorithm repeatedly generates a basis of fixed dimension $\hat{m}$, producing a sequence of approximations to $f(A)\mathbf{b}$. At the end of each iteration, all basis vectors but the last, $\mathbf{v}_{\hat{m}}$, are discarded and $\mathbf{v}_{\hat{m}}$ becomes the initial vector for the next Krylov subspace.

For many network of practical interest, the cost is typically $O(n)$, although the prefactor can vary considerably for different types of networks. A more comprehensive overviews of Krylov subspace methods for approximating $f(A)\mathbf{b}$ can be found in [65, Ch 13]. Additional uses of Krylov subspace methods can be found in [78, Ch. 7] and [61, Ch. 9 and 10].

2 Total Communicability as a Centrality Measure

2.1 Introduction

A standard method of measuring node importance is subgraph centrality [48], which is based on the diagonal entries of a matrix function applied to the adjacency matrix A of the network in question. Here, the matrix exponential e^A is frequently used. While this approach has been successfully used in a number of problems [40, 46, 47], obtaining estimates of the diagonals of e^A for a large network with adjacency matrix A can be quite expensive.

In this chapter, we consider the implications of instead using the row sums of e^A or similar matrix functions as a measure of node centrality, focusing for the sake of brevity on undirected networks. The interpretation of this measure in terms of total communicability of a node is given, and compared to the one for subgraph centrality in section 2.2. In section 2.3, the concept of *total network communicability* is introduced and discussed. Section 2.4 contains experimental comparisons of subgraph centrality and total communicability using various synthetic and real-world networks. Sections 4.8 and 2.6 discuss computational aspects and the use of row sum centrality with other standard matrix functions, respectively. We offer some conclusive remarks in section 2.7.

Centrality measures based on the row sums of matrix functions have long been in

use in network analysis. Note that for the case of the "identity function" $f(A) = A$, and symmetric A (undirected networks), we recover degree centrality. The off-diagonal row sums of e^A have been used in social network analysis to measure the resilience of an individual in the face of hostile attacks from within the network [40, Chapter 6]. More recently, row and column sums of e^A have been applied to the identification of hubs and authorities in directed networks [8]. For resolvent-type functions, such as $f(A) = (I - \alpha A)^{-1}$ (with I the $n \times n$ identity matrix), for suitable values of $\alpha > 0$, we recover the well-known Katz centrality and its variants, also known as α-centrality; see, e.g., [67] and [16, 17, 19, 62]. None of these previous studies, however, considered algorithmic aspects such as computational cost, storage, and so forth.

2.2 Diagonal entries vs. row sums

In [48], the authors introduce the concept of subgraph centrality as a centrality measurement for nodes in a network. This provides a ranking based on the diagonal entries of a matrix function applied to the adjacency matrix. Although there are various choices of function to use, the most common is the matrix exponential. The subgraph centrality of node i is given by $[e^A]_{ii}$ where A is the adjacency matrix of the network. This counts the number of closed walks centered at node i weighting a walk of length k by a penalty factor of $\frac{1}{k!}$. In this way, shorter walks are deemed more important than longer walks.

By contrast, the row sum of e^A for node i is given by $\sum_{j=1}^{n}[e^A]_{ij}$, which counts all walks between node i and all the nodes in the network (node i included), weighting walks of length k by a penalty factor of $\frac{1}{k!}$. Thus, the ith row sum of e^A can be interpreted as the *total subgraph communicability* of node i, and can be interpreted as a measure of the importance of the ith node in the network, since a node with high communicability with a large number of other nodes in the network is likely to be an

important node, and certainly a more important node than one characterized by low total communicability.

An immediate question is how this centrality measure compares with the subgraph centrality of node i in the network. In general, the rankings produced by the total communicability measure will not be the same as those produced by the subgraph centrality measure. The difference between the two rankings is

$$\sum_{j=1}^{n}[e^A]_{ij} - [e^A]_{ii} = \sum_{j\neq i}[e^A]_{ij} = \sum_{j\neq i}\sum_{k=1}^{n} e^{\lambda_k} v_{ki} v_{kj}\,, \tag{2.1}$$

where v_{ik} is the ith element of the normalized eigenvector $\mathbf{v}_k$ of A associated with the eigenvalue λ_k. Note that e^A is always positive definite and that its diagonal entries are often large compared to the off-diagonals. If the diagonal entries of e^A vary over a wide range while its off-diagonal sums remain confined within a more narrow range, the rankings produced by the two methods will not differ by much. However, this depends both on the spectrum of A and the entries of the eigenvectors.

While it appears to be difficult, in general, to establish a relation between the rankings produced by the subgraph centrality and total communicability, for certain types of simple graphs it is easy to show that the two methods will give identical rankings. These include *complete graphs* and *cycles* (where each node has the exact same ranking under both systems), *paths* and *star graphs*. A star graph on n nodes has one central node that is connected to each of the $n-1$ remaining nodes and no other edges. Under both ranking systems, the central node is ranked highest and the remaining nodes all have the same scores. This can be shown either using graph theory or by examining the eigenvalues and eigenvectors of the star graph (more information about the spectra of star graphs can be found in [2]).

One case where the two measures could be expected to give similar rankings is that of networks with a large spectral gap, $\lambda_1 - \lambda_2$, between the first (largest) and second

eigenvalue. We have:

$$\left[e^A\right]_{ii} = e^{\lambda_1} v_{1i}^2 + \sum_{k=2}^{n} e^{\lambda_k} v_{ki}^2$$

and

$$\left[e^A \mathbf{1}\right]_i = e^{\lambda_1} \left(\mathbf{v}_1^T \mathbf{1}\right) v_{1i} + \sum_{k=2}^{n} e^{\lambda_k} \left(\mathbf{v}_i^T \mathbf{1}\right) v_{ki}.$$

Dividing both expressions by the constant e^{λ_1} (which does not affect the rankings) and observing[1] that $\mathbf{v}_1^T \mathbf{1} = \|\mathbf{v}_1\|_1$ shows that for $\lambda_1 \gg \lambda_2$ the two rankings are largely determined by the quantities v_{1i}^2 and $\|\mathbf{v}_1\|_1 v_{1i}$, respectively, and therefore by the entries v_{1i} of the dominant eigenvector of A. Thus, if the difference $\lambda_1 - \lambda_2$ is sufficiently large, the two centrality measures reduce to eigenvector centrality [16] and therefore can be expected to result in very similar rankings, especially for the top nodes. Numerical experiments (not shown here) performed on Erdös–Renyi graphs with large spectral gaps have confirmed this fact.

However, it is difficult to quantify *a priori* how large the spectral gap needs to be for all these rankings to be identical (or even approximately the same). In the section on computational experiments we will see that there can be significant differences between the rankings obtained using subgraph centrality and those using total communicability centrality, even for networks with a relatively large spectral gap.

2.3 Total network communicability

The total communicabilities of individual nodes give a measure of how well each node communicates with the other nodes of the network. In order to measure how effectively communication takes place across the network as a whole, we consider the sum of all

[1] By the Perron–Frobenius Theorem, the dominant eigenvector can be chosen to have nonnegative entries, and positive entries when the graph G is connected.

the total communicabilities. For a network with adjacency matrix A, this is given by

$$C(A) = \sum_{i=1}^{n} \left[e^A \mathbf{1}\right]_i = \sum_{i=1}^{n} \sum_{k=1}^{n} e^{\lambda_k} \left(\mathbf{v}_i^T \mathbf{1}\right) v_{ki} = \mathbf{1}^T e^A \mathbf{1}\,, \tag{2.2}$$

where, as in section 2.2, λ_k is the kth eigenvalue of A and v_{ik} is the ith element of the normalized eigenvector $\mathbf{v}_k$ associated with λ_k. Here we propose to use the *total network communicability*, $C(A)$, as a global measure of the ease of sending information across a network. We emphasize that while $C(A)$ is defined as the sum of all the entries of e^A, it is not necessary to know any of the individual entries of e^A to compute $C(A)$; indeed, very efficient methods exist to compute quadratic forms of the type $\mathbf{v}^T f(A)\mathbf{v}$ for a given function $f(x)$, matrix A and vector $\mathbf{v}$, see [7, 9, 60] and the discussions in Sections 1.7.2 and 4.8.

It is instructive to compare the total communicability of a network with the *Estrada index*, an important graph invariant defined as the sum of all the subgraph centralities:

$$EE(A) = \sum_{i=1}^{n} \left[e^A\right]_{ii} = \sum_{i=1}^{n} e^{\lambda_i} = \mathrm{Tr}(e^A).$$

The following proposition provides simple lower and upper bounds for $C(A)$ in terms of $EE(A)$ and other spectral quantities associated with the underlying network.

Proposition 2.3.1 Let A be the adjacency matrix of a simple network on n vertices. Then,

$$EE(A) \leq C(A) \leq n\, e^{\|A\|_2}\,,$$

where $\|A\|_2$ denotes the spectral norm of A. In particular, for an undirected network we have

$$EE(A) \leq C(A) \leq n\, e^{\lambda_1}.$$

Proof: The lower bound is trivial, as

$$EE(A) = \sum_{i=1}^{n} \left[e^A\right]_{ii} \leq \sum_{i=1}^{n}\sum_{j=1}^{n} \left[e^A\right]_{ij} = \sum_{i=1}^{n} \left[e^A \mathbf{1}\right]_i = C(A).$$

The upper bound follows from noticing that $C(A) = \mathbf{1}^T e^A \mathbf{1} = \left(e^A \mathbf{1}\right)^T \mathbf{1} = \langle e^A \mathbf{1}, \mathbf{1}\rangle$ and applying the Cauchy–Schwarz inequality to the quadratic form $\langle e^A \mathbf{1}, \mathbf{1}\rangle$:

$$|\langle e^A \mathbf{1}, \mathbf{1}\rangle| \leq \|e^A \mathbf{1}\|_2 \|\mathbf{1}\|_2 \leq \|e^A\|_2 \|\mathbf{1}\|_2 \|\mathbf{1}\|_2 \leq n\, e^{\|A\|_2}.$$

For an undirected network A is symmetric and $\lambda_1 = \|A\|_2$. □

Note that the lower bound is attained in the case of the "empty" graph with adjacency matrix $A = 0$, while the upper bound is attained on the complete graph, whose adjacency matrix is $A = \mathbf{1}\mathbf{1}^T - I$.

The bounds from Proposition 2.3.1 also hold for $e^{\beta A}$, $\beta > 0$. For any connected graph with adjacency matrix A, the bounds get tighter as $\beta \to 0+$, since both the lower and upper bound tend to 1. The parameter β can be interpreted as an *inverse temperature* and is a reflection of external disturbances on the network (see, e.g., [46] for details); taking $\beta \to 0+$ is equivalent to "raising the temperature" of the environment surrounding the network.

When appropriately normalized, $C(A)$ can be used to compare the ease of information exchange on different networks. This could be useful, for instance, in the design of communication networks. In the following sections we compute the total communicability for various types of networks. The question arises of what would constitute a reasonable normalization factor. There are several possibilities. Normalizing $C(A)$ by the number n of nodes corresponds to the average total communicability of the network per node. Similarly, normalizing $C(A)$ by the number m of edges would correspond to the average total communicability of the network per edge. We note also that the

minimum value of $C(A)$ is n, corresponding to the empty graph on n nodes ($E = \emptyset$), while the maximum value is $n^2 e^{n-1} - n$, corresponding to the complete graph on n nodes. The expression

$$\hat{C}(A) := \frac{C(A) - n}{ne^{n-1} - n}$$

takes its values in the interval $[0, 1]$, with $\hat{C}(A) = 0$ for "empty" graphs (no communication can take place on such graphs) and $\hat{C}(A) = 1$ on complete graphs (for which the ease of communication between nodes is clearly maximum). Unfortunately, the denominator in this expression grows so fast that for most sparse graphs evaluating $\hat{C}(A)$ results in underflow.

In the experiments below we chose to normalize $C(A)$ by n, the number of nodes, and by m, the number of edges; for the network used in our tests we found that comparing networks based on $C(A)/n$ or on $C(A)/m$ yields exactly the same rankings, therefore we only include results for the former measure.

2.4 Computational studies

In this section we carry out extensive centrality computations for a variety of networks, with the aim of comparing subgraph centrality with total communicability centrality. In particular, we are interested in determining if, or for what type of networks, the two centrality measures provide similar rankings. Moreover, for those networks where the two measures result in rankings that differ significantly, we would like to obtain some insights on why this is the case. Of course it would be desirable to know when one measure should be preferred to the other, but this is a difficult problem since it is not easy to come up with objective criteria for comparing ranking methods (see the discussion in [74, Chapter 16]). We will compare the two methods in terms of computational cost in section 4.8. To measure similarities between the rankings obtained with the two methods, we use the methodology described in Chapter 1, Section 1.6.

Unless otherwise specified, all experiments were performed using Matlab version 7.9.0 (R2009b) on a MacBook Pro running OS X Version 10.6.8, a 2.4 GHZ Intel Core i5 processor and 4 GB of RAM. In this section, we use the Matlab built-in function `expm` for computing the matrix exponential.

2.4.1 Test matrices

The synthetic examples used in the tests were produced using the CONTEST toolbox in Matlab [91, 92]. The graphs tested were of two types: preferential attachment (Barabási–Albert) model and small world (Watts–Strogatz) model. In CONTEST, these graphs and the corresponding adjacency matrices can be built using the functions `pref` and `smallw`, respectively.

The preferential attachment model was designed to produce networks with scale-free degree distributions as well as the small world property [5]. In CONTEST, preferential attachment networks are constructed using the command `pref(n,d)` where n is the number of nodes in the network and $d \geq 1$ is the number of edges each new node is given when it is first introduced to the network. The network is created by adding nodes one by one (each new node with d edges). The edges of the new node connect to nodes already in the network with a probability proportional to the degree of the already existing nodes. This results in a scale-free degree distribution. Note that with this construction, the minimum degree of the network is d. When $d > 1$ this means that the network has no dangling nodes (nodes of degree 1), whereas in many real-life networks one often observes a high number of dangling nodes. In the CONTEST toolbox, the default value is $d = 2$.

In our experiments, we tested various values of d on a network of size $n = 1000$: twenty networks were tested for all values $1 \leq d \leq 10$, as well as for a few larger values. In Table 2.1, the averages of the correlation coefficients between the subgraph centrality rankings and the total subgraph communicability rankings can be found for

Table 2.1: Comparison, using the correlation coefficient, of rankings based on the diagonal entries and row sums of e^A for 1000-node scale-free networks of various parameters built using the `pref` function in the CONTEST Matlab toolbox. The values reported are the averages over 20 matrices with the same parameters. The parameter d is the initial degree of nodes in the network (and consequently the minimum degree of the network).

d	cc
1	0.224
2	0.343
3	0.517
4	0.905
5	0.993
6	0.999
7	0.999
≥ 8	1

Table 2.2: Intersection distance comparisons of rankings based on the diagonal entries and row sums of e^A for 1000-node scale-free networks of various parameters built using the `pref` function in the CONTEST Matlab toolbox. The values reported are the averages over 20 matrices with the same parameters. The parameter d is the initial degree of nodes in the network (and consequently the minimum degree of the network).

d	isim	$\text{isim}_{10\%}$
1	0.174	0.199
2	0.036	0.031
3	0.003	0.005
4	2.04e-4	2.79e-4
5	1.30e-5	1.71e-5
6	9.83e-7	0
7	4.93e-7	0
≥ 8	0	0

various values of d. The intersection distance values can be found in Table 2.2. The intersection distance values were calculated both for the full set of rankings and for the top 10% of ranked nodes.

The results show that correlation between the two metrics increases and the intersection distance value decreases quickly with the value of the parameter d. The intersection distance values for the top 10% of nodes are very close to those for the complete set of nodes. For sufficiently dense networks, the two measures provide essen-

tially identical rankings, producing correlation coefficients close to 1 and intersection distances close to 0.

A second class of synthetic test matrices used in our experiments corresponds to small-world networks (Watts–Strogatz model). The small world model was developed as a way to impose a high clustering coefficient onto classical random graphs [95]. The name comes from the fact that, like classical random graphs, the Watts–Strogatz model produces networks with the small world (that is, small graph diameter) property. To build these matrices, the input is `smallw(n,d,p)` where n is the number of nodes in the network, which are arranged in a ring and connected to their d nearest neighbors on the ring. Then each node is considered independently and, with probability p, a link is added between the node and one of the other nodes in the network, chosen uniformly at random. At the end of this process, all loops and repeated edges are removed. For this set of experiments, the size of the network was fixed at $n = 1000$ and the probability of an extra link was left at the default value of $p = 0.1$ while d was varied.

The values of d tested were: all values $1 \leq d \leq 10$, along with all multiples of 10 up to 200. In each case, twenty networks were created with each value of d. The average correlation coefficients between the subgraph centrality rankings and the total communicability rankings are given in Table 2.3. As before, the correlation coefficients were computed between the complete sets of rankings. The intersection distances, reported in Table 2.4, were computed on both the complete sets of rankings and the top 10% of ranked nodes.

It is evident from these results that for this class of small world networks, the similarity between the two ranking measures is much weaker than for the preferential attachment model, at least as long as the networks remain fairly sparse. The intersection distances are also relatively large, further indicating that the two measures are much more weakly related than in the case of the preferential attachment model. For

Table 2.3: Comparison, using the correlation coefficient, of rankings based on the diagonal entries and row sums of e^A for 1000-node small world networks of various parameters made using the `smallw` function in the CONTEST Matlab toolbox. The values reported are the average over 20 matrices with the same parameters.

d	cc
1	0.177
2	0.089
3	0.037
4	0.033
5	0.031
6	0.048
7	0.039
8	0.046
9	0.031
10	0.054

d	cc
20	0.156
30	0.222
40	0.240
50	0.310
60	0.426
70	0.431
80	0.747
90	0.926
100	0.997
≥ 110	1

Table 2.4: Intersection distance comparison of rankings based on the diagonal entries and row sums of e^A for 1000-node small world networks of various parameters made using the `smallw` function in the CONTEST Matlab toolbox. The values reported are the averages over 20 matrices with the same parameters.

d	isim	$\text{isim}_{10\%}$
1	0.015	0.071
2	0.056	0.160
3	0.089	0.252
4	0.117	0.350
5	0.151	0.479
6	0.178	0.621
7	0.218	0.709
8	0.243	0.731
9	0.262	0.705
10	0.284	0.725

d	isim	$\text{isim}_{10\%}$
20	0.311	0.713
30	0.239	0.535
40	0.133	0.351
50	0.111	0.214
60	0.039	0.120
70	0.014	0.041
80	0.002	0.007
90	1.71e-4	4.05e-4
100	5.88e-6	1.09e-5
≥ 110	0	0

some values of d, the intersection distance between the top 10% of nodes is above 0.7, indicating that there is little consistency among the rankings of the top 10% of nodes under the two measures. As the networks become increasingly dense, however, the correlation between the two measures becomes stronger and the intersection distance eventually decreases.

Table 2.5: Comparison of the total network communicability $C(A)$ of a ring lattice and small world rings with increasing probability of a shortcut. The computed values were averaged over 20 instances.

Graph	number of edges	$C(A)$	normalized $C(A)$
5000 node ring lattice	5000	3.69e04	7.4
`smallw(5000,1,.1)`	5492	4.83e04	9.7
`smallw(5000,1,.2)`	6222	6.22e04	12.4
`smallw(5000,1,.3)`	6495	7.92e04	15.8
`smallw(5000,1,.4)`	6990	9.90e04	19.8
`smallw(5000,1,.5)`	7496	1.24e05	24.8
`smallw(5000,1,.6)`	7999	1.53e05	30.6

2.4.2 Total communicability in small world networks

For networks with low connectivity (or high locality), the total network communicability can be expected to be low compared with networks with higher connectivity. For instance, on a 5000 node ring lattice, the total network communicability is $C(A) =$ 3.69e04 and the normalized $C(A)$ is 7.4. However, when even a few shortcuts are added across the lattice using the Watts–Strogatz small world model, this value jumps considerably. If the probability of a shortcut is $p = 0.1$, the normalized total network communicability (averaged over 20 networks created using input `smallw(5000,1,p)`) is 9.7. If the probability of a shortcut is increased to $p = 0.2$, the normalized total network communicability increases to 12.4. These and additional results can be found in Table 2.5 and Fig. 2.1.

2.4.3 Discussion of test results using synthetic data

The results reported so far can be explained as follows. In a (regular) ring-shaped network, no node is more central than the other nodes and no reasonable centrality measure would be able to assign a (strict) ranking of the nodes. In a small world network obtained by perturbing a regular ring-shaped network, all the nodes have *approximately* the same importance, with the nodes with extra links ("shortcuts") being

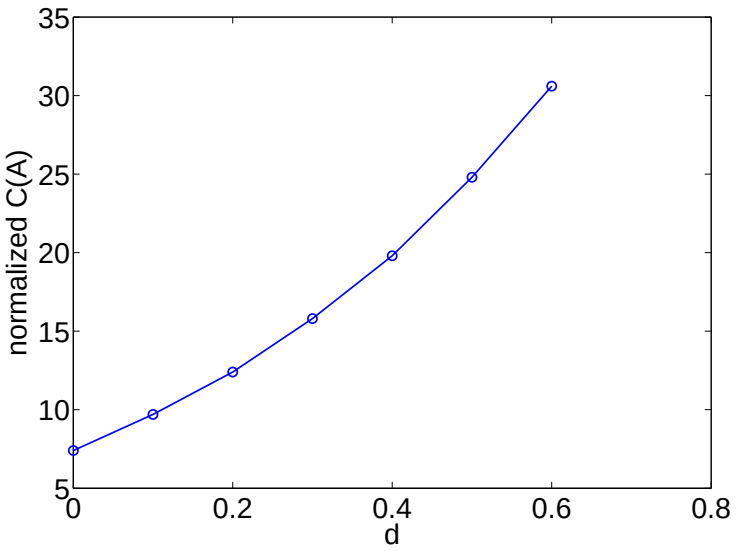

Figure 2.1: Plot of of the total network communicability $C(A)$ for small world graphs with increasing probability d of a shortcut. The computed values were averaged over 20 instances.

slightly more important than the others. When d is small, these shortcuts matter more, but the subgraph centrality scores and the total communicability scores do not have a large range. Due to this, the change in the scores due to moving from the subgraph centrality measure to the total communicability measure can have a high impact on node rankings. This leads to a low correlation and a relatively large intersection distance between the two rankings. When d gets very large, the shortcuts matter less and cause less perturbations between the two sets of rankings. By contrast, in a scale-free preferential attachment network both the subgraph centrality scores and total communicability scores are spread out over a large range, even for small d, and adding the corresponding off-diagonal row sums to the diagonal entries does not change the rankings as much.

2.4.4 Real data

Next, we study correlations between the two ranking methods using various networks corresponding to real data. The networks in this section come from a variety of sources. The Zachary Karate Club network is a classic example in network analysis [98]. The Intravenous Drug User and Yeast PPI networks were provided to us by Prof. Ernesto Estrada. The Yeast PPI network has 440 ones on the diagonal due to the self-interactions of certain proteins. The remainder of the networks can be found in the University of Florida Sparse Matrix Collection [31] under different "groups". The Erdös networks are from the Pajek group. They represent various subnetworks of the Erdös collaboration network. The ca-GrQc and ca-HepTh from the SNAP group are collaboration networks for the arXiv General Relativity and High Energy Physics Theory subsections, respectively. The as-735 network, also from the SNAP group, contains the communication network of a group of Autonomous Systems (AS) measured over 735 days between November 8, 1997 and January 2, 2000. Communication occurs when routers from two Autonomous Systems exchange information. The Minnesota network from the Gleich group represents the Minnesota road network. The order n and number of nonzeros nnz of the corresponding adjacency matrices are given in Table 2.7. These networks exhibit a wide variety of structural properties and together constitute a rather heterogeneous sample of real-world networks. All networks except the Yeast PPI network are simple and all are undirected. The Yeast PPI network is undirected but does have several ones on the diagonal, representing the self-interaction of certain proteins. Table 2.6 reports the correlation coefficients between the two sets of rankings for all the nodes, the top 10% of the nodes and the top 1% of the nodes (limited to the cases where the two methods rank the same nodes in the top 10% and top 1%), as well as the value of the two largest eigenvalues λ_1 and λ_2 of the adjacency matrix.

Table 2.7 reports the correlation coefficients between the two sets of rankings for all the nodes, the top 10% of the nodes and the top 1% of the nodes (limited to the

Table 2.6: Basic data on the real-world networks examined.

Graph	n	nnz	λ_1	λ_2
Zachary Karate Club	34	156	6.726	4.977
Drug User	616	4024	18.010	14.234
Yeast PPI	2224	13218	19.486	16.134
Pajek/Erdos971	472	2628	16.710	10.199
Pajek/Erdos972	5488	14170	14.448	11.886
Pajek/Erdos982	5822	14750	14.819	12.005
Pajek/Erdos992	6100	15030	15.131	12.092
SNAP/ca-GrQc	5242	28980	45.617	38.122
SNAP/ca-HepTh	9877	51971	31.035	23.004
SNAP/as-735	7716	26467	46.893	27.823
Gleich/Minnesota	2642	6606	3.2324	3.2319

Table 2.7: Comparison of rankings based on the diagonal and row sum of e^A for various real-world networks.

Graph	cc	cc_{10}	cc_1
Zachary Karate Club	0.420	–	1
Drug User	0.083	0.976	1
Yeast PPI	0.108	–	1
Pajek/Erdos971	0.523	1	1
Pajek/Erdos972	0.122	–	–
Pajek/Erdos982	0.128	–	–
Pajek/Erdos992	0.143	–	–
SNAP/ca-GrQc	0.021	–	0.995
SNAP/ca-HepTh	0.007	–	–
SNAP/as-735	0.904	0.771	1
Gleich/Minnesota	0.087	–	–

Table 2.8: Intersection distance comparison of rankings based on the diagonal and row sum of e^A for various real-world networks.

Graph	isim	$\text{isim}_{10\%}$	$\text{isim}_{1\%}$
Zachary Karate Club	0.044	0.111	0
Drug User	0.102	0.002	0
Yeast PPI	0.025	0.056	0
Pajek/Erdos971	0.004	0	0
Pajek/Erdos972	0.081	0.075	0.047
Pajek/Erdos982	0.079	0.065	0.044
Pajek/Erdos992	0.077	0.055	0.034
SNAP/ca-GrQc	0.043	0.091	5.49e-4
SNAP/ca-HepTh	0.142	0.319	0.134
SNAP/as-735	1.81e-4	0.001	0
Gleich/Minnesota	0.096	0.341	0.709

cases where the two methods rank the same nodes in the top 10% and top 1%), as well as the value of the two largest eigenvalues λ_1 and λ_2 of the adjacency matrix. A "–" in the table signifies that different lists of top nodes where produced under the two rankings, hence correlation coefficients could not be computed in such cases. Table 2.8 reports the intersection distances between the two sets of rankings for all, for the top 10%, and for the top 1% of the nodes. Table 2.9 reports the normalized Estrada index and normalized total network connectivity for each of the networks. For the Zachary Karate Club, which only has 34 nodes, $cc_1 = 1$ and $\text{isim}_{1\%} = 0$ indicate that the top two ranked nodes under the two rankings are the same. The top node is node 34, which corresponds to the president of the karate club, and the second is node 1, which corresponds to the instructor. These were the two most influential members of the club and fought with each other to the point that eventually the club split into two factions aligned around each of them [98].

The results indicate that there is a good deal of variation between the correlation coefficients for these networks. The correlation coefficient between the rankings of all the nodes ranges from a low of 0.007 for the SNAP/ca-HepTh network to a high of 0.904 for the SNAP/as-735 network. Even for networks that come from similar datasets, the

Table 2.9: Comparison of the normalized Estrada index $EE(A)/n$, the normalized total network connectivity $C(A)/n$, and $e^{\|A\|_2}$ $(= e^{\lambda_1})$ for various real-world networks.

Graph	normalized $EE(A)$	normalized $C(A)$	$e^{\|A\|_2}$
Zachary Karate Club	30.62	608.79	833.81
Drug User	1.12e05	1.15e07	6.63e07
Yeast PPI	1.37e05	3.97e07	2.90e08
Pajek/Erdos971	3.84e04	4.20e06	1.81e07
Pajek/Erdos972	408.23	1.53e05	1.88e06
Pajek/Erdos982	538.58	2.07e05	2.73e06
Pajek/Erdos992	678.87	2.50e05	3.73e06
SNAP/ca-GrQc	1.24e16	8.80e17	6.47e19
SNAP/ca-HepTh	3.05e09	1.06e11	3.01e13
SNAP/as-735	3.00e16	3.64e19	2.32e20
Gleich/Minnesota	2.86	14.13	35.34

correlation coefficients can be very different. For example, the networks in the Pajek group are all subsets of the Erdös collaboration network, but correlations between the two sets of rankings range between 0.122 for the Erdos972 network and 0.583 for the Erdos971 network.

For most of the networks, the correlation coefficient (when defined) increases when only the top 1% of nodes are considered (cc_1), sometimes greatly. Five of the networks (Zachary Karate Club, Drug User, Yeast PPI, Pajek/Erdos971, and SNAP/as-735) produce the exact same rankings on the top 1% of nodes. Another network (SNAP/ca-GrQc) has a correlation coefficient greater than 0.9 on the top 1% of nodes.

The intersection distance values behave in a similar way, although there is not as much variation in the values. Among all the nodes, the smallest intersection distance is 1.81e-4 for the as-735 network and the largest is 0.142 for the ca-HepTh network. These networks also had the largest and smallest correlation coefficients, respectively, for the full set of nodes. For 5 of the 11 networks examined, the intersection distance value decreases when only the top 10% of nodes are considered and for all cases except for the Minnesota road network, it decreases when only the top 1% of nodes are considered.

It is interesting to note that the similarity between the two ranking methods is very different on the ca-GrQc and the ca-HepTh networks. The two networks are both arXiv collaboration networks from subsections of physics so, intuitively, one would assume that they behaved similarly. However, the two rankings are very different on the ca-HepTh network and are highly correlated on the ca-GrQc network. The ca-GrQc network has a spectral gap of approximately 7.5 while the spectral gap of ca-HepTh is approximately 8, only slightly larger. The relative spectral gaps are also comparable. Thus, it is clear that the spectral gap alone cannot be used to differentiate between the two ranking methods. It appears that while the two networks are both physics collaboration networks, there are significant structural differences between the two groups which cause the two ranking systems to behave very differently. Some insight can be gleaned by looking at the degree distributions of the two networks. Although the ca-HepTh network is almost twice as large as the ca-GrQc, the maximum degree on the network is only 65 while the maximum degree on the ca-GrQc network is 81. See Fig. 2.2 for the degree distributions of the two networks. Additionally, the total communicability scores achieved by nodes in the ca-GrQc network range from 2.7 to 8.5e19 (the subgraph centrality scores range from 1.5 to 1.6e18). In contrast, even with many more nodes, the total communicability scores of the ca-HepTh network have a smaller range, from 2.7 to 3.2e13 (the subgraph centrality scores range from 1.5 to 9.7e11). It appears that the wider range of scores in the ca-GrQc network helps to prevent rankings from being changed when the scores are perturbed by the addition of off-diagonal communicabilities. This can be observed when looking at the intersection distances between the two sets of rankings on the networks, which are plotted in Fig. 2.3. Overall, the intersection distances are much lower for the ca-GrQc network than for the ca-HepTh network. Additionally, for $k \leq 34$, isim_k(ca-GrQc)$= 0$, indicating that the first 34 nodes are ranked exactly the same. In contrast, isim_k(ca-HepTh)$= 0$ only for $k \leq 5$, after which there is a large jump in the intersection distances.

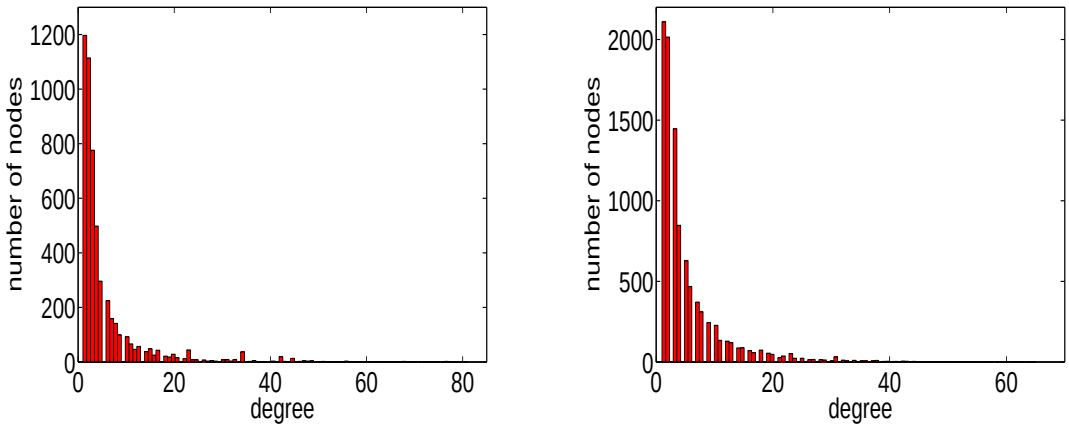

Figure 2.2: The degree distributions of the ca-GrQc (left) and the ca-HepTh (right) collaboration networks.

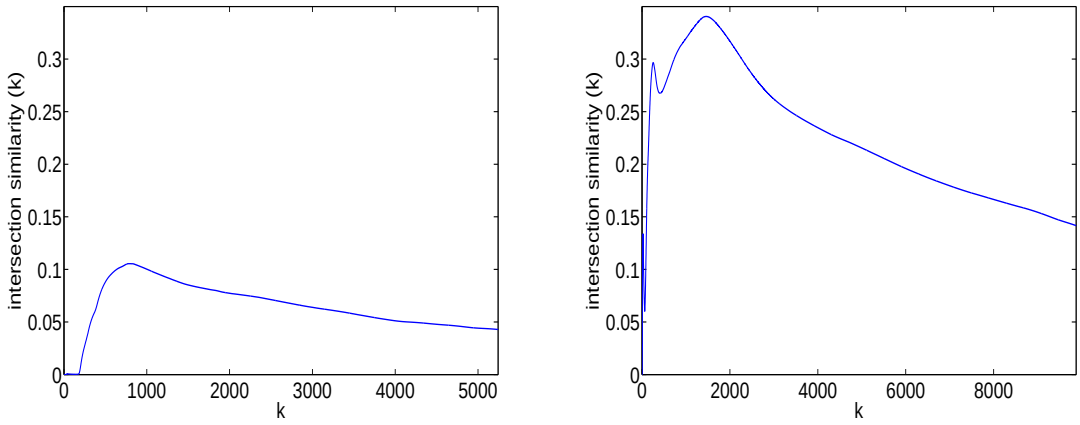

Figure 2.3: The intersection distance values (isim_k) of the ca-GrQc (left) and the ca-HepTh (right) collaboration networks.

Similar behavior can be observed on the various instances of the Erdös collaboration network. Erdos971, which is very small, shows a high correlation between the two rankings; indeed, the rankings of the top 10% of nodes are exactly the same. On the other instances of the collaboration network, however, the rankings are somewhat different, as can be seen from the relatively low values of the correlation coeffcients. The intersection distance values, while not very high, are somewhat higher than for most other networks. The maximum subgraph centrality and total communicability scores

of the Erdos972 network are the smallest of any of the Erdös collaboration subgraphs. The maximum subgraph centrality score is 1.18e05 and the maximum total centrality score is 9.20e06. By comparison, on the (much smaller) Erdos971 network, the maximum subgraph centrality score is 1.11e06. On the Erdos982 network, the maximum subgraph centrality score is 1.71e05 and on the Erdos992 network it is 2.47e05. Although the top 5 nodes of the Erdos972 network are exactly the same under the two ranking schemes, the relatively narrow range of possible scores means that the addition of off-diagonal values to the diagonal ones perturbs the rankings of the other nodes so much as to result in a relatively high value of the intersection distance among the top 1% of nodes.

As before, the spectral gap for these networks does not give much insight into the behavior of the two ranking schemes, unless it is really large; the largest spectral gap for this set of test problems occur for SNAP/as-735, and indeed here we observe a strong correlation and a small intersection distance between the two metrics. Conversely, for the (planar, fairly regular) Gleich/Minnesota network, the spectral gap is smallest and not surprisingly the correlation is very weak and the intersection distance for the top 1% of the nodes, $\mathrm{isim}_{1\%}$, is very high at 0.709.

When examining the (normalized) total network connectivities of the various networks (see Table 2.9), it can be seen that the ease of information sharing across the networks varies widely. Some networks, such as the collaboration networks ca-HepTh and ca-GrQc, have a high normalized $C(A)$ (8.80e17 and 1.06e11, respectively). The value is even higher for the SNAP/as-735 router network ($C(A)/n$ =3.64e19). The Minnesota road network, on the other hand, has a normalized $C(A)$ of only 14.13, indicating that the network is relatively poorly connected, as one would expect in a graph characterized by wide diameter, small bandwidth and high locality.

2.4.5 Identification of essential proteins in PPI network of yeast

One important application of node centrality measures is to rank nodes in protein-protein interaction networks (PPIs) in an attempt to determine which proteins are essential, in the sense that their removal would result in the death of the cell. The goal of such rankings is for as many of the top-ranked nodes as possible to correspond to essential proteins. In [42], various centrality measures were tested on their ability to identify essential proteins in the Yeast PPI network. It was shown that, among the centrality measures tested, subgraph centrality identified the highest percentage of essential proteins ranked in the top 30 nodes, identifying 18 essential proteins (in [42], subgraph centrality was said to identify 19 essential proteins, but this was later corrected [37]). When total communicability is used instead, the top 30 nodes are the same, so the same percentage of essential proteins are identified. The intersection distances between the two sets of rankings are displayed in Fig. 2.4. Here, it can be seen that the intersection distances are small for approximately the top 50 nodes, then they begin to rise. The two rankings are least similar for nodes ranked 200-500, then their similarity increases again. As already noted, total communicability rankings can be calculated much more quickly than subgraph centrality rankings (see also section 4.8). Although there are currently methodologies which do better in protein ranking (see [43] for example), our findings suggest that total communicability does provide valuable information about the relative importance of nodes in the network.

2.4.6 Further discussion of test results using real networks

The results just described indicate that in general the two centrality measures can produce significantly different rankings, even when one restricts the attention to the top 1% of nodes, and even for networks belonging to the same "family". As in the case of synthetic networks, a wider range of values in the two sets of centralities leads to

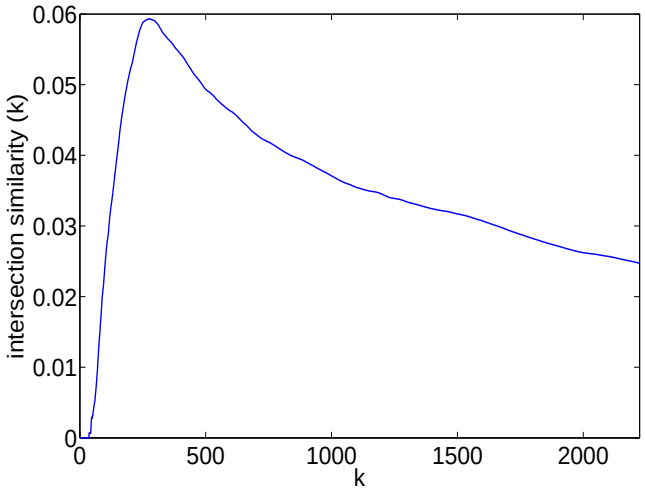

Figure 2.4: The intersection distance values (isim_k) of the Yeast PPI network.

stronger correlations between the corresponding rankings than in the case of a narrow range.

Two extreme cases are represented by the SNAP/as-735 and Gleich/Minnesota data sets. The first one exhibits a large value of the spectral gap, and thus (as expected) a strong correlation between the two rankings; the second one has tiny spectral gap and results in very weakly correlated rankings. For networks that fall somewhere in between these two extremes, the observed correlation coefficients can vary significantly. The subgraph centrality scores measure how "well-connected" a node is in the network as a whole while the communicability score between nodes i and j measures how well information travels between node i and node j. Thus, the total communicability of node i is a measure of how well information travels between node i and any node in the network (node i itself included). Although these two measures are closely related, they are not quite the same. This observation suggests that the two centrality measures reflect somewhat different structural properties of the networks. Thus, they should be applied in concert rather than in alternative of one another, unless computational considerations dictate otherwise.

2.5 Computational aspects

Table 2.10 lists the timings for calculating the matrix exponential directly using `expm` and for estimating the row and diagonal entries as described in Chapter 1, Section 1.7. The subgraph centralities were estimated using the `mmq` toolbox (with 5 iterations of the Lanczos algorithm per node), and the total communicabilities were estimated using the `funm_kryl` toolbox to estimate $e^A\mathbf{1}$ (using a very stringent stopping tolerance of 1e-16). These computations have been performed using Matlab Version 7.9.0 (R2009b) on a 2.4 GHZ Intel Core i5 processor with 4 GB of RAM. In general, the timings with `expm` increase for increasing number of nodes, but structural properties of the underlying graph, like the network diameter, can have a very significant impact on the computing times. For example, the yeast PPI network and the Minnesota road network have approximately the same number n of nodes (2224 and 2642, respectively), yet computing the matrix exponential for the yeast network takes almost 25 times longer than for the Minnesota road network. This appears to be due to the fact that the yeast network has a much smaller diameter than the Minnesota network, therefore the powers A^k of the adjacency matrix fill up much more quickly. Since the algorithm implemented in `expm` involves solving linear systems with polynomials in A as coefficient matrices, the execution time for sparse matrices with small diameter tends to be much higher than for matrices exhibiting a high degree of locality.

For the majority of the networks tested, using the `mmq` toolbox to estimate subgraph centrality was faster than using `expm`, frequently by far. The exceptions (Zachary Karate Club, Drug User, Erdos971, and Minnesota) were the networks with a small number of nodes and/or a high diameter.

The computation of the total communicabilities using `funm_kryl` was by far the fastest method for all networks tested, with the only exception of the tiny Zachary Karate Club network. In principle, this is a clear advantage of total communicability

Table 2.10: Timings (in seconds) to compute centrality rankings based on the diagonal and row sum of e^A for various test problems using different methods.

Graph	`expm`	`mmq`	`funm_kryl`
Zachary Karate Club	0.062	0.138	0.120
Drug User	0.746	2.416	0.363
Yeast PPI	47.794	9.341	0.402
Pajek/Erdos971	0.542	2.447	0.317
Pajek/Erdos972	579.214	35.674	0.410
Pajek/Erdos982	612.920	39.242	0.393
Pajek/Erdos992	656.270	53.019	0.325
SNAP/ca-GrQc	281.814	23.603	0.465
SNAP/ca-HepTh	2710.802	58.377	0.435
SNAP/as-735	2041.439	75.619	0.498
Gleich/Minnesota	1.956	10.955	0.329

over subgraph centrality. However, as we saw, the two methods often result in rather different rankings, therefore we cannot simply replace subgraph centrality with total communicability.

2.5.1 A large-scale example

In addition to the test results discussed above, we performed tests with the digraph of Wikipedia (as of June 6, 2011), where nodes correspond to entries and directed links to hyperlinks from one entry to another. In this case, the entries of $e^A\mathbf{1}$ provide a ranking of the hubs in the networks, see [8]. This graph contains 4,189,503 nodes and 67,197,636 links, and it is prohibitively large for centrality measures based on estimating the diagonals of the matrix exponential. For this reason, we limit ourselves to computations using the `funm_kryl` toolbox to estimate the row sum vector $e^A\mathbf{1}$. The restart parameter was set to 10 and we allowed a maximum of 50 restarts. The run time to obtain the rankings on a parallel system comprising 24 Intel(R) Xeon(R) E5-2630 2.30GHz CPU(s) was 216.7 seconds. This shows that centrality calculations using total communicability are quite feasible even for large networks.

2.6 Resolvent-based centrality measures

There are matrix functions other than the matrix exponential that may be used to calculate subgraph centrality and subgraph communicability scores. The most common of these is the matrix resolvent. Like the matrix exponential, $[(I-\alpha A)^{-1}]_{ii}$ counts the number of closed walks centered at node i and $\sum_{j=1}^{n}[(I-\alpha A)^{-1}]_{ij}$ counts all walks between node i and all other nodes in the network. In this case, however, a walk of length k is penalized by a factor of α^k. One drawback of the use of the matrix resolvent in determining centrality rankings is the need to choose the value of α; also, different values of α can lead to different rankings. For the purposes of the experiments below, we select $\alpha = \frac{0.85}{\lambda_{\max}(A)}$ (similar to the choice of parameter in PageRank [73]).

Resolvent-based total network communicability can also be evaluated. As when using the matrix exponential (cf. section 2.3), the resolvent-based total network communicability is an upper bound for the resolvent-based Estrada index. In the following, $C_r(A) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[(I-\alpha A)^{-1}\right]_{ij}$ denotes the resolvent-based total communicability of a network. The following Proposition can be easily proved along the same lines as Proposition 4.4.1.

Proposition 2.6.1 Let A be the adjacency matrix of a simple, undirected network on n vertices. Then for any $0 < \alpha < \frac{1}{\|A\|_2}$,

$$EE_r(A) := \mathrm{Tr}\left[(I-\alpha A)^{-1}\right] \leq C_r(A) \leq \frac{n}{1-\alpha\|A\|_2}.$$

For an undirected network, $\lambda_{\max}(A) = \lambda_1$ can replace $\|A\|_2$ in the upper bound above.

The resolvent-based subgraph centrality and total communicability rankings were compared on the same two sets of synthetic networks used for the tests in Section 2.4.1.

Table 2.11 lists the average correlation coefficient between the subgraph centrality and total communicability rankings for the nodes in networks constructed using the

Table 2.11: Comparison using correlation coefficients of rankings based on the diagonal entries and row sums of $(I - \alpha A)^{-1}$ for 1000-node scale-free networks of various parameters built using the `pref` function in the CONTEST Matlab toolbox. For each instance, the results are measured for $\alpha = \frac{0.85}{\lambda_{\max}(A)}$. The values reported are the averages over 20 matrices with the same parameters.

d	cc
1	0.292
2	0.370
3	0.442
4	0.486
5	0.536
6	0.583
7	0.607
8	0.638
9	0.667

d	cc
10	0.691
20	0.840
30	0.890
40	0.917
50	0.933
60	0.942
70	0.949
80	0.954
90	0.958
100	0.962

d	cc
110	0.964
120	0.965
130	0.968
140	0.970
150	0.971
160	0.973
170	0.973
180	0.975
190	0.976
200	0.976

Table 2.12: Intersection distance comparison of rankings based on the diagonal entries and row sums of $(I - \alpha A)^{-1}$ for 1000-node scale-free networks of various parameters built using the `pref` function in the CONTEST Matlab toolbox. For each instance, the results are measured for $\alpha = \frac{0.85}{\lambda_{\max}(A)}$. The values reported are the averages over 20 matrices with the same parameters.

d	isim	$\text{isim}_{10\%}$
1	0.186	0.491
2	0.205	0.364
3	0.192	0.235
4	0.179	0.173
5	0.163	0.126
6	0.150	0.102
7	0.137	0.082
8	0.124	0.068
9	0.115	0.059

d	isim	$\text{isim}_{10\%}$
10	0.105	0.051
20	0.055	0.020
30	0.035	0.012
40	0.025	0.007
50	0.019	0.005
60	0.015	0.004
70	0.012	0.003
80	0.010	0.002
90	0.009	0.002
100	0.007	0.001

d	isim	$\text{isim}_{10\%}$
110	0.006	0.001
120	0.005	7.12e-4
130	0.005	6.98e-4
140	0.004	5.74e-4
150	0.004	5.62e-4
160	0.003	3.69e-4
170	0.003	4.25e-4
180	0.003	3.11e-4
190	0.003	3.16e-4
200	0.002	4.00e-4

preferential attachment model (function `pref` in CONTEST) and Table 2.12 lists the intersection distances for all the nodes and for the top 10% of the nodes. For small values of d ($1 \leq d \leq 3$), the correlation coefficients between the two sets of rankings using the matrix resolvent are close to those using the matrix exponential. However, when using the matrix exponential the average correlation coefficient was found to be

Table 2.13: Comparison using correlation coefficients of rankings based on the diagonal entries and row sums of $(I - \alpha A)^{-1}$ for 1000-node small world networks of various parameters built using the `smallw` function with $p = 0.1$ in the CONTEST Matlab toolbox. For each instance, the results are measured for $\alpha = \frac{0.85}{\lambda_{\max}(A)}$. The values reported are the averages over 20 matrices with the same parameters.

d	cc
1	0.065
2	0.023
3	0.052
4	0.052
5	0.052
6	0.051
7	0.062
8	0.037
9	0.050

d	cc
10	0.063
20	0.078
30	0.080
40	0.135
50	0.144
60	0.141
70	0.144
80	0.133
90	0.248
100	0.190

d	cc
110	0.294
120	0.246
130	0.275
140	0.311
150	0.312
160	0.321
170	0.301
180	0.293
190	0.354
200	0.300

greater than 0.9 for all $d \geq 4$, and exactly 1 for all $d \geq 8$. Using the matrix resolvent the correlation coefficient grows as d increases, but somewhat more slowly than for the matrix exponential. The intersection distances are also larger for all values of d when the matrix resolvent is used, although they also decrease as d increases. Moreover, we did not find a single instance where the two methods produced *exactly* the same rankings.

For the small world networks, all values $1 \leq d \leq 10$ as well as as all multiples of 10 with $20 \leq 10 \leq 200$ were tested. For each d, twenty networks were tested. The averages of the correlation coefficients between the subgraph centrality and total communicability rankings can be found in Table 2.13 and the average intersection distances for both all the nodes and the top 10% of the nodes can be found in Table 2.14. As was the case for the matrix exponential, the two methods (diagonal entries and row sums) using the matrix resolvent exhibit much weaker correlations for this class of networks than for the preferential attachment networks; indeed, the correlations tend to be even smaller for the resolvent than for the exponential. For $d = 1$, the average correlation is 0.065 and the average intersection distance was 0.040 using the resolvent, compared

Table 2.14: Intersection distance comparison of rankings based on the diagonal entries and row sums of $(I - \alpha A)^{-1}$ for 1000-node small world networks of various parameters built using the `smallw` function with $p = 0.1$ in the CONTEST Matlab toolbox. For each instance, the results are measured for $\alpha = \frac{0.85}{\lambda_{\max}(A)}$. The values reported are the averages over 20 matrices with the same parameters.

d	isim	$\text{isim}_{10\%}$
1	0.040	0.149
2	0.070	0.189
3	0.085	0.241
4	0.091	0.269
5	0.098	0.301
6	0.104	0.318
7	0.126	0.361
8	0.135	0.414
9	0.149	0.413

d	isim	$\text{isim}_{10\%}$
10	0.156	0.435
20	0.207	0.508
30	0.198	0.517
40	0.204	0.571
50	0.207	0.621
60	0.191	0.588
70	0.181	0.582
80	0.189	0.607
90	0.156	0.597
100	0.179	0.585

d	isim	$\text{isim}_{10\%}$
110	0.147	0.541
120	0.148	0.553
130	0.160	0.554
140	0.142	0.560
150	0.123	0.542
160	0.121	0.539
170	0.124	0.517
180	0.125	0.512
190	0.114	0.504
200	0.123	0.504

to a correlation of 0.177 and an intersection distance of 0.015 using the exponential. For the values of d tested, the highest average correlation coefficient was 0.354, for $d = 190$. When looking at the intersection distances for other values of d, the picture is somewhat different. Comparing Table 2.14 with Table 2.4, we see that for small d the intersection distance between the two ranking schemes tends to be somewhat higher with the matrix exponential than with the resolvent. However, as d increases the intersection distance eventually drops with the matrix exponential, but not with the resolvent. This is true both when looking at the ranking of all the nodes and when looking at only the top 10%.

Next, we consider tests with real-world networks. As shown in Table 2.15, the correlation coefficients between the two ranking systems for the whole set of nodes were higher (in a majority of cases) using the matrix resolvent than they were using the matrix exponential. (Again, a "–" signifies that correlation coefficients could not be computed due to the fact that the two ranking schemes produced different lists of nodes.) Only the Erdos971, as-735, and the Minnesota networks had a higher correlation coefficient between the two ranking systems under the exponential than under the

Table 2.15: Comparison using correlation coefficients of rankings based on the diagonal entries and row sums of $(I - \alpha A)^{-1}$ with $\alpha = \frac{0.85}{\lambda_{\max}(A)}$ for various real-world networks.

Graph	cc	cc_{10}	cc_1
Zachary Karate Club	0.589	1	1
Drug User	0.189	–	–
Yeast PPI	0.177	–	–
Pajek/Erdos971	0.233	–	1
Pajek/Erdos972	0.215	–	–
Pajek/Erdos982	0.207	–	–
Pajek/Erdos992	0.197	–	—
SNAP/ca-GrQc	0.070	–	–
SNAP/ca-HepTh	0.072	–	–
SNAP/as-735	0.204	–	—
Gleich/Minnesota	0.019	–	–

Table 2.16: Intersection distance comparison of rankings based on the diagonal entries and row sums of $(I - \alpha A)^{-1}$ with $\alpha = \frac{0.85}{\lambda_{\max}(A)}$ for various real-world networks.

Graph	isim	$\text{isim}_{10\%}$	$\text{isim}_{1\%}$
Zachary Karate Club	0.061	0	0
Drug User	0.125	0.145	0.069
Yeast PPI	0.204	0.363	0.187
Pajek/Erdos971	0.080	0.050	0
Pajek/Erdos972	0.110	0.273	0.263
Pajek/Erdos982	0.109	0.269	0.264
Pajek/Erdos992	0.109	0.271	0.247
SNAP/ca-GrQc	0.047	0.122	0.033
SNAP/ca-HepTh	0.058	0.159	0.236
SNAP/as-735	0.247	0.513	0.271
Gleich/Minnesota	0.102	0.301	0.557

Table 2.17: Comparison of the normalized resolvent-based Estrada index $EE_r(A)/n$ and total network connectivity $C_r(A)/n$ for various real-world networks. Here, $f(A) = (I - \alpha A)^{-1}$ with $\alpha = \frac{0.85}{\lambda_{\max}(A)}$.

Graph	normalized $EE_r(A)$	normalized $C_r(A)$
Zachary Karate Club	1.21	5.13
Drug User	1.03	2.36
Yeast PPI	1.03	2.17
Pajek/Erdos971	1.03	2.44
Pajek/Erdos972	1.01	1.70
Pajek/Erdos982	1.01	1.66
Pajek/Erdos992	1.01	1.60
SNAP/ca-GrQc	1.00	1.21
SNAP/ca-HepTh	1.01	1.24
SNAP/as-735	1.00	1.86
Gleich/Minnesota	1.27	3.44

matrix resolvent. This can be understood when looking at the normalized Estrada indexes and total network communicabilities in Table 2.17. The smaller the factor α, the more it minimizes the contribution of the network data from A to the scores produced by the diagonal entries or row sums of $(I - \alpha A)^{-1}$. This can also be seen by noticing that as $\alpha \to 0$, $(I - \alpha A)^{-1}$ approaches the identity. In these experiments, $\alpha = \frac{0.85}{\lambda_{\max}(A)}$. However, this also means that for the networks tested with a large maximum eigenvalue (such and ca-GrQc, ca-HepTh, and as-735) α is quite small, causing the resulting subgraph centrality scores to be small and, consequently, close together. In the case of a network with a small maximum eigenvalue (such as the Minnesota network), the effect of α is not as pronounced. The compression of the score values means that a perturbation of the scores (such as occurs when switching from subgraph centrality scores to total communicability scores) has a large effect on the node rankings, especially for the higher ranked nodes.

When only the top 1% of nodes were considered, the exponential subgraph centrality and exponential total communicability rankings were much closer together than their resolvent counterparts, where often the top 1% of nodes were not even the same.

This seems to indicate that when using the resolvent, the subgraph centrality and total communicability tend to rank the less important nodes more similarly than they do under the matrix exponential. Under the matrix exponential, the two rankings seem to agree more closely on the important nodes than they do when using the resolvent. This can also be seen when looking at the intersection distance, which gives more weight to differences in the top ranked nodes than in the lower ranked nodes. For all networks except ca-HepTh, the intersection distance between the two rankings is smaller when using the exponential than when using the resolvent. When looking at the top 1% of nodes, the intersection distances are also smaller (often much smaller) in the case of the exponential, for all except three of the networks. The exceptions are the Minnesota road network (which has a large intersection distance on the top 1% of nodes for both the exponential and the resolvent) and the Zachary Karate Club and Erdos971 networks (which have $\text{isim}_{1\%} = 0$ for both cases).

Another observation that can be made is that the resolvent-based total network communicability $C_r(A)$ is unable to discriminate between highly connected networks and poorly connected ones, in stark contrast with the exponential-based one. For instance, in the case of the Minnesota road network α is relatively large (since λ_1 is small for this graph), hence the off-diagonal contributions to $C_r(A)$ are more significant than for other networks where λ_1 is large (thus forcing a small value of α, leading to a resolvent very close to the identity matrix). Thus, only the exponential-based total network communicability should be used when comparing different networks in terms of ease of communication.

When the identification of essential proteins in the Yeast PPI network is considered using resolvent-based total communicability, the results are comparable to those using the exponential. The resolvent-based total communicability rankings with $\alpha = \frac{0.85}{\lambda_{\max}(A)}$ identified 17 essential proteins in the top 30 (as compared to 18 identified by exponential subgraph centrality and total communicability). The resolvent-based subgraph

centrality, however, identified 19 essential proteins in the top 30, slightly outperforming the other methods.

Concerning the computational complexity, when dealing with large networks the use of the conjugate gradient method (possibly with some type of preconditioning) to solve the linear system $(I - \alpha A)\mathbf{x} = \mathbf{1}$ is orders of magnitude faster than trying to estimate the diagonal entries of $(I - \alpha A)^{-1}$. For certain networks, Chebyshev semi-iteration can be even faster [11]. Thus, as was the case for the matrix exponential, rankings based on total communicability (row sums) are a lot cheaper than the rankings based on subgraph centrality (diagonals). Once again, however, the two ranking methods in general produce different rankings, so one should not choose between the two based solely on computational cost.

2.7 Summary and conclusions

We have examined the use of total communicability as a method for ranking the importance of nodes in a network. Like the subgraph centrality ranking, the total communicability ranking using the matrix exponential counts the number of walks starting at a given node, weighing walks of length k by a penalization factor of $\frac{1}{k!}$. However, instead of only counting closed walks, it counts all walks between the given node and every node in the network. If the matrix resolvent is used, the weight on the walks becomes α^k for a chosen parameter α in a certain range. There are various classes of graphs on which it can be shown that the two exponential-based rankings are always identical or in very good agreement; for instance, certain types of simple regular graphs and Erdös–Renyi random graphs with large spectral gap. However, as is well known, these classes are not realistic models of real-world complex networks.

The two sets of rankings (total communicability and subgraph centrality) have been used to rank the nodes of networks corresponding to both real and synthetic data sets.

The synthetic data sets were constructed using the preferential attachment (Barabási–Albert) and the small world (Watts–Strogatz) models, corresponding to the functions `pref` and `smallw` of the CONTEST toolbox for Matlab. Good agreement between the two ranking methods was observed on the networks obtain with the preferential attachment method, especially as the density of the graphs increased. More pronounced differences between the rankings produced with the two methods were observed in the case of small world networks. Overall, the two importance rankings matched more closely when the matrix exponential was used than when under the matrix resolvent.

We also presented the results of experiments with real-world networks including social networks, citation networks, PPI networks, and infrastructure (transportation) networks. Here we found that overall, the two (complete) sets of rankings were closer to each other when the matrix resolvent was used instead of the matrix exponential. However, when only the top 1% of nodes was examined, the rankings matched more closely when the matrix exponential was used. This suggests that, for the networks tested, the resolvent-based rankings match more closely on "unimportant" (low-ranked) nodes and the exponential-based rankings exhibit more agreement on the "important" (top-ranked) nodes.

In general, there is no simple way to compare two ranking schemes and determine that one is "better" than the other. However, the total communicability rankings take into account more of the network topology than the subgraph centrality rankings (all walks starting at node i versus all closed walks starting at node i). This added information often (but not always) changes the ranking of the nodes to a certain degree, although there are many cases where there is still a strong similarity between the two sets of rankings. The main benefit of using total communicability to rank the nodes is that the ranking can be estimated very quickly using Krylov subspace methods. Indeed, as the Wikipedia graph calculation described in section 2.5.1 shows, for very large networks only the total communicability (row sum) method is computationally feasible,

the subgraph centrality ranking being prohibitively expensive to compute. Even if total communicability cannot always be recommended as a cheaper alternative to subgraph centrality, it provides valuable information about the network and can be used along with other ranking schemes.

Finally, we have introduced the total communicability of a network as a global measure of connectivity and of the ease of information flow on a given network. This measure can be computed quickly even for very large networks, and could be of interest in the design of communication networks.

3 Robustness of Parameterized Centrality Rankings

3.1 Introduction

It has been observed that often the rankings produced by various centrality measures are strongly correlated on many real-world networks [8, 10, 62]. In this chapter, we analyze the relationships between the following centrality measures:

- degree centrality: $C_d(i) = d_i$,
- eigenvector centrality: $C_{ev}(i) = \mathbf{q}_1(i)$,
- exponential subgraph centrality: $SC_i(\beta) = [e^{\beta A}]_{ii}$,
- resolvent subgraph centrality: $RC_i(\alpha) = [(I - \alpha A)^{-1}]_{ii}$,
- total communicability: $TC_i(\beta) = [e^{\beta A}\mathbf{1}]_i = \mathbf{e}_i^T e^{\beta A}\mathbf{1}$
- Katz centrality: $K_i(\alpha) = [(I - \alpha A)^{-1}\mathbf{1}]_i = \mathbf{e}_i^T(I - \alpha A)^{-1}\mathbf{1}$

There are many more ranking methods we could have considered for this analysis (including some that are considered in other parts of this thesis), yet we restricted our analysis to the above six. The choice of which of the many centrality measures to study and why is something that must be considered carefully (see [25], for example). In this chapter, we restricted our analysis to commonly (and successfully) used centrality measures that had a formulation in terms of a function on the adjacency matrix of the

network, and that were analytically tractable. We additionally restricted our scope to centrality measures that we could demonstrate were related to each other.

We show that the parameters α (in the case of resolvent subgraph and Katz centrality) and β (in the case of exponential subgraph centrality and total communicability) act as tunable parameters, interpolating between degree and eigenvector centrality. This interpolation stabilizes to eigenvector centrality very quickly when the spectral gap, $\lambda_1 - \lambda_2$, is large. This results helps to explain why degree and eigenvector centrality are strongly correlated on many real-world networks.

3.2 Relationship between various centrality measures

One difficulty in measuring the "importance" of a node in a network using centrality is that it is not clear which of the many centrality measures should be used. Additionally, it is not clear a priori whether two centrality measures will give the same node rankings on a given network. When using exponential or resolvent based centrality measures, the need to choose the value of a parameter (β and α respectively) adds another layer of difficulty. Different choices of α and β produce different centrality scores and can lead to different node rankings. However, experimentally, it has been seen that different centrality measures often provide rankings that are highly correlated [8, 10, 62]. Moreover, in most cases, the rankings do not change too much for different choices of α and β.

This robustness of certain centrality rankings can be explained, in part, by the following two theorems, which relate degree and eigenvector centrality to exponential subgraph and resolvent centrality, respectively:

Theorem 3.2.1 Let $G = (V, E)$ be a connected, undirected network with adjacency matrix A. Let $SC_i(\beta) = [e^{\beta A}]_{ii}$ be the subgraph centrality of node i and $\mathbf{SC}$ be the vector of subgraph centralities. Then,

(i) as $\beta \to 0+$, the rankings produced by $\mathbf{SC}(\beta)$ converge to those produced by $\mathbf{C_d}$, the vector of degree centralities,

(ii) as $\beta \to \infty$, the rankings produced by $\mathbf{SC}(\beta)$ converge to those produced by $\mathbf{C_{ev}}$, the vector of eigenvector centralities.

Proof: To prove (i), consider the Taylor expansion of $SC_i(\beta)$:

$$SC_i(\beta) = [e^{\beta A}]_{ii} = 1 + \beta A_{ii} + \frac{\beta^2 [A^2]_{ii}}{2!} + \frac{\beta^3 [A^3]_{ii}}{3!} + \cdots]_{ii} = 1 + 0 + \frac{\beta^2}{2!} d_i + \frac{\beta^3}{3!} [A^3]_{ii} + \cdots$$

Let $\boldsymbol{\phi}(\beta) = \frac{2!}{\beta^2}[\mathbf{SC}(\beta) - 1]$. The rankings produced by $\boldsymbol{\phi}(\beta)$ will be the same as those produced by $\mathbf{SC}(\beta)$, as the scores for each node have been shifted and scaled in the same way. Now,

$$\phi_i(\beta) = \frac{2!}{\beta^2}[SC_i(\beta) - 1] = d_i + \frac{2\beta}{3!}[A^3]_{ii} + \frac{2\beta^2}{4!}[A^4]_{ii} + \cdots$$

which tends to d_i as $\beta \to 0+$. Thus, as $\beta \to 0+$, the rankings produced by the subgraph centrality scores reduce to those produced by the degrees.

To prove (ii), consider the expansion of $SC_i(\beta)$ in terms of the eigenvalues and eigenvectors of A:

$$SC_i(\beta) = \sum_{k=1}^{n} e^{\beta \lambda_k} \mathbf{q}_k(i)^2 = e^{\beta \lambda_1} \mathbf{q}_1(i)^2 + \sum_{k=2}^{n} e^{\beta \lambda_k} \mathbf{q}_k(i)^2.$$

Let $\boldsymbol{\psi}(\beta) = \frac{1}{e^{\beta \lambda_1}} \mathbf{SC}(\beta)$. As in the proof of (i), the rankings produced by $\boldsymbol{\psi}(\beta)$ are the same as those produced by $\mathbf{SC}(\beta)$, since the scores for each node have been scaled in the same way. Next,

$$\psi_i(\beta) = \mathbf{q}_1(i)^2 + \sum_{k=2}^{n} e^{\beta(\lambda_k - \lambda_1)} \mathbf{q}_k(i)^2.$$

Since $\lambda_1 > \lambda_k$ for $2 \leq k \leq n$, as $\beta \to \infty$, $\psi_i(\beta) \to \mathbf{q}_1(i)^2$. By the Perron-Frobenius

Theorem, $\mathbf{q}_1 > 0$, so the rankings produced by $\mathbf{q}_1(i)^2$ are the same as those produced by $\mathbf{q}_1(i)$. Thus, as $\beta \to \infty$, the rankings produced by $SC(\beta)$ converge to those produced by C_{ev}. □

Theorem 3.2.2 Let $G = (V, E)$ be a connected, undirected network with adjacency matrix A. Let the resolvent subgraph centrality of node i be given by $RC_i(\alpha) = [(I - \alpha A)^{-1}]_{ii}$ and $\mathbf{RC}$ be the vector of resolvent centralities. Then,

(i) as $\alpha \to 0+$, the rankings produced by $\mathbf{RC}(\alpha)$ converge to $\mathbf{C_d}$, the vector of degree centralities,

(ii) as $\alpha \to \frac{1}{\lambda_1}-$, the rankings produced by $\mathbf{RC}(\alpha)$ converge to $\mathbf{C_{ev}}$, the vector of eigenvector centralities.

The proof of Theorem 3.2.2 follows the same arguments as that of Theorem 3.2.1

Similar relationships hold between degree and eigenvector centralities and the centrality measures based on row sums (exponential total communicability and Katz centrality). These relationships can be found in the following two theorems.

Theorem 3.2.3 Let $G = (V, E)$ be a connected, undirected network with adjacency matrix A. Let $TC_i(\beta) = [e^{\beta A}\mathbf{1}]_i$ be the total communicability of node i and $\mathbf{TC}(\beta)$ be the vector of total communicabilities. Then,

(i) as $\beta \to 0+$, the rankings produced by $\mathbf{TC}(\beta)$ converge to those produced by $\mathbf{C_d}$, the vector of degree centralities,

(ii) as $\beta \to \infty$, the rankings produced by $\mathbf{TC}(\beta)$ converge to those produced by $\mathbf{C_{ev}}$, the vector of eigenvector centralities.

Theorem 3.2.4 Let $G = (V, E)$ be an connected, undirected network with adjacency matrix A. Let the Katz centrality of node i be given by $K_i(\alpha) = [(I - \alpha A)^{-1}\mathbf{1}]_i$ and let $\mathbf{K}(\alpha)$ denote the vector of Katz centralities. Then,

(i) as $\alpha \to 0+$, the rankings produced by $\mathbf{K}(\alpha)$ converge to $\mathbf{C_d}$, the vector of degree centralities,

(ii) as $\alpha \to \frac{1}{\lambda_1}-$, the rankings produced by $\mathbf{K}(\alpha)$ converge to $\mathbf{C_{ev}}$, the vector of eigenvector centralities.

The proofs of Theorems 3.2.3 and 3.2.4 follow similar arguments as the proof of Theorem 3.2.1.

3.3 Interpretation

The centrality scores which we are considering in this paper are all based on walks in the network. The degree centrality of a node i counts the number of walks of length one starting at i (the degree of i). In contrast, the eigenvector centrality of node i gives the limit as k goes to infinity of the percentage of walks of length k which start at node i (see [40, p. 127] and [30]). Thus, the degree centrality of node i measures the local influence of i and the eigenvector centrality measures the global influence of i.

Exponential subgraph centrality and total communicability take both local and global influence into account, weighting walks of length k by $\frac{\beta^k}{k!}$, $\beta > 0$. As β tends toward 0, the weights corresponding to larger k decay faster and shorter walks become more important in the centrality rankings. In the limit as $\beta \to 0$, walks of length one dominate the centrality scores and the rankings converge to the degree centrality rankings. As β increases to infinity, given a fixed walk length k, the weights of walks of length k increase more rapidly than those of shorter walks. In the limit as $\beta \to \infty$, walks of infinite length dominate and the centrality rankings converge to those of eigenvec-

tor centrality. When resolvent subgraph centrality and Katz centrality are considered, walks of length k are weighted by a factor of α^k, $0 < \alpha < \frac{1}{\lambda_1}$. Again, as $\alpha \to 0$, shorter walks dominate the rankings and, in the limit, the rankings converge to those produced by degree centrality. As $\alpha \to \frac{1}{\lambda_1}$, longer walks dominate and, in the limit, the rankings converge to those produced by eigenvector centrality.

In these parameterized centrality rankings, the parameters α and β should be viewed as a method for tuning between rankings based on local influence (short walks) and those based on global influence (long walks). In applications where local influence is most important, degree centrality will often be equivalent to any of the parameterized centrality rankings with α or β very small. Similarly, when global influence is the only important factor, parameterized centrality rankings with α or β large will often be equivalent to eigenvector centrality. Exponential subgraph centrality and total communicability, along with resolvent subgraph centrality and Katz centrality, are most useful when both local and global influence need to be considered in the ranking of nodes in a complex network. In order to achieve this, moderate values of β and α must be used. Differences between the rankings produced by these methods with various choices of parameter and those produced by degree and eigenvector centrality in both synthetic and real world networks can be found in Section 3.5.

The rate at which these rankings converge to eigenvector centrality (as $\alpha \to \frac{1}{\lambda_1}$ and $\beta \to \infty$) depends on the spectral gap, $\lambda_1 - \lambda_2$. If the (relative) spectral gap is large, all four of the parameterized centrality rankings will converge to eigenvector centrality very quickly as α and β increase. Thus, in networks with a large enough spectral gap, eigenvector centrality should be used instead of a method based on the exponential or resolvent of the adjacency matrix. However, it is difficult to tell *a priori* when $\lambda_1 - \lambda_2$ is "large enough".

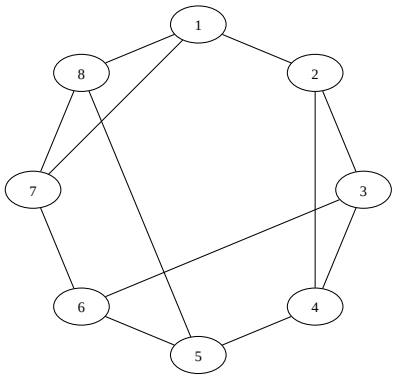

Figure 3.1: A 3-regular graph on 8 nodes which is not walk-regular.

3.4 A small example

Consider the eight vertex, 3-regular graph shown in Figure 3.1. This graph is not walk-regular. This can be verified by noting that nodes 1, 7, and 8 make up a triangle, as do nodes 2, 3, and 4, but nodes 5 and 6 are not involved in any triangles. Consequently, for all nodes $i \neq 5, 6$, there are two closed walks of length 3 beginning at node i while nodes 5 and 6 are not involved in any closed walks of length three.

Due to the regularity of this network, all nodes are equivalent under degree centrality and under eigenvector centrality. That is, both on the local level (walks of length one) and the global level (walks of infinite length), all nodes have the same importance. Furthermore, since all the powers A^p of the adjacency matrix A of a k-regular graph have constant row sums ($= k^p$), Katz and total communicability centrality are also unable to discriminate between nodes, regardless of α and β.

In contrast, due to the fact that the graph is not walk-regular, subgraph centrality (whether exponential or resolvent-based) is able to discriminate between the nodes and thus to provide rankings. For example, using the diagonals of e^A leads to a 4-way tie at the top (nodes 3,4,7,8), followed by nodes 1 and 2 (tied) followed by nodes 5 and 6 (tied). The same ranking is obtained using the diagonal entries of $(I - \alpha A)^{-1}$ with, say, $\alpha = 0.25$ (the dominant eigenvalue of A is $\lambda_1 = 3$).

Hence, this simple academic example shows that subgraph centrality has greater discriminatory power than the other ranking methods considered here.[1]

3.5 Numerical Experiments on real data

In this section, we examine the relationship between the centrality measures under consideration on a selection of real world networks from a variety of sources. Unless otherwise specified, all of the numerical experiments were performed in Matlab version 7.9.0 (R2009b) on a MacBook Pro running OS X Version 10.7.5 with a 2.4 GHZ Intel Core i5 processor and 4GB of RAM. The networks used are those described in Section 2.4.4. Basic data on these networks can be found in Table 2.6.

3.5.1 Exponential subgraph centrality and total communicability

We examined the effects of changing β on the exponential subgraph centrality and total communicability rankings of nodes in a variety of undirected real world networks, as well as their relation to degree and eigenvector centrality. The measures were calculated for eleven networks (descriptions of which can be found in Table 2.6). Although the only restriction on β is that it must be greater than zero, there is often an implicit upper limit that may be problem dependent. For the analysis in this section, we impose the following limits: $0.1 \leq \beta \leq 10$. To examine the sensitivity of the exponential subgraph centrality and total communicability rankings, we calculate both sets of scores and rankings for various choices of β. The values of β tested are: 0.1, 0.5, 1, 2, 5, 8 and 10.

The rankings produced by the matrix exponential based centrality measures for all choices of β were compared to those produced by degree centrality and eigenvector

[1] It is obvious that if a graph is walk-regular, none of the centrality measures considered in this paper can discriminate between its nodes. For subgraph centrality, the converse of this statement is an open conjecture of Estrada; see, e.g., [38, 44, 87, 89] as well as [6] for further discussion of this and related questions.

centrality, using the intersection distance method as described in Section 3.5. Plots of the intersection distances for the rankings produced by various choices of β with those produced by degree or eigenvector centrality can be found in Figures 3.2 and 3.3. The intersection distances for rankings produced by successive choices of β can be found in Figure 3.4.

In Figure 3.2, the rankings produced by exponential subgraph centrality and total communicability are compared to those produced by degree centrality. For small values of β, both sets of rankings based on the matrix exponential are very close to those produced by degree centrality (low intersection distances). When $\beta = 0.1$, the largest intersection distance between the degree centrality rankings and the exponential subgraph centrality rankings for the networks examined is slightly less than 0.2 (for the Minnesota road network). The largest intersection distance between the total communicability rankings with $\beta = 0.1$ and the degree centrality rankings is 0.3 (for the as-735 network). In general, the (diagonal based) exponential subgraph centrality rankings tend to be slightly closer to the degree rankings than the (row sum based) total communicability rankings for low values of β. As β increases, the intersection distances increase then level off. The rankings of nodes in networks with a very large (relative) spectral gap, such as the karate, Erdos971 and as-735 networks, stabilize extremely quickly, as expected. The one exception to the stabilization is the intersection distances between the degree centrality rankings and exponential subgraph centrality (and total communicability rankings) of nodes in the Minnesota road network. This is expected, as the small spectral gap for the Minnesota road network means that it will take longer for the exponential subgraph centrality (and total communicability) rankings to stabilize as β increases.

The rankings produced by exponential subgraph centrality and total communicability are compared to those produced by eigenvector centrality for various values of β in Figure 3.3. When β is small, the intersection distances are large but, as β increases, the

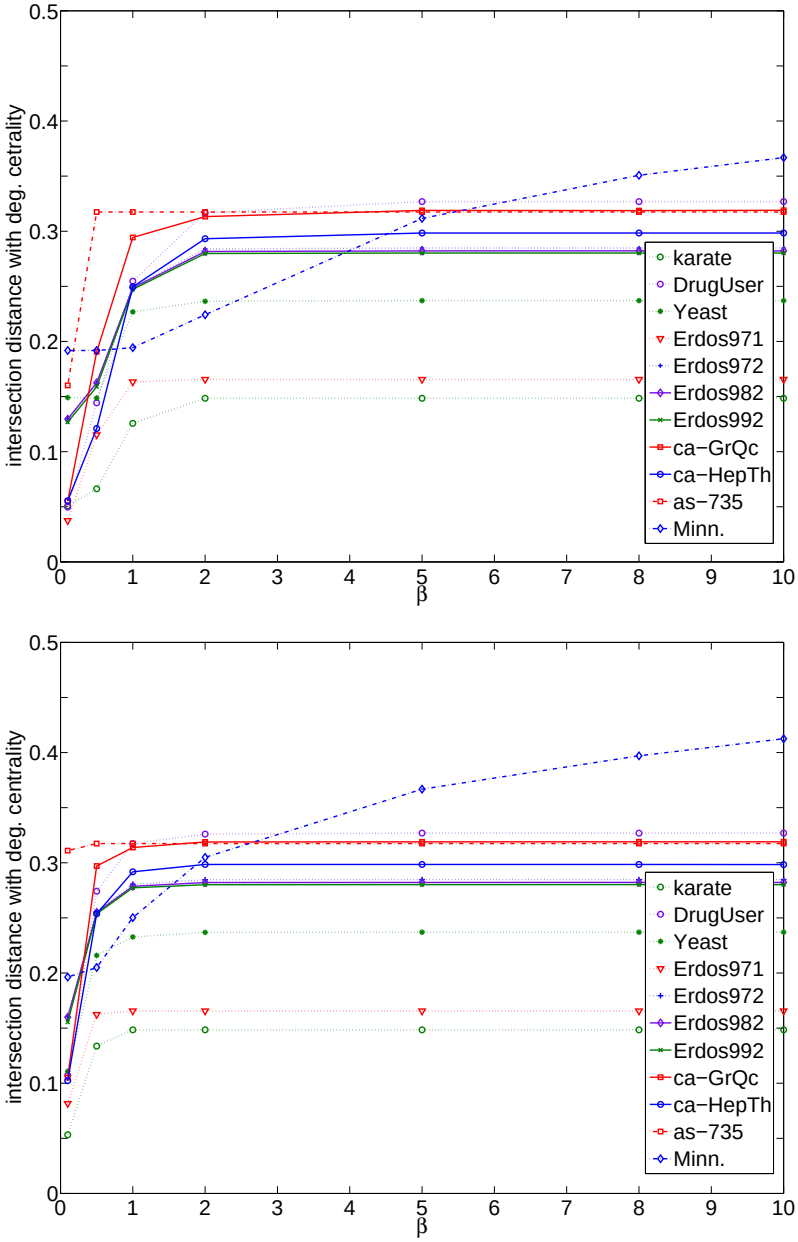

Figure 3.2: The intersection distances between degree centrality and the exponential subgraph centrality (top) or total communicability (bottom) rankings of the nodes in the networks in Table 2.6.

intersection distances quickly decrease. When $\beta = 2$, they are essentially zero for all but one of the networks examined. Again, the outlier is the Minnesota road network.

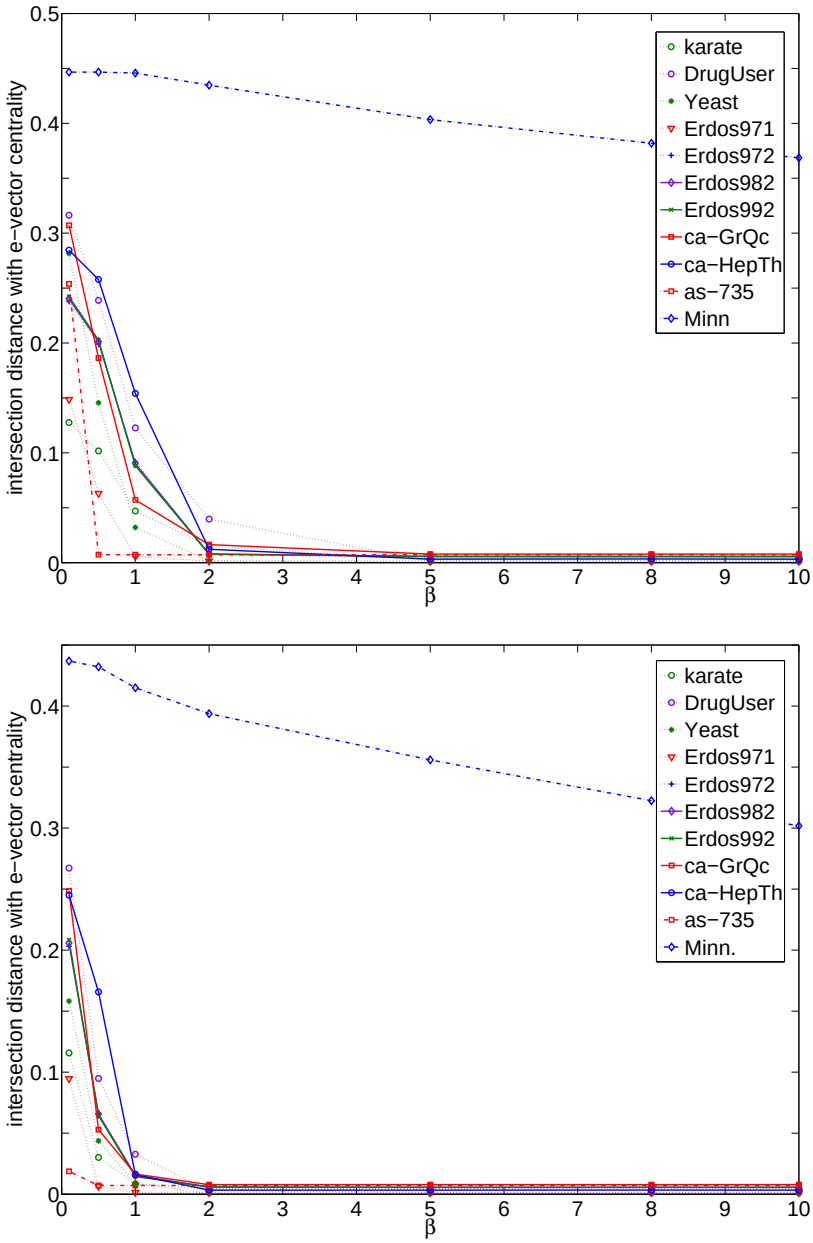

Figure 3.3: The intersection distances between eigenvector centrality and the exponential subgraph centrality (top) or total communicability (bottom) rankings of the nodes in the networks in Table 2.6.

For this network, the intersection distances between the matrix exponential based centrality rankings and the eigenvector centrality rankings still decrease as β increases,

but at a much slower rate than for the other networks. This is also expected, as the spectral gap for the Minnesota road network is extremely small (< 0.001). Again, the rankings of the nodes in the karate, Erdos971, and as-735 networks, which have very large relative spectral gaps, stabilize extremely quickly.

In Figure 3.4, the intersection distances between the rankings produced by exponential subgraph centrality and total communicability are compared for successive choices of β. Overall, these intersection distances are quite low (the highest is 0.25 and occurs for the exponential subgraph centrality rankings of the as-735 network when β increases from 0.1 to 0.5). For all the networks examined, the largest intersection distances between successive choices of β occur as β increases to two. For higher values of β, the intersection distance drops, which corresponds to the fact that the rankings are converging to those produced by eigenvector centrality. In general, there is less change in the rankings produced by the total communicability scores for successive values of β than for the rankings produced by the exponential subgraph centrality scores.

If the intersection distances are restricted to the top 10 nodes, they are even lower. For the karate, Erdos992, and ca-GrQc networks, the intersection distance for the top 10 nodes between successive choices of β is always less than 0.1. For the DrugUser, Yeast, Erdos971, Erdos982, and ca-HepTh networks, the intersection distances are somewhat higher for low values of β, but by the time $\beta = 2$, they are all equal to 0 as the rankings have converged to those produced by the eigenvector centrality. For the Erdos972 network, this occurs slightly more slowly. The intersection distances between the rankings of the top 10 nodes produced by $\beta = 2$ and $\beta = 5$ are 0.033 and for all subsequence choices of β are 0. In the case of the Minnesota Road network, the intersection distances between the top 10 ranked nodes never stabilize to 0, as is expected (see Figure B.1 in Appendix B).

For the networks examined, when $\beta < 0.5$, the exponential subgraph centrality and total communicability rankings are very close to those produced by degree centrality.

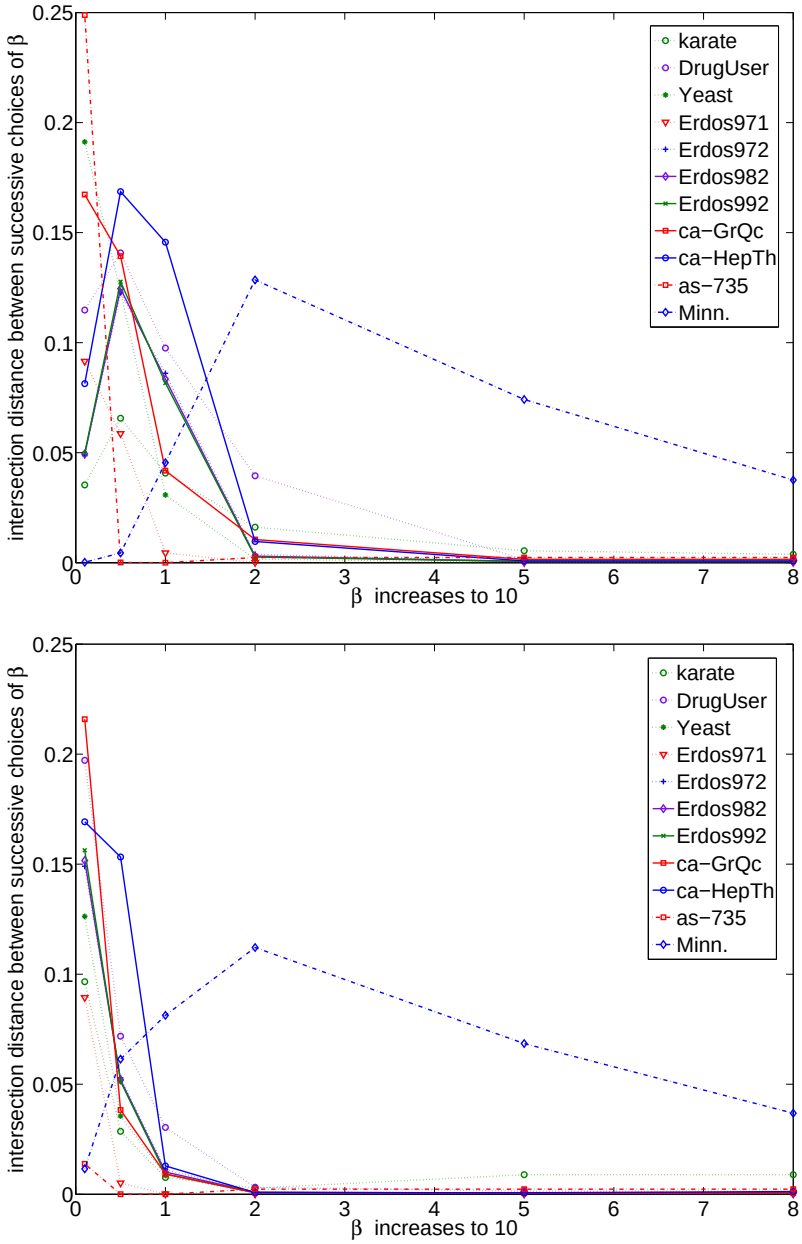

Figure 3.4: The intersection distances between the exponential subgraph centrality (top) or total communicability (bottom) rankings produced by successive choices of β. Each line corresponds to a network in Table 2.6.

When $\beta \geq 2$, they are essentially identical to the rankings produced by eigenvector centrality. Thus, the most additional information about node rankings (i.e. information

that is not contained in the degree or eigenvector centrality rankings) is obtained when $0.5 < \beta < 2$. This supports the intuition developed in Section 3.3 that moderate values of β should be used to gain the most benefit from the use of matrix exponential based centrality rankings.

3.5.2 Resolvent subgraph and Katz centrality

In this section, we investigate the effect of changes in α on the resolvent subgraph centrality and Katz centrality in the networks listed in 2.6, as well as the relationship of these centrality measures to degree and eigenvector centrality. We calculate the scores and node rankings produced by $\mathbf{C_d}$ and $\mathbf{C_{ev}}$, as well as those produced by $\mathbf{RC}(\alpha)$ and $\mathbf{K}(\alpha)$ for various values of α. The values of α tested are given by $\alpha = 0.01 \cdot \frac{1}{\lambda_1}$, $0.05 \cdot \frac{1}{\lambda_1}$, $0.1 \cdot \frac{1}{\lambda_1}$, $0.25 \cdot \frac{1}{\lambda_1}$, $0.5 \cdot \frac{1}{\lambda_1}$, $0.75 \cdot \frac{1}{\lambda_1}$, $0.9 \cdot \frac{1}{\lambda_1}$, $0.95 \cdot \frac{1}{\lambda_1}$, and $0.99 \cdot \frac{1}{\lambda_1}$.

As in Section 3.5.1, the rankings produced by degree centrality and eigenvector centrality were compared to those produced by matrix resolvent based centrality measures for all choices of α using the intersection distance method. The results are plotted in Figures 3.5 and 3.6. The rankings produced by successive choices of α are also compared and these intersection distances are plotted in Figure 3.7.

Figure 3.5 shows the intersection distances between the degree centrality rankings and those produced by resolvent subgraph centrality or Katz centrality for the values of α tested. When α is small, the intersection distances between the matrix resolvent based centrality rankings and the degree centrality rankings are low. For $\alpha = 0.01 \cdot \frac{1}{\lambda_1}$, the largest intersection distance between the degree centrality rankings and the resolvent subgraph centrality rankings is slightly less than 0.2 (for the Minnesota road network). The largest intersection distance between the degree centrality rankings and the Katz centrality rankings is also slightly less than 0.2 (again, for the Minnesota road network). The relatively large intersection distances for the node rankings on the Minnesota road network when $\alpha = 0.01 \cdot \frac{1}{\lambda_1}$ is due to the fact that both the degree

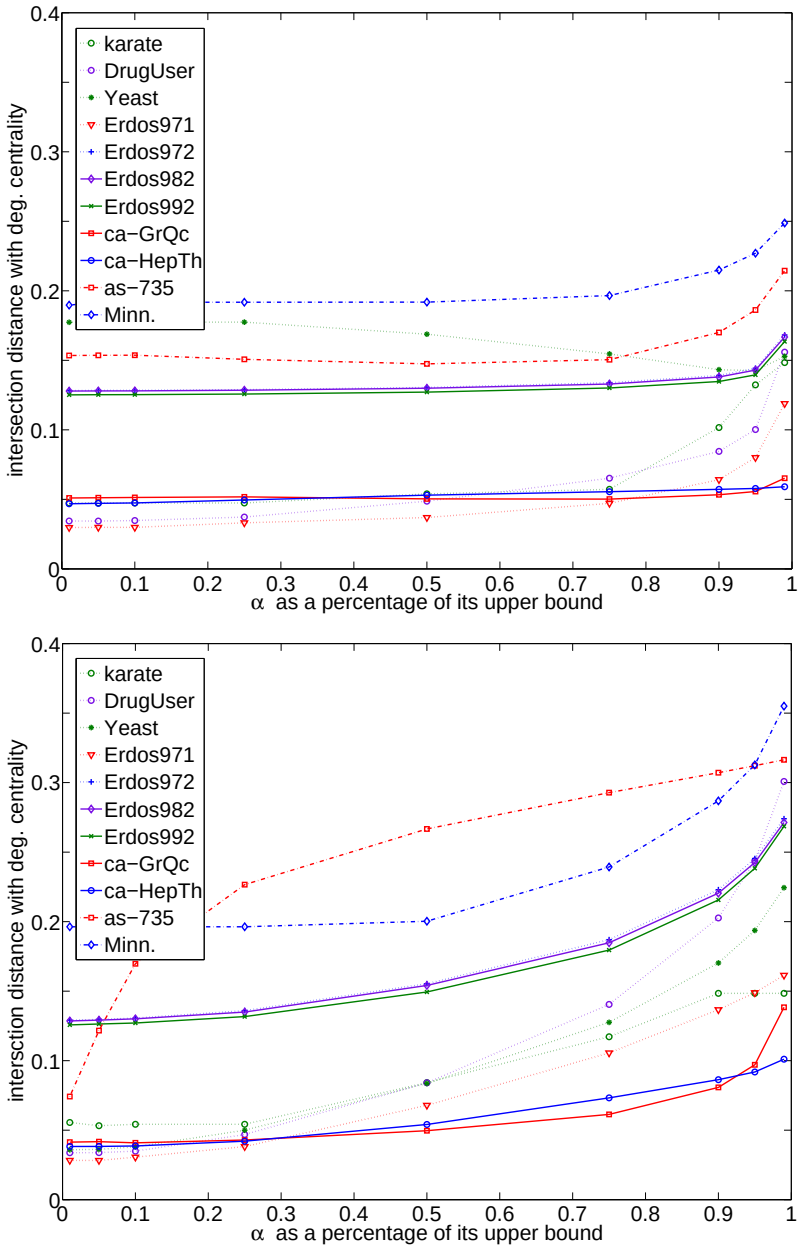

Figure 3.5: The intersection distances between degree centrality and the resolvent subgraph centrality (top) or Katz centrality (bottom) rankings of the nodes in the networks in Table 2.6.

centrality and the resolvent subgraph (or Katz) centrality scores for the nodes are very close. Thus, small changes in the score values can lead to large changes in the rank-

ings. As α increases towards $\frac{1}{\lambda_1}$, the intersection distances increase. This increase is more rapid for the Katz centrality rankings than for the resolvent subgraph centrality rankings.

In Figure 3.6, the resolvent subgraph centrality and Katz centrality rankings for various values of α are compared to the eigenvector centrality rankings on the networks described in Table 2.6. For small values of α, the intersection distances tend to be large. As α increases, the intersection distances decrease for both resolvent subgraph centrality and Katz centrality on all of the networks examined. This decrease is faster for the (row sum based) Katz centrality rankings than for the (diagonal based) resolvent subgraph centrality rankings. The network with the highest intersection distances between the eigenvector centrality rankings and those based on the matrix resolvent, and slowest decrease of these intersection distances as α increases, is the Minnesota road network. As was the case when matrix exponential based scores were examined, this is expected due to this network's small spectral gap. The node rankings in networks with large relative spectral gaps (karate, Erdos971, as-735) converge to the eigenvector centrality rankings most quickly.

The intersection distance between the resolvent subgraph and Katz centrality rankings produced by successive choices of α are plotted in Figure 3.7. All of these intersection distances are extremely small (the largest is < 0.08), indicating that the rankings do not change much as α increases. However, as α increases, the rankings corresponding to successive values of α tend to be slightly less similar to each other. The exception to this is the Katz centrality rankings for the as-735 network which become more similar as α increases.

Again, if the analysis is restricted to the top 10 nodes, the intersection distances between the rankings produced by successive choices of α are very small. For the karate, Erdos971, Erdos982, Erdos992, ca-GrQc, and Minnesota road networks, the intersection distances between the top 10 ranked nodes for successive choices of α are

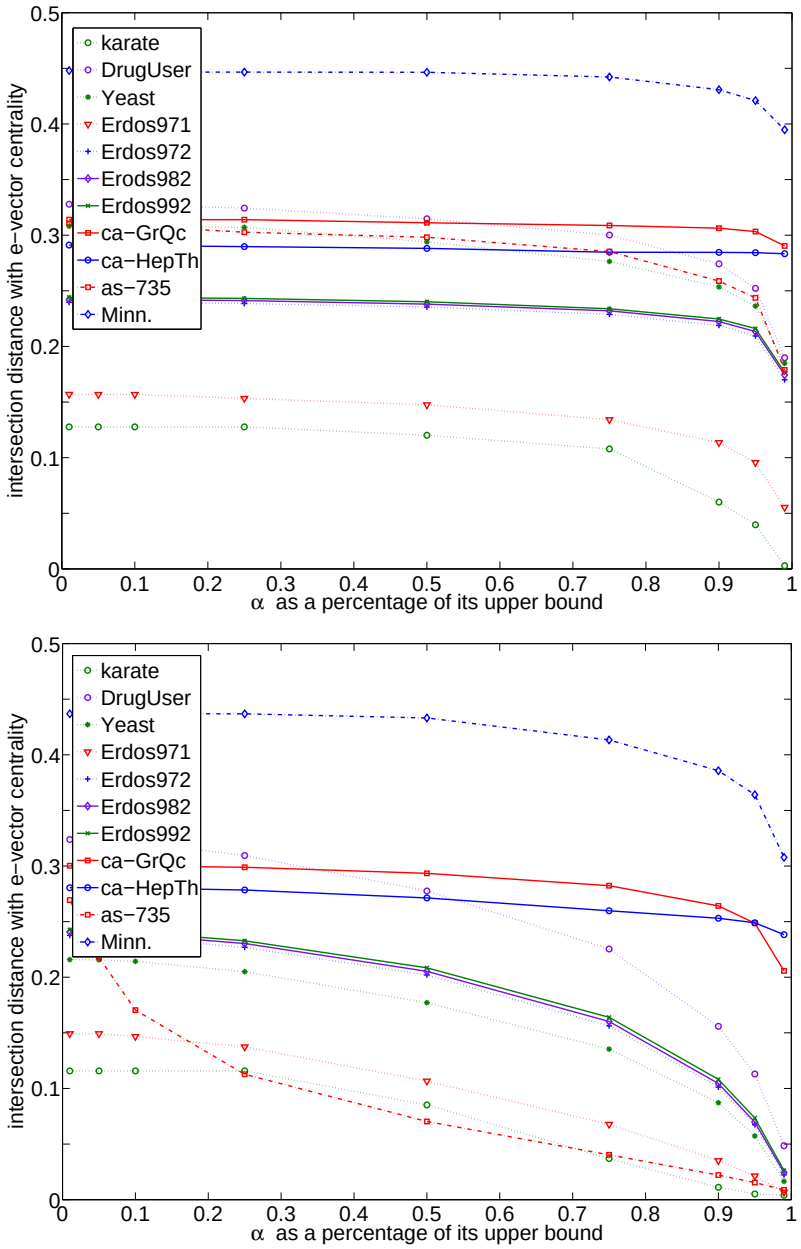

Figure 3.6: The intersection distances between eigenvector centrality and the resolvent subgraph centrality (top) or Katz centrality (bottom) rankings of the nodes in the networks in Table 2.6.

always less than or equal to 0.1 and often equal to zero. For the ca-HepTh network, the top 10 ranked nodes are exactly the same for all choices of α. For the DrugUser, Yeast,

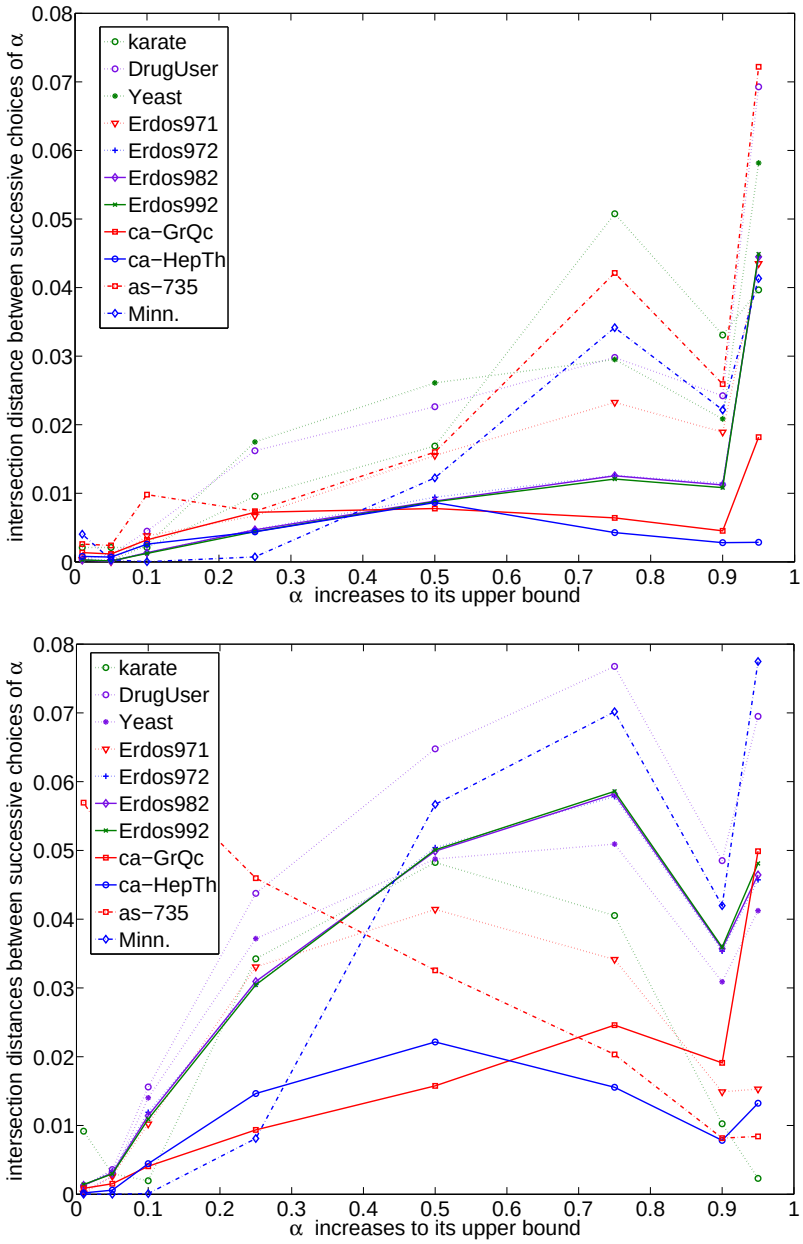

Figure 3.7: The intersection distances between resolvent subgraph centrality (top) or Katz centrality (bottom) rankings produced by successive choices of α. Each line corresponds to a network in Table 2.6.

and Erdos972 networks, they are always less than 0.2 (see Figure B.2 in Appendix B).

For the eleven networks examined, the resolvent subgraph and Katz centrality rank-

ings tend to be close to the degree centrality rankings when $\alpha < 0.5 \cdot \frac{1}{\lambda_1}$. It is interesting to note that as α increases, these rankings stay close to the degree centrality rankings until α is approximately one half of its upper bound. Additionally, the matrix resolvent based rankings are close to the eigenvector centrality rankings when $\alpha > 0.9 \cdot \frac{1}{\lambda_1}$. Thus, the most information is gained by using matrix resolvent based centrality measures when $0.5 \cdot \frac{1}{\lambda 1} \leq \alpha \leq 0.9 \cdot \frac{1}{\lambda_1}$. This supports the intuition from Section 3.3 that moderate values of α provide the most additional information about node ranking beyond that provided by degree and eigenvector centrality.

3.6 Centrality robustness in directed networks

In Section 3.2, we examined the relationship between the rankings produced by centrality measures based on the matrix exponential and resolvent with those produced by degree and eigenvector centrality on undirected networks. However, the measurement of node "importance" becomes more complicated on directed networks. In addition to the difficulties associated with ranking nodes in undirected networks (which centrality measure to use, how various centrality measures are related, etc.), the fact that A is no longer symmetric means that all adjacency matrix based centrality measures can be applied to either A or A^T. In terms of degree centrality, nodes now have both in- and out-degrees and, in terms of eigenvector centrality, A now has both a dominant left and a dominant right eigenvector.

The application of centrality measures to A and A^T correspond to two different types of node importance in directed networks. Since edges can only be traversed in one direction, in terms of information flow there is a difference between the ability of a node to spread information and its ability to gather information. In [8, 68], these different abilities were captured by assigning each node a hub and authority score. In [39], the two aspects of information spread are referred to as broadcast and receive

centrality. Broadcast centralities measure the ability of nodes in a network to broadcast information along directed walks. Here, we will examine broadcast centralities based on the row sums of $f(A)$ where f is the matrix exponential or resolvent. Receive centralities measure the ability of nodes in a network to receive information along directed walks. We will examine receive centralities based on the column sums of $f(A)$ (which correspond to the row sums of $f(A^T)$), where f is, again, the matrix exponential or resolvent.

We do not consider broadcast and receive centralities based on the diagonal entries of matrix functions. This is due to the fact that the diagonal entries of $f(A)$ and $f(A^T)$ are the same. Thus, these centralities measures cannot distinguish between the two types of node "importance" in a directed network.

The relationship between the broadcast and receive total communicabilities with the in- and out-degrees and left and right eigenvectors of A are described in the following theorem:

Theorem 3.6.1 Let $G = (V, E)$ be a strongly connected, directed network with adjacency matrix A. Let $TC_i^b(\beta) = [e^{\beta A}\mathbf{1}]_i$ be the broadcast total communicability of node i and $\mathbf{TC}^b(\beta)$ be the vector of broadcast total communicabilities. Let $TC_i^r(\beta) = [e^{\beta A^T}\mathbf{1}]_i$ be the receive total communicability of node i and $\mathbf{TC}^r(\beta)$ be the vector of receive total communicabilities. Then,

(i) as $\beta \to 0+$, the rankings produced by $\mathbf{TC}^b(\beta)$ converge to those produced by the out-degrees of the nodes in the network,

(ii) as $\beta \to 0+$, the rankings produced by $\mathbf{TC}^r(\beta)$ converge to those produced by the in-degrees of the nodes in the network,

(iii) as $\beta \to \infty$, the rankings produced by $\mathbf{TC}^b(\beta)$ converge to those produced by $\mathbf{x}_1$, where $\mathbf{x}_1$ is the dominant right eigenvector.

(iv) as $\beta \to \infty$, the rankings produced by $\mathbf{TC}^r(\beta)$ converge to those produced by $\mathbf{y}_1$, where $\mathbf{y}_1$ is the dominant left eigenvector.

Proof: The proofs of (i) and (ii) follow the same procedure as the proof of Theorem 3.2.1 applied to A and A^T, respectively.

If A is diagonalizable, the proof of (iii) can be seen by considering the expansion of $TC_i^b(\beta)$ in terms of X and Λ:

$$TC_i^b(\beta) = [e^{\beta A}\mathbf{1}]_i = [Xe^{\beta\Lambda}X^{-1}\mathbf{1}]_i = \sum_{k=1}^{n} e^{\beta\lambda_k}\mathbf{x}_k(i)(\mathbf{y}_k^*\mathbf{1})$$

$$= e^{\beta\lambda_1}\mathbf{x}_1(i)(\mathbf{y}_1^T\mathbf{1}) + \sum_{k=2}^{n} e^{\beta\lambda_k}\mathbf{x}_k(i)(\mathbf{y}_k^*\mathbf{1}).$$

Let $\boldsymbol{\psi}^b(\beta) = \frac{1}{e^{\beta\lambda_1}(\mathbf{y}_1^T\mathbf{1})}\mathbf{TC}^b(\beta)$. Similarly to the proof of (i), the rankings produced by $\boldsymbol{\psi}^b(\beta)$ are equivalent to those produced by $\mathbf{TC}^b(\beta)$, since the scores for each node have been scaled in the same way. Now,

$$\psi_i^b(\beta) = \mathbf{x}_1(i) + \sum_{k=2}^{n} \frac{e^{\beta(\lambda_k-\lambda_1)}}{\mathbf{y}_1^T\mathbf{1}}\mathbf{x}_k(i)(\mathbf{y}_k^*\mathbf{1}).$$

As $\beta \to \infty$, $\psi_i^b(\beta) \to \mathbf{x}_1(i)$. Note that by the Perron-Frobenius Theorem, $\mathbf{x}_1 > 0$, so the rankings produced by $\mathbf{x}_1(i)$ are all real and positive.

If A is not diagonalizable, then from 1.3, we know:

$$TC_i(\beta) = \sum_{k=1}^{s}\sum_{j=0}^{l_k-1} \frac{\beta^j e^{\beta\lambda_k}}{j!}[(A-\lambda_k I)^j G_k\mathbf{1}]_i$$

where s is the number of distinct eigenvalues of A, l_k is the geometric multiplicity of the kth distinct eigenvalue, and G_k is the oblique projector onto $\mathcal{N}((A-\lambda_k I)^{l_k})$ along $\mathcal{R}((A-\lambda_k I)^{l_k})$. Due to the fact that λ_1 is simple by the Perron-Frobenius theorem, this

becomes:

$$TC_i(\beta) = e^{\beta\lambda_1}\mathbf{x}_1(i)(\mathbf{y}_1^T\mathbf{1}) + \sum_{k=2}^{s}\sum_{j=0}^{l_k-1}\frac{\beta^j e^{\beta\lambda_k}}{j!}[(A-\lambda_k I)^j G_k\mathbf{1}]_i.$$

Again, let $\boldsymbol{\psi}^b(\beta) = \frac{1}{e^{\beta\lambda_1(\mathbf{y}_1^T\mathbf{1})}}\mathbf{TC}^b(\beta)$. The rankings produced by $\boldsymbol{\psi}^b(\beta)$ will be the same as those produced by $\mathbf{TC}^b(\beta)$. Now,

$$\psi_i^b(\beta) = \mathbf{x}_1(i) + \sum_{k=2}^{s}\sum_{j=0}^{l_k-1}\frac{\beta^j e^{\beta(\lambda_k-\lambda_1)}}{j!(\mathbf{y}_1^T\mathbf{1})}[(A-\lambda_k I)^j G_k\mathbf{1}]_i$$

which converges to $\mathbf{x}_1(i)$ as $\beta \to \infty$. □

A similar theorem holds for the broadcast and receive Katz centralities.

3.7 Numerical experiments on directed networks

In this section, we examine the relationship between the matrix exponential and resolvent based broadcast centrality measures with the out-degrees and the dominant right eigenvectors of two real world directed networks. A similar analysis can be done on the relationship between the receive centrality measures and the in-degrees and dominant left eigenvectors. Both networks can be found in the University of Florida Sparse Matrix Collection [31]. As was done in Section 3.5, the rankings are compared using the intersection distance method. The first network we examine is wb-cs-Stanford, a network of hyperlinks between the Stanford CS webpages in 2001. It is in the Gleich group of the UF collection. The second network is the wiki-Vote network, which is a network of who votes for whom in elections for Wikipedia editors to become administrators. It is in the SNAP group of the UF collection.

Since the theorems in Section 3.6 only hold for strongly connected networks with irreducible adjacency matrices, our experiments were performed on the largest strongly connected component of the above networks. Basic data on these strongly connected

Table 3.1: Basic data on the largest strongly connected component of the real-world directed networks examined.

Graph	n	nnz	λ_1	λ_2
Gleich/wb-cs-Stanford	2759	13895	35.618	12.201
SNAP/wiki-Vote	1300	39456	45.145	27.573

components can be found in Table 3.1. In both of the networks examined, the two largest eigenvalues of the largest strongly connected component were real. However, this is not always the case. Both networks are simple.

3.7.1 Total communicability

As in Section 3.5.1, we examined the effects of changing β on the broadcast total communicability rankings of nodes in two directed real world networks, as well as their relation to the out-degrees and dominant right eigenvectors of the networks. The measures were calculated for the networks described in Table 3.1. To examine the sensitivity of the broadcast total communicability rankings, we calculate the scores and rankings for various choices of β. The values of β tested are: 0.1, 0.5, 1, 2, 5, 8 and 10.

The broadcast rankings produced by total communicability for all choices of β were compared to those produced by the out-degree rankings and the rankings produced by $\mathbf{x}_1$ using the intersection distance method as described in Section 3.5. Plots of the intersection distances for the rankings produced by various choices of β with those produced by the out-degrees and right dominant eigenvector can be found in Figures 3.8 and 3.9. The intersection distances for rankings produced by successive choices of β can be found in Figure 3.10.

In Figure 3.8, the intersection distances between the rankings produced by broadcast total communicability are compared to those produced by the out-degrees of nodes in the network. As β approaches 0, the intersection distances decrease for both networks. As β increases to 10, the intersection distances initially increase, then stabilize

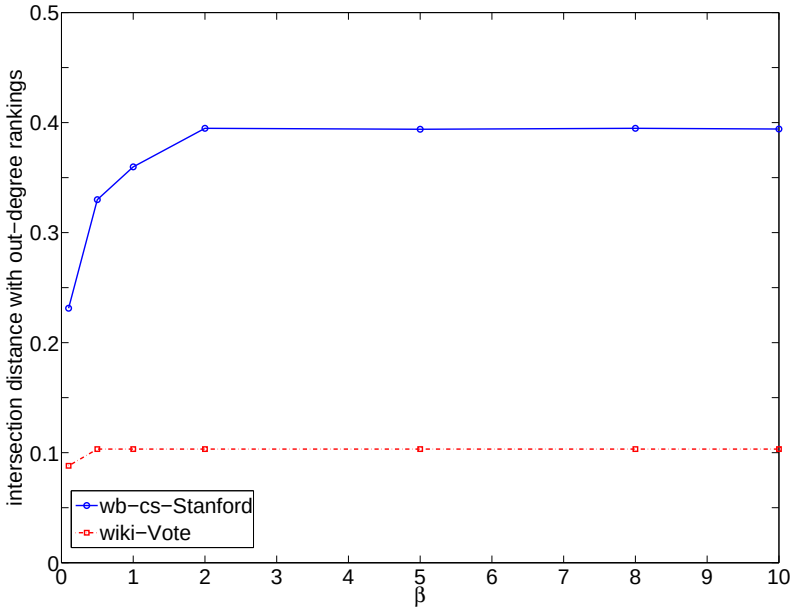

Figure 3.8: The intersection distances between the out-degree rankings and the broadcast total communicability rankings of the nodes in the networks in Table 3.1.

as the rankings converge to those produced by $\mathbf{x}_1$.

The intersection distances between the rankings produced by broadcast total communicability are compared to those produced by $\mathbf{x}_1$ in Figure 3.9. For both networks, the intersection distances quickly decrease as β increases. In the wiki-Vote network, the intersection distances between the compared rankings are 0 by the time $\beta = 0.5$. For the wb-cs-Stanford network, by the time β has reached five, the intersection distance between the broadcast total communicability rankings and those produced by $\mathbf{x}_1(i)$ have decreased to about 0.04. The rankings then stabilize at this intersection distance. This is due to a group of nodes that have nearly identical total communicability scores.

In Figure 3.10, the intersection distances between the broadcast total communicability rankings for successive choices of β are plotted. These intersection distances are slightly lower than those observed in the undirected case, with a maximum of approximately 0.14, which occurs in the wb-cs-Stanford network when β increases from 0.01 to 0.05. By the time $\beta = 0.5$, the of rankings on the wiki-Vote network have stabilized

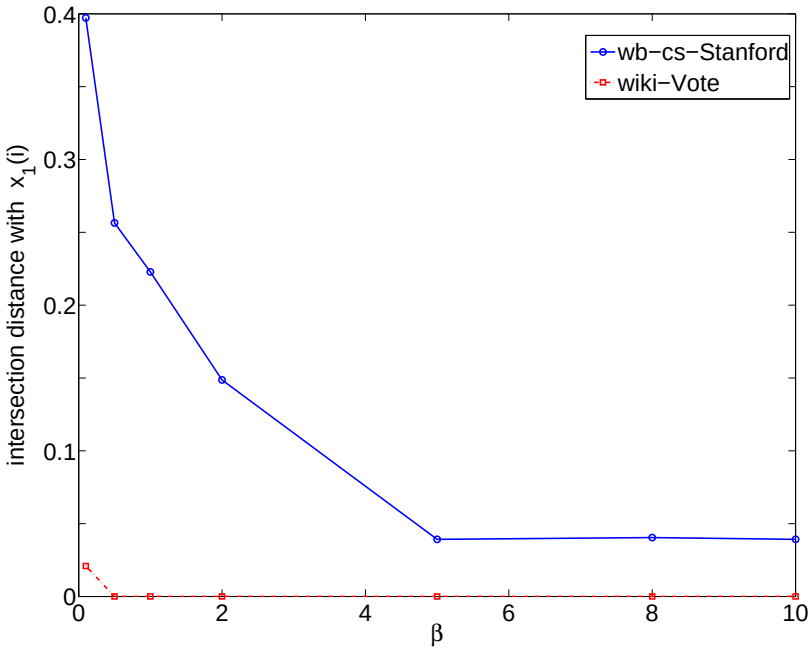

Figure 3.9: The intersection distances between the rankings produced by $\mathbf{x}_1$ and the broadcast total communicability rankings of the nodes in the networks in Table 3.1.

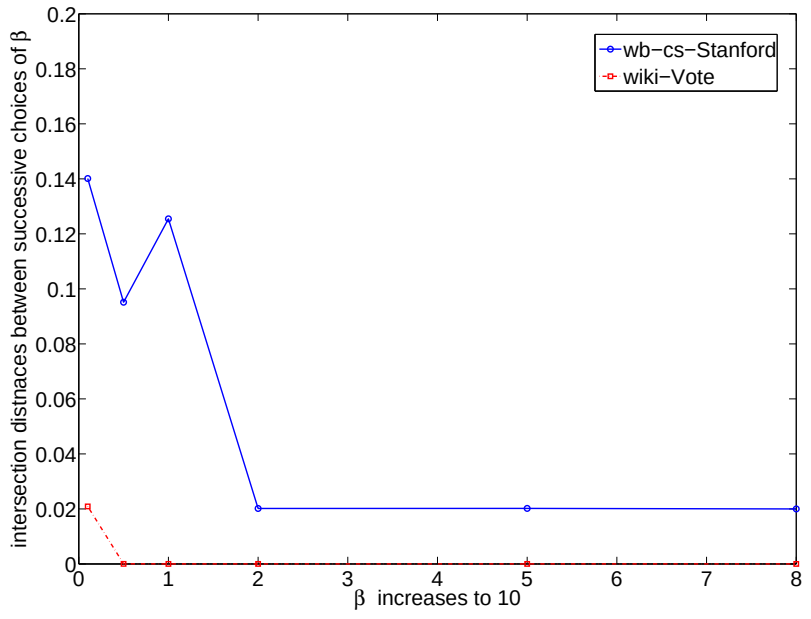

Figure 3.10: The intersection distances between the broadcast total communicability rankings produced by successive choices of β. Each line corresponds to a network in Table 3.1.

and all subsequent intersection distances are 0. For both the broadcast total communicability rankings on the wb-cs-Stanford network, the intersection distances decrease (non-monotonically) as β increases until they stabilize at approximately 0.02.

When this analysis is restricted to the top 10 nodes, the intersection distances are extremely small. For the wb-cs-Stanford network, the largest intersection distance between the top 10 ranked nodes for successive choices of β is 0.11 (when β increases from 0.1 to 0.5). For the wiki-Vote network, the intersection distance between the top 10 total communicability scores is 0.01 when β increases from 0.1 to 0.5 and zero otherwise (see Figure B.3 in Appendix B).

The differences between the out-degree rankings and the broadcast total communicability rankings are greatest when $\beta \geq 0.5$. The differences between the left and right eigenvector based rankings and the broadcast rankings are greatest when $\beta < 2$ (although in the case of the wiki-Vote network, they have converged by the time $\beta = 0.5$). Thus, like in the case of the undirected networks, moderate values of β give the most additional ranking information beyond that provided by the out-degrees and the left and right eigenvalues.

3.7.2 Katz centrality

In this section, we investigate the effect of changes in α on the broadcast Katz centrality rankings of nodes in the networks listed in Table 3.1 and relationship of these centrality measures to the rankings produced by the out-degrees and the dominant right eigenvectors of the network. We calculate the scores and node rankings produced by $\mathbf{K}^b(\alpha)$ for various values of α. The values of α tested are given by $\alpha = 0.01 \cdot \frac{1}{\lambda_1}$, $0.05 \cdot \frac{1}{\lambda_1}$, $0.1 \cdot \frac{1}{\lambda_1}$, $0.25 \cdot \frac{1}{\lambda_1}$, $0.5 \cdot \frac{1}{\lambda_1}$, $0.75 \cdot \frac{1}{\lambda_1}$, $0.9 \cdot \frac{1}{\lambda_1}$, $0.95 \cdot \frac{1}{\lambda_1}$, and $0.99 \cdot \frac{1}{\lambda_1}$.

The rankings produced by the out-degrees and the dominant right eigenvectors were compared to those produced by Katz centrality for all choices of α using the intersection distance method, as was done in Section 3.7.1. The results are plotted in

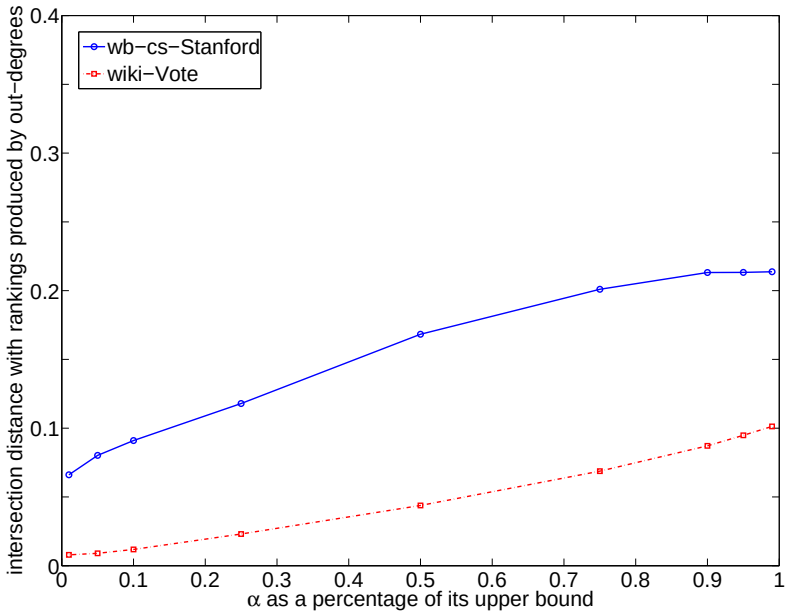

Figure 3.11: The intersection distances between the rankings produced by the out-degrees and those produced by broadcast Katz centrality rankings of the nodes in the networks in Table 3.1.

Figures 3.11 and 3.12.

As α increases from $0.01 \cdot \frac{1}{\lambda_1}$ to $0.99 \cdot \frac{1}{\lambda_1}$, the intersection distances between the scores produced by the broadcast Katz centralities and the out-degrees increase. When α is small, the broadcast Katz centrality rankings are very close to those produced by the out-degrees (low intersection distances). On the wb-cs-Stanford network, when $\alpha = 0.01 \cdot \frac{1}{\lambda_1}$, the intersection distance between the two rankings is approximately 0.06. On the wiki-Vote network, it is approximately 0.01. As α increases, the intersection distances also increase. By the time $\alpha = 0.99 \cdot \frac{1}{\lambda_1}$, the intersection distance between the two sets of node rankings on the wb-cs-Stanford network is above 0.2 and on the wiki-Vote network it is approximately 0.1.

In Figure 3.12, the rankings produced by broadcast Katz centrality are compared to those produced by $\mathbf{x}_1$. Overall, The intersection distances between the two sets of rankings are lower on the wiki-Vote network than they are on the wb-cs-Stanford

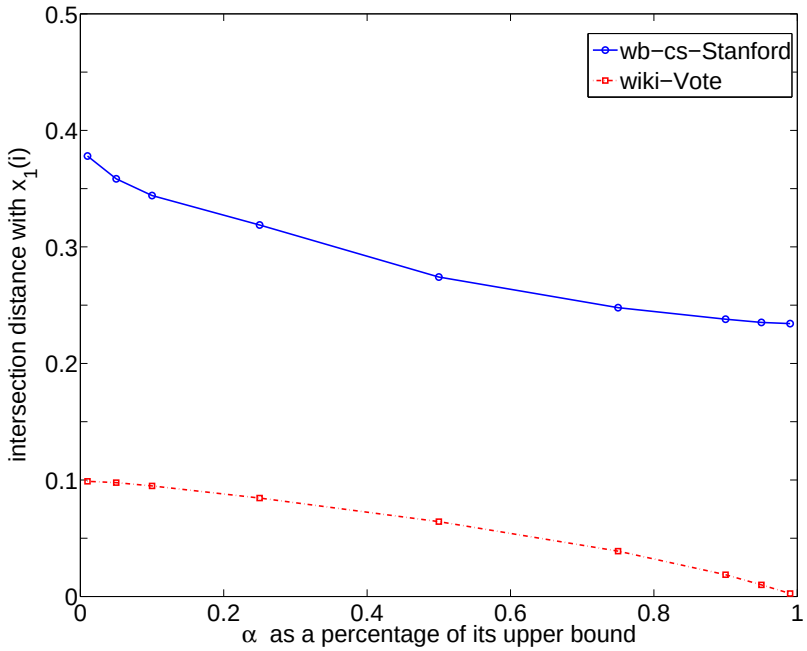

Figure 3.12: The intersection distances between the rankings produced by $\mathbf{x}_1$ and the broadcast Katz centrality (right) rankings of the nodes in the networks in Table 3.1.

network. As α increases from $0.01 \cdot \frac{1}{\lambda_1}$ to $0.99 \cdot \frac{1}{\lambda_1}$, the intersection distances between the two sets of rankings on the wiki-Vote network decrease from 0.1 to essentially 0. On the wb-cs-Stanford network, they decrease from approximately 0.47 to 0.24.

The intersection distances between the rankings produced by the broadcast Katz centralities for successive values of α are plotted in Figure 3.13. As was the case in the undirected networks examined, these rankings are more stable in regards to the choice of α than the total communicability rankings were in regards to the choice of β. Here, the maximum intersection distance is less than 0.1. When only the top 10 ranked nodes are similar, the intersection distances have a maximum of 0.06 (on the wb-cs-Stanfrod network when α increases from $0.25 \cdot \frac{1}{\lambda_1}$ to $0.5 \cdot \frac{1}{\lambda_1}$). For both networks,the intersection distances between the rankings on the top 10 nodes for successive choices of α are quite small (see Figure B.4 in Appendix B).

The broadcast Katz centrality rankings are only far from those produced by the out-

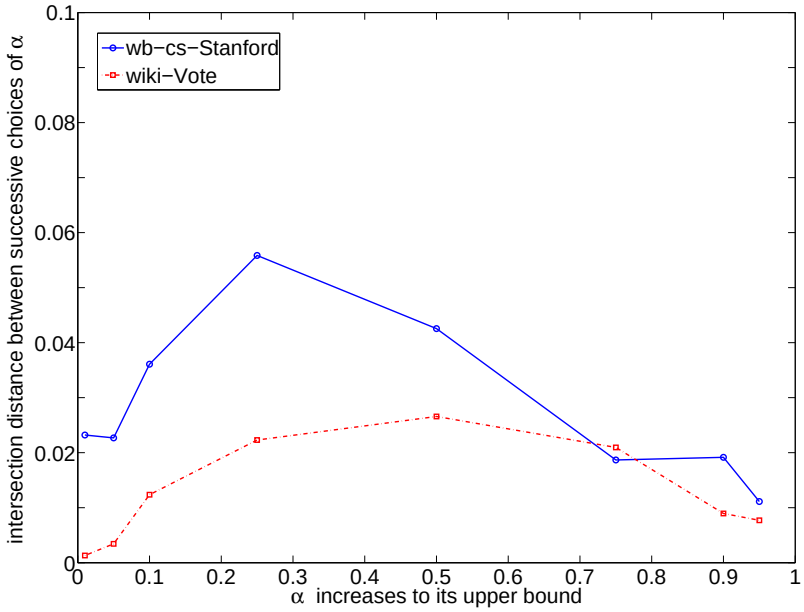

Figure 3.13: The intersection distances between the broadcast Katz centrality rankings produced by successive choices of α. Each line corresponds to a network in Table 3.1.

degrees when $\alpha \geq 0.5 \cdot \frac{1}{\lambda_1}$. They are farthest from those produced by the dominant right eigenvector of A when $\alpha < 0.9 \cdot \frac{1}{\lambda_1}$. Thus, as was seen in the case of undirected networks, the most additional information is gained when moderate values of α, $0.5 \cdot \frac{1}{\lambda_1} \leq \alpha < 0.9 \cdot \frac{1}{\lambda_1}$, are used to calculate the matrix resolvent based centrality scores.

3.8 Summary and conclusions

We have analyzed the relationship between centrality measures based on the diagonal entries and row sums of the matrix exponential and resolvent with degree and eigenvector centrality. We have shown that the parameters α (in the case of matrix resolvent-based centrality rankings) and β (in the case of matrix exponential-based centrality ranks) act as tunable parameters between the degree centrality rankings and eigenvector centrality rankings. That is, when α and β tend toward their lower bounds, the rankings produced by exponential subgraph centrality, total communicability, re-

solvent subgraph centrality, and Katz centrality converge to those produced by degree centrality. Similarly, as α and β tend to their upper bounds, these rankings converge to those produced by eigenvector centrality. We also demonstrated a similar relationship between the broadcast and receive centralities (based on the row and column sums of the matrix exponential and resolvent) and the rankings produced by the in- and out-degrees and the left and right eigenvectors. This relationship helps to explain the observed correlation between the degree and eigenvector centrality rankings on many real-world complex networks (which often have a large spectral gap).

Additionally, we have presented the results from experiments that used a large variety of choices of α and β to compute rankings on a number of directed and undirected real world networks. We computed the intersection distances of these rankings with those produced by degree and eigenvector centrality and with each other. Here, we found that, as expected, the rankings were close to degree centrality when α and β were close to their lower bounds and were close to eigenvector centrality when α and β were close to their upper bounds. The rankings were the least similar to both degree and eigenvector centrality when $0.5 \cdot \frac{1}{\lambda 1} \leq \alpha \leq 0.9 \cdot \frac{1}{\lambda_1}$ and $0.5 < \beta < 2$.

Our results also allow us to provide guidelines for the choice of the parameters α and β in order to produce rankings that are the most different from the degree and eigenvector centrality rankings and, therefore, most useful in terms of adding more information to the analysis of a complex network. Although the question of the optimal choice of parameter is not, and may never be, completely resolved, the results presented in this chapter contribute to the ability to make intelligent parameter choices.

4 Ranking Hubs and Authorities using Matrix Functions

4.1 Introduction

As we have seen previously, subgraph centrality is one of the most commonly used centrality measures. The interpretation of centrality described in [47] applies mostly to undirected networks. However, many important real-world networks (the World Wide Web, the Internet, citation networks, food webs, certain social networks, etc.) are directed. One goal of this chapter is to extend the notions of centrality and communicability described in [45, 47] to directed networks, with an eye towards developing new ranking algorithms for, e.g., document collections, web pages, and so forth. We further compare our approach with some standard algorithms, such as HITS (see [68]) and a few others. Methods of quickly determining hub and authority rankings using Gauss-type quadrature rules are also discussed.

4.2 HITS reformulation

One of the classical methods for ranking nodes in a directed network is the Hypertext Induced Topics Search (HITS) algorithm, first introduced by J. Kleinberg in [68], which ranks nodes as hubs and authorities. This algorithm provides the motivation for the extension of subgraph centrality to directed graphs given in section 4.4. The standard presentation of the HITS algorithm can be found in Chapter 1, Section 1.7.

In a digraph the adjacency matrix A is generally nonsymmetric, however, the two matrices used in the HITS algorithm (A^TA and AA^T) are symmetric. Note that, setting

$$\mathcal{A} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix},$$

a symmetric matrix is obtained. Now,

$$\mathcal{A}^2 = \begin{pmatrix} AA^T & 0 \\ 0 & A^TA \end{pmatrix}; \quad \mathcal{A}^3 = \begin{pmatrix} 0 & AA^TA \\ A^TAA^T & 0 \end{pmatrix}.$$

In general,

$$\mathcal{A}^{2k} = \begin{pmatrix} (AA^T)^k & 0 \\ 0 & (A^TA)^k \end{pmatrix}; \quad \mathcal{A}^{2k+1} = \begin{pmatrix} 0 & A(A^TA)^k \\ (A^TA)^kA^T & 0 \end{pmatrix}.$$

Applying HITS to this matrix $\mathcal{A}$, $\mathcal{A}^T = \mathcal{A}$ so $\mathcal{A}^T\mathcal{A} = \mathcal{A}\mathcal{A}^T = \mathcal{A}^2$ and introducing the vector $\mathbf{u}^{(k)} = \begin{pmatrix} \mathbf{y}^{(k)} \\ \mathbf{x}^{(k)} \end{pmatrix}$ for $k = 1, 2, 3, \ldots$, equation (3.2) becomes

$$\mathbf{u}^{(k)} = \mathcal{A}^2\mathbf{u}^{(k-1)} = \begin{pmatrix} AA^T & 0 \\ 0 & A^TA \end{pmatrix} \mathbf{u}^{(k-1)}, \tag{4.1}$$

followed by normalization of the two vector components of $\mathbf{u}^{(k)}$ so that each has 2-norm equal to 1. Now, if A is an $n \times n$ matrix, $\mathcal{A}$ is $2n \times 2n$ and vector $\mathbf{u}^{(k)}$ is in $\mathbb{R}^{2n}$. The first n entries of $\mathbf{u}^{(k)}$ correspond to the hub rankings of the nodes, while the last n entries give the authority rankings. Under suitable assumptions (see the discussion in [73, Chapter 11.3]), as $k \to \infty$ the sequence $\{\mathbf{u}^{(k)}\}$ converges to the dominant nonnegative eigenvector of $\mathcal{A}$, which yields the desired hub and authority rankings.

Hence, in HITS only information obtained from the dominant eigenvector of $\mathcal{A}$ is

used. It is natural to expect that taking into account spectral information corresponding to the remaining eigenvalues and eigenvectors of $\mathcal{A}$ may lead to improved results.

Among the limitations of HITS, we mention the possible dependence of the rankings on the choice of the initial vectors $\mathbf{x}^{(0)}$, $\mathbf{y}^{(0)}$, see [52] for examples of this.

4.3 Subgraph centralities and communicabilities

In [47], the authors review the use of (exponential) subgraph centrality and communicability for ranking nodes in undirected networks. However, these measures are less meaningful when applied to directed networks. Although the matrix exponential is certainly well-defined for any matrix, whether symmetric or not, the *interpretation* of the notion of subgraph centrality for directed networks can be problematic. To see this, consider the directed path graph consisting of n nodes, with edge set $E = \{(1,2),(2,3),\ldots,(n-1,n)\}$ and adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{4.2}$$

The entries of e^A are given by

$$[e^A]_{ij} = \begin{cases} 1/(j-i)!, & \text{if } j \geq i, \\ 0, & \text{else.} \end{cases}$$

In particular, the diagonal entries of e^A are all equal to 1. Therefore, it is impossible to distinguish any of the nodes from the others on the basis of this centrality measure; yet, it is clear that the first and last node are rather special, and certainly more "peripheral"

(less “central”) than the other nodes. Part of the problem, of course, is that the path digraph contains no closed walks. In the next section we show one way to extend the notion of subgraph centrality to digraphs that is immune from such shortcomings, and correctly differentiates between nodes in the example above. (On the other hand, it is interesting to note that the interpretation of the off-diagonal entries of e^A in terms of communicabilities is straightforward for the directed path. All entries of e^A below the main diagonal are zero, reflecting the fact that information can only flow from a node to higher-numbered nodes. Also, the entries of e^A decay rapidly away from the main diagonal, reflecting the fact that the “ease" of communication between a node and a higher numbered one decreases rapidly with the distance.) Although the diagonal entires of e^A do not provide a meaningful ranking of the nodes in the network, here the row and column sums are able to identify nodes as hubs and authorities.

Another issue when extending the notions of subgraph centrality and communicability to directed graphs is that computational difficulties may arise. While the computations involved do not pose a problem for small networks, many real-world networks are large enough that directly computing the exponential of the adjacency matrix is prohibitive. In [7], techniques for bounding and estimating individual entries of the matrix exponential using Gaussian quadrature rules are discussed; see also [18] and section 4.8 below. The ability to find upper and lower bounds for the entries requires that the matrix be symmetric, thus these bounds cannot be directly computed using the adjacency matrix of a directed network. Again, these difficulties can be circumvented using the approach discussed in the next section.

4.4 An extension to digraphs

Although the techniques described in [7] cannot be directly applied to non-symmetric matrices, setting

$$\mathcal{A} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \tag{4.3}$$

produces a symmetric matrix $\mathcal{A}$ and, thus, upper and lower bounds of individual entries of $e^{\mathcal{A}}$ can be computed. In Proposition 4.4.1 below we relate $e^{\mathcal{A}}$ to the underlying hub and authority structure of the original digraph. By $B^{\dagger}$ we denote the Moore–Penrose generalized inverse of matrix B.

Proposition 4.4.1 Let $\mathcal{A}$ be as described in equation (4.3). Then,

$$e^{\mathcal{A}} = \begin{pmatrix} \cosh\left(\sqrt{AA^T}\right) & A\left(\sqrt{A^TA}\right)^{\dagger}\sinh\left(\sqrt{A^TA}\right) \\ \sinh\left(\sqrt{A^TA}\right)\left(\sqrt{A^TA}\right)^{\dagger}A^T & \cosh\left(\sqrt{A^TA}\right) \end{pmatrix}.$$

Proof: Let $A = U\Sigma V^T$ be the SVD of the original (non-symmetric) adjacency matrix A. Then, $\mathcal{A}$ can be decomposed as $\mathcal{A} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}\begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix}\begin{pmatrix} U^T & 0 \\ 0 & V^T \end{pmatrix}$. Hence,

$$e^{\mathcal{A}} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}\exp\begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix}\begin{pmatrix} U^T & 0 \\ 0 & V^T \end{pmatrix}. \tag{4.4}$$

Now,

$$\exp\begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix} = \cosh\begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix} + \sinh\begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix}$$

$$= \begin{pmatrix} \cosh(\Sigma) & 0 \\ 0 & \cosh(\Sigma) \end{pmatrix} + \begin{pmatrix} 0 & \sinh(\Sigma) \\ \sinh(\Sigma) & 0 \end{pmatrix}.$$

Thus,

$$\exp\begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix} = \begin{pmatrix} \cosh(\Sigma) & \sinh(\Sigma) \\ \sinh(\Sigma) & \cosh(\Sigma) \end{pmatrix}. \tag{4.5}$$

Putting together equations (4.4) and (4.5),

$$e^{\mathcal{A}} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \cosh(\Sigma) & \sinh(\Sigma) \\ \sinh(\Sigma) & \cosh(\Sigma) \end{pmatrix} \begin{pmatrix} U^T & 0 \\ 0 & V^T \end{pmatrix}$$

$$= \begin{pmatrix} \cosh\left(\sqrt{AA^T}\right) & A\left(\sqrt{A^TA}\right)^{\dagger} \sinh\left(\sqrt{A^TA}\right) \\ \sinh\left(\sqrt{A^TA}\right)\left(\sqrt{A^TA}\right)^{\dagger} A^T & \cosh\left(\sqrt{A^TA}\right) \end{pmatrix}.$$

The identities involving the off-diagonal blocks can be easily checked using the SVD of A. □

4.4.1 Interpretation of diagonal entries

In the context of undirected networks, the interpretation of the entries of the matrix exponential in terms of subgraph centralities and communicabilities is well-established, see e.g. [47]. In the case of directed networks and $e^{\mathcal{A}}$, things are not as clear. The network behind $\mathcal{A}$ can be thought of as follows: take the vertices from the original network A and make two copies of them, V and V'. Then, undirected edges exist between the two sets based on the following rule: $E' = \{(i, j')|\text{there is a directed edge from } i \text{ to } j \text{ in the original network}\}$. This creates a bipartite graph with $2n$ nodes: $1, 2, \ldots, n, n+1, n+2, \ldots, 2n$. We denote by $V(\mathcal{A})$ this set of nodes. The use of bipartization to treat rectangular and structurally unsymmetric matrices is of course standard in numerical linear algebra.

In the undirected case, each node had only one role to play in the network: any information that came into the node could leave by any edge. In the directed case, there are two roles for each node: that of a hub and that of an authority. It is unlikely

that a high ranking hub will also be a high ranking authority, but each node can still be seen as acting in both of these roles. In the network $\mathcal{A}$, the two aspects of each node are separated. Nodes $1, 2, \ldots, n$ in $V(\mathcal{A})$ represent the original nodes in their role as hubs and nodes $n+1, n+2, \ldots, 2n$ in $V(\mathcal{A})$ represent the original nodes in their role as authorities.

Given a directed network, an *alternating walk* of length k, starting with an out-edge, from node v_1 to node v_{k+1} is a list of nodes $v_1, v_2, ..., v_{k+1}$ such that there exists edge (v_i, v_{i+1}) if i is odd and edge (v_{i+1}, v_i) if i is even:

$$v_1 \to v_2 \leftarrow v_3 \to \cdots$$

An *alternating walk* of length k, starting with an in-edge, from node v_1 to node v_{k+1} in a directed network is a list of nodes $v_1, v_2, ..., v_{k+1}$ such that there exists edge (v_{i+1}, v_i) if i is odd and edge (v_i, v_{i+1}) if i is even:

$$v_1 \leftarrow v_2 \to v_3 \leftarrow \cdots$$

From graph theory (see also [29]), it is known that $[AA^TA\ldots]_{ij}$ (where there are k matrices being multiplied) counts the number of alternating walks of length k, starting with an out-edge, from node i to node j, whereas $[A^TAA^T\ldots]_{ij}$ (where there are k matrices being multiplied) counts the number of alternating walks of length k, starting with an in-edge, from node i to node j. That is, $[(AA^T)^k]_{ij}$ and $[(A^TA)^k]_{ij}$ count the number of alternating walks of length $2k$.

In the original network A, if node i is a good hub, it will point to many good authorities, which will in turn be pointed at by many hubs. These hubs will also point to many authorities, which will again be pointed at by many other hubs. Thus, if i is a good hub, it will show up many times in the sets of hubs described above. That is, there should be many even length alternating walks, starting with an out-edge, from

node i to itself. Giving a walk of length $2k$ a weight of $\frac{1}{(2k)!}$, these walks can be counted using the (i,i) entry of the matrix

$$I + \frac{AA^T}{2!} + \frac{AA^TAA^T}{4!} + \cdots + \frac{(AA^T)^k}{(2k)!} + \cdots$$

Letting $A = U\Sigma V^T$ be the SVD of A, this becomes:

$$U\left(I + \frac{\Sigma^2}{2!} + \frac{\Sigma^4}{4!} + \cdots + \frac{\Sigma^{2k}}{(2k)!} + \cdots\right)U^T$$

$$= U\cosh(\Sigma)U^T = \cosh(\sqrt{AA^T})\,.$$

The *hub centrality* of node i (in the original network) is thus given by

$$[e^{\mathcal{A}}]_{ii} = [\cosh(\sqrt{AA^T})]_{ii}.$$

This measures how well node i transmits information to the authoritative nodes in the network.

Similarly, if node i is a good authority, there will be many even length alternating walks, starting with an in-edge, from node i to itself. Giving a walk of length $2k$ a weight of $\frac{1}{(2k)!}$, these walks can be counted using the (i,i) entry of $\cosh(\sqrt{A^TA})$.

Hence, the *authority centrality* of node i is given by

$$[e^{\mathcal{A}}]_{n+i,n+i} = [\cosh(\sqrt{A^TA})]_{ii}.$$

It measures how well node i receives information from the hubs in the network.

Note that the traces of the two diagonal blocks in $e^{\mathcal{A}}$ are identical, so each accounts for half of the Estrada index of the bipartite graph. Also, recalling the well-known fact

that the eigenvalues of $\mathcal{A}$ are $\pm\sigma_i$ where σ_i denotes the singular values of A, we have

$$\mathrm{Tr}\,(e^{\mathcal{A}}) = \sum_{i=1}^{n} e^{\sigma_i} + \sum_{i=1}^{n} e^{-\sigma_i} = 2\sum_{i=1}^{n} \cosh\,(\sigma_i),$$

an identity that can also be obtained directly from the expression for $e^{\mathcal{A}}$ given in Proposition 4.4.1.

Returning to the example of the directed path graph with adjacency matrix A given by (4.2), one finds that using the diagonal entries of $e^{\mathcal{A}}$ to rank the nodes gives node 1 as the least authoritative node, and node n as the one with the lowest hub ranking, with all the other nodes being tied. Thus we see that, while e^{A} fails to differentiate between the nodes of this graph, using $e^{\mathcal{A}}$ yields a very reasonable hub/authority ranking of the nodes.

4.4.2 Interpretation of off-diagonal entries

Although not used in the remainder of this paper, for the sake of completeness we give here an interpretation of the off-diagonal entries of $e^{\mathcal{A}}$. As we will see, this interpretation is rather different from the one usually given for the off-diagonal entries of e^{A}, and provides information of a different nature on the structure of the underlying graph.

In discussing the off-diagonal entries of $\mathcal{A}$, there are three blocks to consider. First, there are the off-diagonal entries of the upper-left block, $\cosh(\sqrt{AA^T})$, then there are the off-diagonal entries of the lower-right block, $\cosh(\sqrt{A^TA})$. Finally, there is the off-diagonal block, $A\left(\sqrt{A^TA}\right)^{\dagger}\sinh\left(\sqrt{A^TA}\right)$ (the fourth block in $e^{\mathcal{A}}$ being its transpose).

From section 4.4.1, $[e^{\mathcal{A}}]_{ij} = [\cosh(\sqrt{AA^T})]_{ij}$, $1 \le i, j \le n$, counts the number of even length alternating walks, starting with an out-edge, from node i to node j, weighting walks of length $2k$ by a factor of $\frac{1}{(2k)!}$. When $i \neq j$, these entries measure how similar nodes i and j are as hubs. That is, if nodes i and j point to many of the same nodes, there will be many short even length alternating walks between them.

The *hub communicability* between nodes i and j, $1 \leq i, j \leq n$, is given by

$$[e^{\mathcal{A}}]_{ij} = [\cosh(\sqrt{AA^T})]_{ij}$$

This measures how similar nodes i and j are in their roles as hubs. That is, a large value of hub communicability between nodes i and j indicates that they point to many of the same authorities. In other words, they point to nodes which are authorities on the same subjects.

Similarly, $[e^{\mathcal{A}}]_{n+i,n+j} = [\cosh(\sqrt{A^TA})]_{ij}$, $1 \leq i, j \leq n$, counts the number of even length alternating walks, starting with an in-edge, from node i to node j, also weighing walks of length $2k$ by a factor of $\frac{1}{(2k)!}$. When $i \neq j$, these entries measure how similar the two nodes are as authorities. If i and j are pointed at by many of the same hubs, there will be many short even length alternating walks between them.

The *authority communicability* between nodes i and j, $1 \leq i, j, \leq n$, is given by

$$[e^{\mathcal{A}}]_{i+n,j+n} = [\cosh(\sqrt{A^TA})]_{ij}$$

This measures how similar nodes i and j are in their roles as authorities. That is, a large value of authority communicability between nodes i and j means that they are pointed to by many of the same hubs and, as such, are likely to contain information on the same subjects.

Let us now consider the off-diagonal blocks of $\mathcal{A}$. Here, $[\sinh(\sqrt{A^TA})]_{ij}$ counts the number of odd length alternating walks, starting with an out-edge, from node i to node j, weighing walks of length $2k+1$ by $\frac{1}{(2k+1)!}$. This measures the communicability between node i as a hub and node j as an authority.

The *hub-authority communicability* between nodes i and j (that is, the communica-

bility between node i as a hub and node j as an authority) is given by:

$$[e^{\mathcal{A}}]_{i,n+j} = [A\left(\sqrt{A^TA}\right)^{\dagger}\sinh\left(\sqrt{A^TA}\right)]_{ij}$$

$$= [\sinh\left(\sqrt{A^TA}\right)\left(\sqrt{A^TA}\right)^{\dagger}A^T]_{ji} = [e^{\mathcal{A}}]_{n+j,i}.$$

A large hub-authority communicability between nodes i and j means that they are likely in the same "part" of the directed network: node i tends to point to nodes that contain information similar to that on which node j is an authority.

4.4.3 Relationship with HITS

As described in 4.2, the HITS ranking of nodes as hubs and authorities uses only information from the dominant eigenvector of $\mathcal{A}$. Here we show that when using the diagonal of $e^{\mathcal{A}}$, we exploit information contained in all the eigenvectors of $\mathcal{A}$; moreover, the HITS rankings can be regarded as an approximation of those given by the diagonal entries of $e^{\mathcal{A}}$.

Assume the eigenvalues of $\mathcal{A}$ can be ordered as $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_{2n}$. Then, $\mathcal{A}$ can be written as $\mathcal{A} = \sum_{i=1}^{2n}\lambda_i\mathbf{u}_i\mathbf{u}_i^T$ where $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{2n}$ are the normalized eigenvectors of $\mathcal{A}$. Taking the exponential of $\mathcal{A}$, we get:

$$e^{\mathcal{A}} = \sum_{i=1}^{2n} e^{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T = e^{\lambda_1}\mathbf{u}_1\mathbf{u}_1^T + \sum_{i=2}^{2n} e^{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T.$$

Now, the hub and authority rankings come from the diagonal entries of $e^{\mathcal{A}}$:

$$\mathrm{diag}\,(e^{\mathcal{A}}) = e^{\lambda_1}\mathrm{diag}\,(\mathbf{u}_1\mathbf{u}_1^T) + \sum_{i=2}^{2n} e^{\lambda_i}\mathrm{diag}\,(\mathbf{u}_i\mathbf{u}_i^T).$$

Rescaling the hub and authority scores by e^{λ_1} does not alter the rankings; hence, we

can instead consider

$$\mathrm{diag}\,(e^{-\lambda_1}e^{\mathcal{A}}) = \mathrm{diag}\,(e^{\mathcal{A}-\lambda_1 I}) = \mathrm{diag}\,(\mathbf{u}_1\mathbf{u}_1^T) + \sum_{i=2}^{2n} e^{\lambda_i-\lambda_1}\mathrm{diag}\,(\mathbf{u}_i\mathbf{u}_i^T).$$

Now, the diagonal entries of the rank-one matrix $\mathbf{u}_1\mathbf{u}_1^T$ are just the squares of the (nonnegative) entries of the dominant eigenvector of $\mathcal{A}$; hence, the rankings provided by the first term in the expansion of $e^{\mathcal{A}}$ in the eigenbasis of $\mathcal{A}$ are precisely those given by HITS.

It is also clear that if $\lambda_1 \gg \lambda_2$, then the rankings provided by the diagonal entries of $e^{\mathcal{A}}$ are unlikely to differ much from those of HITS, since the weights $e^{\lambda_i-\lambda_1}$ will be tiny, for all $i = 2,\ldots,2n$. Conversely, if the gap between λ_1 and the rest of the spectrum is small ($\lambda_1 \approx \lambda_2$), then the contribution from the remaining eigenvectors, $\sum_{i=2}^{2n} e^{\lambda_i-\lambda_1}\mathrm{diag}\,(\mathbf{u}_i\mathbf{u}_i^T)$, may be non-negligible relative to the first term and therefore the resulting rankings could differ significantly from those obtained using HITS. In section 4.7 we will see examples of real networks illustrating both scenarios.

Summarizing, use of the matrix exponential for ranking hubs and authorities amounts to using the (squared) entries of *all* the eigenvectors of $\mathcal{A}$, weighted by the exponential of the corresponding eigenvalues. Of course, in place of the exponential, a number of other functions could be used; see the discussion in the next section. As shown above, the HITS ranking scheme uses the leading term only, corresponding to the approximation $e^{\mathcal{A}} \approx e^{\lambda_1}\mathbf{u}_1\mathbf{u}_1^T$. Between these two extremes one could also use approximations of the form

$$e^{\mathcal{A}} \approx \sum_{i=1}^{k} e^{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\,, \tag{4.6}$$

where $1 < k < n$; indeed, in most cases of practical interest a modest value of k ($\ll n$) will be sufficient for a very good approximation, since the eigenvalues of $\mathcal{A}$ are often observed to decay rapidly from a certain index k onward. We return on this topic in section 4.8.

4.5 Other ranking schemes

In this section we discuss a few other schemes that have been proposed in the literature, and compare them with the hub and authority centrality measures based on the exponential of $\mathcal{A}$.

4.5.1 Resolvent-based measures

As already noted, another function that has been successfully used to define centrality and communicability measures for an undirected network is the matrix resolvent, $(I - \alpha A)^{-1}$. This approach was pioneered early on by Katz [67], and variants thereof have since been used by numerous authors; see, e.g., [18, 21, 46, 47, 55, 94]. It is hardly necessary to mention the close relationship existing between the resolvent and the exponential function, which can be expressed via the Laplace transform. For the adjacency matrix $\mathcal{A}$ of a bipartite graph given by (4.3), the resolvent is easily determined to be

$$(I - \alpha\mathcal{A})^{-1} = \begin{pmatrix} (I - \alpha^2 AA^T)^{-1} & \alpha A(I - \alpha^2 A^T A)^{-1} \\ \alpha(I - \alpha^2 A^T A)^{-1} A^T & (I - \alpha^2 A^T A)^{-1} \end{pmatrix}. \tag{4.7}$$

The condition on α can be expressed as $0 < \alpha < 1/\sigma_1$, where $\sigma_1 = \|A\|_2$ denotes the largest singular value of A, the adjacency matrix of the undirected network. This ensures that the matrix in (4.7) is well-defined and nonnegative, with positive diagonal entries. The diagonal entries of $(I - \alpha^2 AA^T)^{-1}$ provide the hub scores, those of $(I - \alpha^2 A^T A)^{-1}$ the authority scores. A drawback of this approach is the need to select the parameter α, and the fact that different values of α may lead to different rankings. We have performed numerical experiments with this approach and we found that for certain values of α, particularly those close to the upper limit $1/\sigma_1$, the hub and authority rankings obtained with the resolvent function are not too different from those obtained with the matrix exponential, which is not surprising in light of the results

from Chapter 3. Moreover, not surprisingly, as the value of α is reduced, one obtains hub and authority rankings that are strongly correlated with the out- and in-degree of the nodes, respectively. Overall, because the resolvent tends to weigh short walks more heavily than the exponential, and since longer walks contribute relatively little to the centrality scores, it is fair to say that the exponential is less "degree biased" than the resolvent function. Also, since (for $\beta = 1$) the exponential rankings do not depend on a tunable parameter, they provide unambiguous rankings.

We note that "Katz" authority and hub scores may also be obtained by considering the column and row sums of the (nonsymmetric) matrix resolvent $(I - \alpha A)^{-1}$, where A is the adjacency matrix of the original digraph and $\alpha > 0$ is again assumed to be small enough for the corresponding Neumann series to converge. Indeed, the row sums of $(I - \alpha A)^{-1}$ count the number of (weighted) walks out of each node, while the column sums count the number of (weighted) walks into each node. Denoting by $\mathbf{1}$ the vector of all ones, hub and authority rankings can be obtained by solving the two linear systems

$$(I - \alpha A)\mathbf{y} = \mathbf{1} \quad \text{and} \quad (I - \alpha A^T)\mathbf{x} = \mathbf{1}\,, \tag{4.8}$$

respectively. Here the parameter α must satisfy $0 < \alpha < 1/\rho(A)$, where $\rho(A)$ denotes the spectral radius of A. The results of numerical experiments comparing the Katz scores with those based on the exponential of $\mathcal{A}$ are given in section 4.7. Here we observe that these Katz scores are also dependent on the choice of the parameter α, and similar considerations to those made for $(I - \alpha\mathcal{A})^{-1}$ apply.

A natural analogue to this approach is the use of row and column sums of the exponential e^A to rank hubs and authorities. Some results obtained with this approach are discussed in section 4.7. We note that this method is different from the *Exponentiated Inputs HITS Method* of [52]. This method is a modification to HITS, which was developed in order to correct the issue of non unique results in certain networks. If

the dominant eigenvalue of A^TA (and, consequently of AA^T) is not simple, then the corresponding eigenspace is multidimensional. This means that the choice of the initial vector affects the convergence of the HITS algorithm and different hub and authority vectors can be produced using different initial vectors. This can occur when A^TA is reducible, that is when the original network is not strongly connected. In [52], Farahat *et. al.* propose a modification to the HITS algorithm which amounts to replacing A and A^T with $e^A - I$ and $(e^A - I)^T$ in the HITS iteration. They prove that, as long as the original network is weakly connected, the dominant eigenvalue of $(e^A - I)^T(e^A - I)$ is simple. Thus, HITS with this exponentiated input produces unique hub and authority rankings. However, a result of this replacement is that nodes with 0 in-degree (or a low in-degree) are less important in the calculation of authority scores than nodes with a high in-degree. When there are many nodes with 0 in-degree or whose edges point to only a few other nodes, dropping these edges can greatly affect the HITS rankings.

An obvious disadvantage of this algorithm is its cost, since it requires iterating with a matrix exponential and its transpose. It can be implemented using only matrix- vector products involving $\tilde{A}$ and $\tilde{A}^T$ by means of techniques, like Krylov subspace methods, for evaluating the action of a matrix function on a given vector; see, e.g., [65, Chapter 13]. This approach leads to a nested iteration scheme, with HITS as the outer iteration and the Krylov method as the inner one.

4.5.2 PageRank and Reverse PageRank

As mentioned in Section 1.5.2, the (now) classical PageRank algorithm provides a means of finding the authoritative nodes in a digraph. Details about the PageRank algorithm can be found in Chapter 1, Section 1.7. It was pointed out in [54] that applying PageRank to the digraph obtained by reversing the direction of the edges provides a natural way to rank the hubs; this is usually referred to as *Reverse PageRank*. In other words, authority rankings are obtained by applying PageRank to the "Google" matrix

derived from A, and hub rankings are obtained by the same process applied to A^T. Like HITS, PageRank and Reverse PageRank are eigenvector-based ranking algorithms that do not take into account information about the network contained in the non-dominant eigenvectors. As already mentioned, it has been argued [79] that eigenvector-based algorithms tend to be degree-biased. Furthermore, like the Katz-type algorithms, the rankings obtained depend on the choice of a tuneable damping parameter. While the success of PageRank in finding authoritative nodes is well known and very well documented, much less is known about the effectiveness of Reverse PageRank in identifying hubs; some references are [4, 28, 96, 97]. We present the results of a few numerical experiments with PageRank and Reverse PageRank in section 4.7.

4.6 Examples on small digraphs

In this section and the next we illustrate the proposed method on some simple networks of small size, as well as on some larger data sets corresponding to real networks. We also compare our approach with HITS and other rankings schemes, including Katz, PageRank and Reverse PageRank. Here, we compare out and in-degree counts, HITS, and our proposed method to obtain hub and authority rankings in a few small digraphs. The purpose of this section is mostly pedagogical.

4.6.1 Example 1

Consider the small directed network in Fig. 4.1 (left panel). The adjacency matrix is given by

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

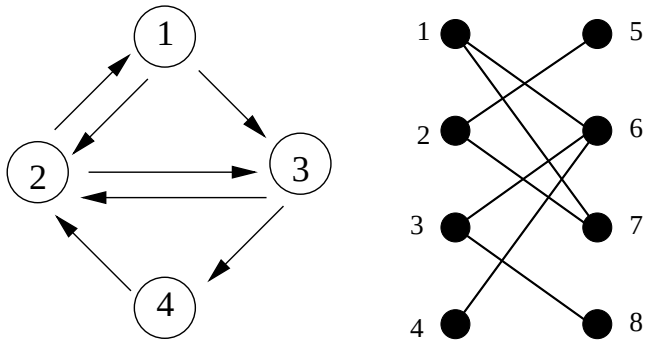

Figure 4.1: The original directed network from Example 1, with adjacency matrix A (left) and the bipartite network with adjacency matrix $\mathcal{A}$ (right).

The corresponding bipartite graph is shown in Fig. 4.1 (right panel). If hubs and authorities are determined simply using in-degree and out-degree counts, the result is as follows:

node	out-degree	in-degree
1	2	1
2	2	3
3	2	2
4	1	1

Under this ranking, the hub ranking of the nodes is: $\{1, 2, 3\,(\text{tie}); 4\}$. The authority ranking of the nodes is: $\{2; 3; 1, 4\,(\text{tie})\}$. We obtain somewhat different results using the HITS algorithm. The eigenvectors of AA^T and A^TA corresponding to the largest eigenvalue $\lambda_{\max} \approx 3.9563$, which is simple, yield the following rankings for hubs and authorities:

node	hub rank	authority rank
1	.3383	.0965
2	.1729	.4618
3	.2798	.2854
4	.2091	.1562

Here, the ranking for hubs is: $\{1; 3; 4; 2\}$. The ranking for authorities is: $\{2; 3; 4; 1\}$. Note that node 2, which was given a top hub score by looking just at the out-degrees, is judged by HITS as the node with the lowest hub score.

Using $e^{\mathcal{A}}$ as described above, the rankings for hub centralities and authority centralities are:

node	hub centrality = $[e^{\mathcal{A}}]_{ii}$	authority centrality = $[e^{\mathcal{A}}]_{4+i,4+i}$
1	2.3319	1.5906
2	2.2289	3.0209
3	2.2812	2.2796
4	1.6414	1.5922

With this method, the hub ranking of the nodes is: $\{1; 3; 2; 4\}$. The authority ranking is: $\{2; 3; 4; 1\}$. On this example, our method produces the same authority ranking as HITS. The hub ranking, however, is slightly different: both methods identify node 1 as the one with the highest hub score, followed by node 3; however, our method assigns the lowest hub score to node 4 rather than node 2. This is arguably a more meaningful ranking.

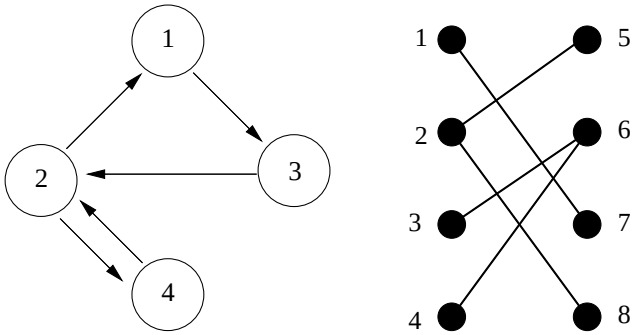

Figure 4.2: The original directed network from Example 2, with adjacency matrix A (left) and the bipartite network with adjacency matrix $\mathcal{A}$ (right).

4.6.2 Example 2

Consider the small directed network in Fig. 4.2 (left panel). The adjacency matrix is given by

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The corresponding bipartite graph is shown in Fig. 4.2 (right panel). If hubs and authorities are determined only using in-degrees and out-degrees, the result is as follows:

node	out-degree	in-degree
1	1	1
2	2	2
3	1	1
4	1	1

Under this criterion, the hub and authority rankings are both $\{2; 1, 3, 4\,(\text{tie})\}$. While it is intuitive that node 2 should be given a high score (both as an authority and as

a hub), just looking at the degrees does not allow one to distinguish the remaining nodes.

Consider now the use of HITS. The largest eigenvalue of AA^T (and A^TA) is $\lambda_{\max} = 2$ and it has multiplicity two. Thus, different starting vectors for the HITS algorithm may produce different rankings, as discussed in [52]. Starting from a constant authority vector $x^{(0)}$, as suggested in [68], produces the following scores:

node	hub rank	authority rank
1	.0000	.3333
2	.5000	.3333
3	.2500	.0000
4	.2500	.3333

The ranking for hubs is: $\{2; 3, 4\,(\text{tie}); 1\}$. The ranking for authorities is the following: $\{1, 2, 4\,(\text{tie}); 3\}$.

If the ranking is determined using $e^{\mathcal{A}}$ as described above, the resulting scores are:

node	hub centrality = $[e^{\mathcal{A}}]_{ii}$	authority centrality = $[e^{\mathcal{A}}]_{4+i,4+i}$
1	1.5431	1.5891
2	2.1782	2.1782
3	1.5891	1.5431
4	1.5891	1.5891

With this method, the hub ranking of the nodes is the same as in HITS: $\{2; 3, 4\,(\text{tie}); 1\}$. However, in the authority ranking, node 2 is the clear winner rather than being part of a three-way tie for first place: $\{2; 1, 4\,(\text{tie}); 3\}$. In this example, the method based on the matrix exponential is able to identify a top authority node by making use of additional spectral information.

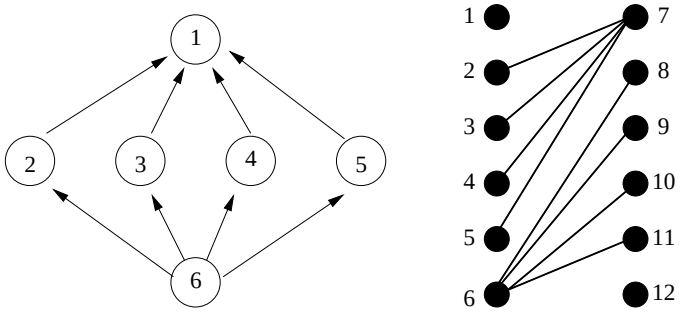

Figure 4.3: The original directed network from Example 3, with adjacency matrix A (left) and the bipartite network with adjacency matrix $\mathcal{A}$ (right).

4.6.3 Example 3

Let G be the small directed network in Fig. 4.3. The adjacency matrix is given by

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

If hubs and authorities are determined using only in-degrees and out-degrees, the result is:

node	out-degree	in-degree
1	0	4
2	1	1
3	1	1
4	1	1
5	1	1
6	4	0

The hub ranking of the nodes using degrees is: {6; 2,3,4,5 (tie); 1}. The authority ranking is {1; 2,3,4,5 (tie); 6}.

If the HITS algorithm is used, the resulting rankings are similar, but not exactly the same. Starting with a constant authority vector $x^{(0)}$, the results are:

node	hub rank	authority rank
1	.000	.200
2	.125	.200
3	.125	.200
4	.125	.200
5	.125	.200
6	.500	.000

The hub ranking of the nodes is: $\{6; 2, 3, 4, 5\,(\text{tie}); 1\}$. The authority ranking is: {1,2,3,4,5 (tie); 6}. Here, HITS does not differentiate between node 1 and nodes 2, 3, 4, and 5 in terms of the authority score, even though node 1 has by far the highest in-degree. This appears as a failure of HITS, since it is intuitive that node 1 should be regarded as very authoritative.

When $e^{\mathcal{A}}$ is used to calculate the hub and authority scores, node 1 does get a higher authority ranking than all the other nodes:

node	hub centrality = $[e^{\mathcal{A}}]_{ii}$	authority centrality = $[e^{\mathcal{A}}]_{6+i,6+i}$
1	1.0000	3.7622
2	1.6905	1.6905
3	1.6905	1.6905
4	1.6905	1.6905
5	1.6905	1.6905
6	3.7622	1.0000

Note that, if desired, the value 1 can be subtracted from these scores since it does not affect the relative ranking of the nodes. The hub ranking is {6; 2,3,4,5 (tie); 1}, and the authority ranking is: {1; 2,3,4,5 (tie); 6}.

4.7 Application to web graphs

Similarly to HITS, and in analogy to subgraph centrality for undirected networks, the rankings produced by the values on the diagonal of $e^{\mathcal{A}}$ can be used to rank websites as hubs and authorities in web searches (many other applications are of course also possible). Three of the data sets considered here are small web graphs consisting of web sites on various topics and can be found at [93] along with the website associated with each node; see also [20]. The experiments for this paper were run on the "Expanded" version of the data sets. Each data set is named after the corresponding topic.[1] In addition, we include results for the *wb-cs-stanford* data set from the University of Florida Sparse Matrix Collection [31]. This digraph represents a subset of the Stanford University web. In this section, the hub and authority rankings obtained from $e^{\mathcal{A}}$ are compared with those from HITS, Katz (using (4.8) with $\alpha = 1/(\rho(A) + 0.1)$), the row and column sums of the exponential e^A of the nonsymmetric matrix A, and PageRank/Reverse PageRank. For the latter we use the standard value $\alpha = 0.85$ for the damping parameter. All experiments are performed using Matlab Version 7.9.0 (R2009b) on a MacBook Pro running OS X Version 10.6.8, a 2.4 GHZ Intel Core i5 processor and 4 GB of RAM. For the purpose of these tests we use the built-in Matlab function `expm` to compute the matrix exponentials, and backslash to compute the Katz scores. Other approximations of $e^{\mathcal{A}}$ are discussed in section 4.8.

[1] It should be noted, however, that in the node list for the adjacency matrix, the node labeling begins with 1 and in the list of websites associated with the nodes found at [93], node labeling begins at 0. Thus, node i in the adjacency matrix is associated with website $i - 1$.

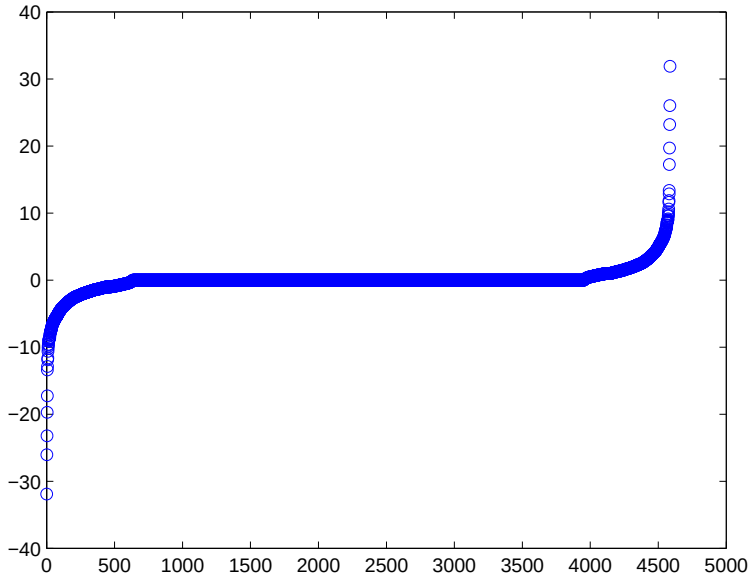

Figure 4.4: Plot of the eigenvalues of the expanded abortion matrix $\mathcal{A}$.

4.7.1 Abortion data set

The *abortion* data set contains $n = 2293$ nodes and $m = 9644$ directed edges. The expanded matrix $\mathcal{A} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$ has order $N = 2n = 4586$ and contains $2m = 19288$ nonzeros. The maximum eigenvalue of $\mathcal{A}$ is $\lambda_N \approx 31.91$ and the second largest eigenvalue is $\lambda_{N-1} \approx 26.04$. In this matrix, the largest eigenvalue is fairly well-separated from the second largest so that one would expect the HITS rankings (which only use information from the dominant eigenpair of $\mathcal{A}$) to be reasonably close to the rankings from $e^{\mathcal{A}}$ (which use information from all of the eigenvalues and corresponding eigenvectors). A plot of the eigenvalues of the expanded abortion data set matrix can be found in Fig. 4.4. Note the high multiplicity of the zero eigenvalue in this matrix, as well as in the adjacency matrices of the computational complexity and death penalty data sets considered below. Due to this, the numerical rank of the matrix is very low and, as such, the diagonal values of e^A can be estimated using only a few eigenvalues (see section 4.8 for further discussion on this).

Table 4.1: Top 10 hubs of the abortion web graph, ranked using $[e^{\mathcal{A}}]_{ii}$, HITS, Katz, e^A row sums and Reverse PageRank with $\alpha = 0.85$.

$[e^{\mathcal{A}}]_{ii}$	HITS	Katz	e^A rs	RPR
48	48	80	80	125
1021	1006	1431	1431	2184
1007	1007	1432	1432	79
1006	1021	1387	1426	81
1053	1053	1388	1425	48
1020	1020	1389	1415	1424
987	960	1397	1388	1447
990	968	1425	1389	78
985	969	1426	1397	134
989	970	1415	1387	1445

Table 4.2: Top 10 authorities of the abortion web graph, ranked using $[e^{\mathcal{A}}]_{ii}$, HITS, Katz, e^A column sums and PageRank with $\alpha = 0.85$.

$[e^{\mathcal{A}}]_{ii}$	HITS	Katz	e^A cs	PR
967	939	1430	1430	1609
958	958	1387	1387	1941
939	967	1425	1425	1948
962	961	1426	1426	1608
963	962	1429	1417	587
964	963	1396	1409	1610
961	964	1405	1429	2045
965	965	1406	1406	317
966	966	1409	1396	2191
587	1582	1417	1405	753

The top 10 hubs and authorities of the abortion data set, as determined using the diagonal entries of $e^{\mathcal{A}}$, HITS with constant initial vector, the row/column sums of $(I - cA)^{-1}$ ("Katz"), the row/column sums of e^A and Reverse PageRank/PageRank are shown in Tables 4.1 and 4.2. We observe that there is a good deal of agreement between the $e^{\mathcal{A}}$ rankings and the HITS ones: indeed, both methods identify the web-sites labeled 48, 1021, 1007, 1006, 1053, 1020 as the top 6 hubs, and both pick web site 48 as the top one. Also, there are 7 web sites identified by both methods as being among the top 10 authorities. The top authority identified by HITS is ranked third by

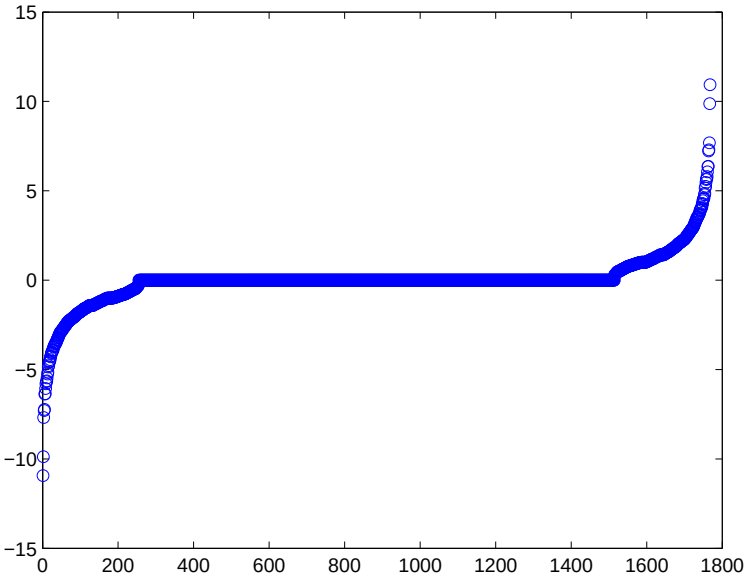

Figure 4.5: Plot of the eigenvalues of the expanded computational complexity matrix $\mathcal{A}$.

$e^{\mathcal{A}}$, and conversely the top authority identified by $e^{\mathcal{A}}$ is third in the HITS ranking. The Katz rankings and those based on e^{A} show considerable agreement with one another, but are very different from the HITS ones and from those based on $e^{\mathcal{A}}$. Node 48, which is the top-ranked hub according to HITS and $e^{\mathcal{A}}$, is now not even among the top 100. Conversely, node 80, which is ranked the top hub by Katz and e^{A}, is not in the top 100 nodes according to HITS or to $e^{\mathcal{A}}$. This is not too surprising, since the metrics based on A and those based on $\mathcal{A}$ are obtained by counting rather different types of graph walks. Finally, for this network Reverse PageRank and PageRank return rankings with almost no overlap with any of the other methods.

4.7.2 Computational complexity data set

The *computational complexity* data set contains $n = 884$ nodes and $m = 1616$ directed edges. The expanded matrix $\mathcal{A}$ has order $N = 2n = 1768$ and contains $2m = 2232$ nonzeros. The maximum eigenvalue of $\mathcal{A}$ is $\lambda_N \approx 10.93$ and the second largest eigen-

value is $\lambda_{N-1} \approx 9.86$. Here, the (relative) spectral gap between the first and the second eigenvalue is smaller than in the previous example; consequently, we expect the rankings produced using $e^{\mathcal{A}}$ and HITS to be less similar than for the abortion data set. A plot of the eigenvalues of the expanded computational complexity data set matrix can be found in Fig. 4.5.

The top 10 hubs and authorities of the computational complexity data set, determined by the various ranking methods, can be found in Tables 4.3 and 4.4. As expected, we see less agreement between HITS and the diagonals of $e^{\mathcal{A}}$. Concerning the hubs, both methods agree that the web site labelled 57 is by far the most important hub on the topic of computational complexity. However, the method based on $e^{\mathcal{A}}$ identifies as the second most important hub the web site corresponding to node 17, which is ranked only 39th by HITS. The two methods agree on the next three hubs, but after that they return completely different results. The difference is even more pronounced for the authority rankings. The method based on $e^{\mathcal{A}}$ clearly identifies web site 1 as the most authoritative one, whereas HITS relegates this node to 8th place. The top authority acording to HITS, web site 719, places 5th in the ranking obtained by $e^{\mathcal{A}}$. The two methods agree on only two other web sites as being in the top 10 authorities (717 and 727). The Katz rankings and those based on e^{A} show little overlap for this data set, although node 57 is clearly considered an important hub by all measures. A natural question is how much these results are affected by the choice of the parameter c used to compute the Katz scores. We found experimentally that, in contrast to the situation for the other data sets, small changes in the value of α can significantly affect the Katz ranking for this particular data set. Changing the value of c to $\alpha = 1/(\rho(A) + 0.3)$ results in hub and authority rankings that are much closer to those given by the column/row sums of e^{A}. The potential sensitivity to α is a clear drawback of the Katz-based approach compared to the methods based on the matrix exponential. Coming to (Reverse) PageRank, it is interesting to note that for this data set it provides rankings

Table 4.3: Top 10 hubs of the computational complexity web graph, ranked using $[e^{\mathcal{A}}]_{ii}$, HITS, Katz, e^A row sums and Reverse PageRank with $\alpha = 0.85$.

$[e^{\mathcal{A}}]_{ii}$	HITS	Katz	e^A rs	RPR
57	57	56	57	57
17	634	709	56	56
644	644	57	17	17
643	721	697	51	51
634	643	705	634	21
106	544	690	21	11
119	632	714	255	255
529	801	708	173	12
86	640	712	709	13
162	639	715	45	45

Table 4.4: Top 10 authorities of the computational complexity web graph, ranked using $[e^{\mathcal{A}}]_{ii}$, HITS, Katz, e^A column sums and PageRank with $\alpha = 0.85$.

$[e^{\mathcal{A}}]_{ii}$	HITS	Katz	e^A cs	PR
1	719	688	673	673
315	717	685	1	664
673	727	673	664	534
148	723	690	534	45
719	808	56	45	2
717	735	686	473	1
2	737	664	315	376
45	1	1	376	341
727	722	45	688	50
534	770	534	599	51

that are at least in partial agreement with some of the other measures, especially those based on e^A. Looking at the authority scores, we also notice a good degree of overlap among all methods, except HITS. Due to the small spectral gap, HITS is probably the least reliable of all ranking methods on this particular data set.

4.7.3 Death penalty data set

The *death penalty* data set contains $n = 1850$ and $m = 7363$ directed edges. The expanded matrix $\mathcal{A}$ has order $N = 2n = 3700$ and contains $m = 14726$ nonzeros. The

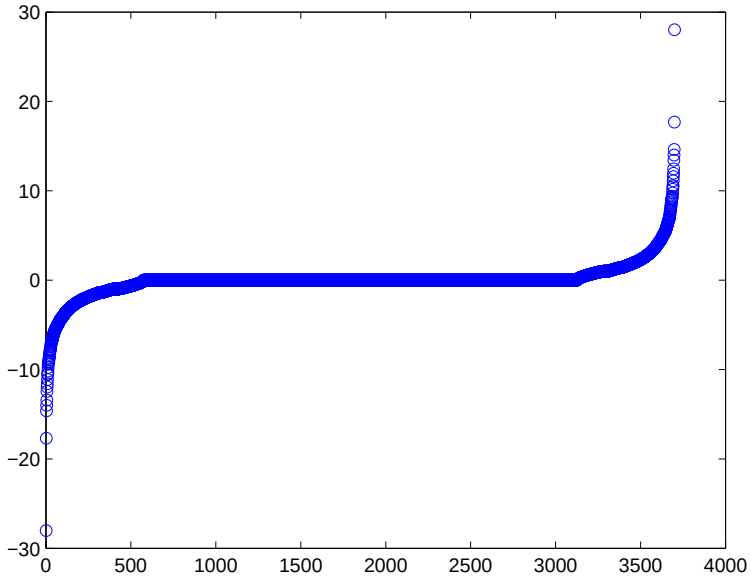

Figure 4.6: Plot of the eigenvalues of the expanded death penalty matrix $\mathcal{A}$.

maximum eigenvalue of $\mathcal{A}$ is $\lambda_N \approx 28.02$ and the second largest eigenvalue $\lambda_{N-1} \approx 17.68$. In this case, the largest and second largest eigenvalues are quite far apart, and the relative gap is larger than in the previous examples. A plot of the eigenvalues of the expanded death penalty matrix can be found in Fig. 4.6.

Due to the presence of a large spectral gap, much of the information used in forming the rankings of $e^{\mathcal{A}}$ is also used in the HITS ranking, and we expect the two methods to produce similar results; see section 4.4.3. Indeed, as shown in Table 4.5 (hubs) and Table 4.6 (authorities), in this case the top 10 rankings produced by the two methods are actually identical.

Looking at the Katz scores and those based on e^{A}, we see in this case a great deal of overlap between these two, but almost completely different rankings compared to HITS and $e^{\mathcal{A}}$ (although node 210 is clearly an important hub by any standard). Note that node 1632 is both the top hub and the top authority according to Katz and to e^{A}. PageRank and Reverse PageRank show a limited amount of overlap with the other measures; nevertheless, nodes 210 and 1632 are also found to be important hubs and

Table 4.5: Top 10 hubs of the death penalty web graph, ranked using $[e^{\mathcal{A}}]_{ii}$, HITS, Katz, e^A row sums and Reverse PageRank with $\alpha = 0.85$.

$[e^{\mathcal{A}}]_{ii}$	HITS	Katz	e^A rs	RPR
210	210	1632	1632	210
637	637	133	133	1632
413	413	1671	1671	70
1586	1586	552	552	95
552	552	1651	1651	135
462	462	1673	210	133
930	930	1328	1673	55
542	542	1653	1653	958
618	618	210	1328	1077
1275	1275	1709	1709	315

nodes 1632, 1 and 4 are found to be authoritative, in agreement with some of the other measures.

4.7.4 Stanford web graph

The *wb-cs-stanford* data set from the University of Florida sparse matrix collection contains $n = 9914$ nodes and $m = 36854$ directed edges. The expanded matrix $\mathcal{A}$ has order $N = 2n = 19828$ and contains $m = 73708$ nonzeros. The maximum eigenvalue of $\mathcal{A}$ is $\lambda_N \approx 38.38$ and the second largest is $\lambda_{N-1} \approx 32.12$, hence there is a sizeable gap. Tables 4.7-4.8 report the results obtained with the various ranking schemes.

The first thing to observe is the remarkable agreement between the HITS, $e^{\mathcal{A}}$, Katz, and e^A rankings of both hubs and authorities. This in stark contrast with the results for the previous three data sets. Moreover, many of the nodes that are ranked highly as hubs are also ranked highly as authorities. A plausible explanation of these observations is that the adjacency matrix A for this digraph is much closer to being symmetric than in the other cases. Indeed, the percentage of "bidirectional" edges in the wb-cs-stanford graph is 47.63%; the corresponding percentages for the abortion, computational complexity and death penalty graphs are just 2.72%, 2.97% and 4.02%,

Table 4.6: Top 10 authorities of the death penalty web graph, ranked using $[e^A]_{ii}$, HITS, Katz, e^A column sums and PageRank with $\alpha = 0.85$.

$[e^A]_{ii}$	HITS	Katz	e^A cs	PR
4	4	1632	1632	993
1	1	1662	1662	667
6	6	1697	1697	3
7	7	1689	1689	736
10	10	1653	1653	735
16	16	1671	1671	1632
2	2	1675	1675	42
3	3	1684	1684	1
44	44	798	789	4
27	27	1652	1654	1212

Table 4.7: Top 10 hubs of the wb-cs-stanford web graph, ranked using $[e^A]_{ii}$, HITS, Katz, e^A row sums and Reverse PageRank with $\alpha = 0.85$.

$[e^A]_{ii}$	HITS	Katz	e^A rs	RPR
6562	6562	6562	6562	251
6838	6838	6837	6837	252
6840	6837	6838	6838	253
6837	6839	6839	6839	254
6839	6840	6840	6840	271
6616	6616	6669	6669	2240
6765	6615	6668	6668	2241
6615	6765	6670	6670	2242
6669	6669	6616	6616	2243
6731	6731	6615	6615	348

respectively.

Interestingly, the (Reverse) PageRank results are now drastically different fron the ones provides by all the other measures in nearly all cases. The only (partial) exception is that PageRank finds nodes 6837, 6839 and 6840 to be among the top 10 authorities; these three nodes are identified as the three most authoritative ones by the remaining methods.

Table 4.8: Top 10 authorities of the wb-cs-stanford web graph, ranked using $[e^{\mathcal{A}}]_{ii}$, HITS, Katz, e^A column sums and PageRank with $\alpha = 0.85$.

$[e^{\mathcal{A}}]_{ii}$	HITS	Katz	e^A cs	PR
6837	6837	6837	6837	2264
6840	6839	6839	6839	8226
6839	6840	6840	6840	8059
6838	6838	6838	6838	8057
6617	6617	6573	6573	4485
6615	6615	6574	6575	5707
6766	6614	6575	6576	8225
6764	6616	6576	6577	6837
6616	6764	6577	6578	6839
6614	6766	6578	6579	6840

4.8 Approximating the matrix exponential

For the purpose of ranking hubs and authorities in a directed network, only the main diagonal of $e^{\mathcal{A}}$ is required. As was discussed in Chapter 1, Section 1.7, this can be done without having to compute *all* the entries in $e^{\mathcal{A}}$. When applying the approach based on Gauss quadrature rules to the $2n \times 2n$ matrix $\mathcal{A}$, only matrix-vector products with A and its transpose are required, just like in the HITS algorithm. If only the hub scores are wanted, it is also possible to apply the described techniques to the symmetric matrix AA^T using the function $f(\lambda) = \cosh(\sqrt{\lambda})$; the same applies if only the authority scores are wanted, working this time with the matrix $A^T A$. The problem with this approach is that only estimates (rather than increasingly accurate lower and upper bound) can be obtained, due to the fact that the function $f(\lambda) = \cosh(\sqrt{\lambda})$ is not strictly completely monotonic on the positive real axis. We refer to [9] for details. In our experiments we always work with the matrix $\mathcal{A}$, since we are interested in computing both hub and authority scores.

Table 4.9: The number of iterations necessary for the top 10 hubs or authorities to be determined (not necessarily in the correct order).

Dataset	hub (lower bound)	hub (upper bound)
Abortion	> 40	> 40
Comp. Complex.	3	3
Death Penalty	5	3
Stanford	8	8

Dataset	authority (lower bound)	authority (upper bound)
Abortion	2	2
Comp. Complex.	4	5
Death Penalty	4	2
Stanford	7	8

4.8.1 Test results

Accurate evaluation of *all* the diagonal entries of e^A using quadrature rules may be too expensive for truly large-scale graphs. In most applications, fortunately, it is not necessary to rank all the nodes in the network: only the top few hubs and authorities are likely to be of interest. When using quadrature rules, the number of quadrature nodes (Lanczos iterations) required to correctly rank the nodes as hubs or authorities varies and depends on both the eigenvalues of e^A and how close the diagonal entries are in value. If the rankings of the nodes are very close, it can take many iterations for the ordering to be exactly determined. However, since estimates for diagonal entries are calculated individually, once the top 10 (say) nodes have been identified, additional iterations can be performed only on these nodes in order to determine their exact ranking.

Our approach exploits the monotonicity of the Gauss-Radau bounds: as soon as the lower bound for node i is above the upper bounds for other nodes, we know that node i will be ranked higher than those othe nodes. This observation leads to a simple algorithm for identifying the top-k nodes. The number of Lanczos iterations per node necessary to identify the top $k = 10$ hubs and authorities, using Gauss-Radau

Table 4.10: The number of iterations necessary for the top 10 hubs or authorities to be ranked in the top 30.

Dataset	hub (lower bound)	hub (upper bound)
Abortion	5	4
Comp. Complex.	2	2
Death Penalty	2	2
Stanford	7	4

Dataset	authority (lower bound)	authority (upper bound)
Abortion	2	2
Comp. Complex.	4	2
Death Penalty	2	2
Stanford	2	4

lower and upper bounds, for the four data sets from section 4.7 is given in Table 4.9. Our implementation is based on Meurant's Matlab code [77], From the table it can be seen that, in most cases, only 2-5 iterations per node are needed. An exception is the determination of the top 10 hubs of the abortion data set, for which the number of iterations is large (> 40). This is due to a cluster of nodes (nodes 960 and 968-990) that have nearly identical hub rankings. These nodes' scores agree to 15 significant digits. However, for most applications, if a subset of nodes are so closely ranked, their exact ordering may not be so important. Table 4.10 reports the number of Lanczos iterations neeeded for the top $k = 10$ hubs and authorities to be ranked at least in the top 30. Here, the number of iterations per node needed is never more than 7. The total cost is thus $O(n)$ Lanczos iterations, again leading to an $O(n^2)$ overall complexity. Various enhancements can be used to reduce costs, including the use of sparse-sparse mat-vecs in the Lanczos iteration, and the exclusion of nodes with zero out-degree (for hub computations) and zero in-degree (for authority computations) from the top-k calculations. It is also safe to assume that in most cases of interest, one can also exclude nodes with in- and out-degree 1 from the computations, leading to further savings.

4.9 Conclusions and outlook

In this chapter we have presented a new approach to ranking hubs and authorities in directed networks using functions of matrices. Bipartization is used to transform the original directed network into an undirected one with twice the number of nodes. The adjacency matrix of the bipartite graph is symmetric, and this allows the use of subgraph centrality (and communicability) measures for undirected networks. We showed that the diagonal entries of the matrix exponential provide hub and authority rankings, and we gave an interpretation for the off-diagonal entries (communicabilities). Unlike HITS, the results are independent of any starting vectors; and unlike the Katz-based ranking schemes, there is no dependency on an arbitrary parameter.

Several examples, both synthetic and corresponding to real data sets, have been used to demonstrate the effectiveness of the proposed ranking algorithms relative to HITS and to other ranking schemes based on the matrix resolvent and on the exponential of the adjacency matrix of the original digraph. Our experiments indicate that our method results in rankings that are frequently different from those computed by HITS, at least in the absence of large gaps between the dominant singular value of the adjacency matrix A and the remaining ones. This is to be expected, since our method uses information from all the singular spectrum of the network, not just the dominant singular triplets.

As usual in this field, there is no simple way to compare different ranking schemes, and therefore it is impossible to state with certainty that a ranking scheme will give "better" results than a different scheme in practice. It is, however, certainly the case that the method based on the exponential of $\mathcal{A}$ takes into account more spectral information than HITS does; moreover, the rankings so obtained are unambiguous, in that they do not depend on the choice of an initial guess or on a tuneable parameter. As we saw, the latter is a weak spot of Katz-like methods, and a similar case can be made for PageRank

and Reverse PageRank.

Compared to HITS, the new technique has a higher computational cost. We showed how Gaussian quadrature rules can be used to quickly identify the top ranked hubs and authorities for networks involving thousands of nodes. We note that such schemes require a symmetric input matrix and are not readily applicable to nonsymmetric matrices, since in this case one can only hope for estimates instead of lower and upper bounds.

Future work should include a more efficient implementation and tests on larger networks. It is likely that the proposed approach based on Gaussian quadrature will prove to be too expensive for truly large-scale networks with millions of nodes. We hope to explore techniques similar to those presented in [18, 53] and [88] in order to extend our methodology to truly large-scale networks. Another relevant question is the study of the rate of convergence of the Lanczos algorithm for estimating bilinear forms associated with adjacency matrices of graphs of different types.

5 Conclusions

In this thesis, we have studied problems arising from the calculation of node centrality rankings, specifically those based on functions of matrices. We introduced the total communicability of a node i, given by the ith row sum of the exponential of the adjacency matrix of the network, as a viable centrality method. We demonstrated that the rankings provided by total communicability are often in strong agreement with those produced by subgraph centrality, especially for highly ranked nodes. This was shown through extensive numerical tests on both synthetic and real-world networks. These tests also demonstrated that total communicability is as good as subgraph centrality in identifying essential nodes in a Yeast PPI network [10, 42, 43]. We also showed that total communicability scores can be estimated extremely quickly using Krylov subspace methods, allowing total communicability to be used to rank nodes on very large networks where the use of subgraph centrality is infeasible. Additionally, we introduced the total communicability of a network, given by the average node total communicability, as a global measure of network connectivity. Future work in this area includes providing more precise guidelines on when subgraph centrality and total communicability rankings will be in strong agreement as well as investigating the use of total network communicability in the design of communication networks.

We also analyzed the relationship between four parameterized centrality measures based on the matrix exponential and resolvent with degree and eigenvector centrality. The centrality measures examined were subgraph centrality and total communicability (given by $[e^{\beta A}]_{ii}$ and $[e^{\beta A}\mathbf{1}]_i$, respectively, with $\beta > 0$) and resolvent and Katz central-

ity (given by $[(I - \alpha A)^{-1}]_{ii}$ and $[(I - \alpha A)^{-1}\mathbf{1}]_i$, $0 < \alpha < \frac{1}{\lambda_1}$). We proved that as α and β approach 0, the rankings produced by all four centrality measures converge to those produced by degree centrality. As α approaches $\frac{1}{\lambda_1}$ and β approaches infinity, the rankings converge to those produced by eigenvector centrality. The speed of this convergence is based on the size of the spectral gap $\lambda_1 - |\lambda_2|$. These results help to explain the observed correlation between degree and eigenvector centrality on many real-world complex networks. We also conducted extensive numerical experiments to determine how the convergence to degree and eigenvector centrality rankings behaved on real-world networks. Based on these experiments, we were able to provide guidelines on the choice of α and β to ensure the rankings produced are as different as possible from those produced by degree and eigenvector centrality.

We propose an extension of the measure of subgraph centrality to directed networks based on a bipartization of the network and explicitly determine the exponential of this bipartized network. We show how this extension can be used to rank hubs and authorities using the diagonal entries of the exponential of the bipartized network. We also show that the off-diagonal entries contain information about the ability of pairs of nodes to communicate in various manners. We used this method to rank hubs and authorities on a variety of directed web networks and compared the rankings to those produced by HITS, PageRank, and Reverse PageRank. Our proposed method is more expensive than the other methods tested. Although there are currently methods to estimate the centralities based on Gaussian quadrature, they are still too expensive to use on very large networks. Future work includes the investigation of techniques to more quickly estimate these centralities and their use in the study of real-world applications.

Appendices

A Algorithms

HITS algorithm, adapted from [73, p. 118]

Input: A: adjacency matrix; $\mathbf{x}^{(0)}$: starting vector; ε: convergence tolerance

Output: $\mathbf{x}$: authority vector; $\mathbf{y}$: hub vector

initialization: $\mathbf{x} = \mathbf{x}^{(0)}$;

$r = 1$;

while $r \geq \varepsilon$ **do**

$\mathbf{x}_o = \mathbf{x}$;

$\mathbf{x} = A\mathbf{x}$;

$\mathbf{x} = A^T\mathbf{x}$;

$\mathbf{x} = \frac{1}{\|\mathbf{x}\|_2}\mathbf{x}$;

$r = \|\mathbf{x} - \mathbf{x}_o\|_1$;

end

$y = A\mathbf{x}$;

Algorithm 1: HITS Algorithm

PageRank Algorithm, adapted from [73, p. 42]

Input: H: a row normalized hyperlink matrix; $\boldsymbol{\pi}^{(0)}$: starting vector; α: scaling parameter ; ε: convergence tolerance

Output: π: PageRank vector

initialization: $\boldsymbol{\pi} = \boldsymbol{\pi}^{(0)}$;

$r = 1$;

Replace zero rows of H with $\frac{1}{n}\mathbf{1}^T$;

while $r \geq \varepsilon$ **do**

$\mathbf{x} = \boldsymbol{\pi}$;

$\boldsymbol{\pi}^T = \alpha\boldsymbol{\pi}^T H + (1-\alpha)\frac{1}{n}\mathbf{1}^T\mathbf{1}$;

$r = \|\boldsymbol{\pi} - \mathbf{x}\|_1$;

end

Algorithm 2: PageRank Algorithm

Lanczos Algorithm, adapted from [76, p. 8] Input: A: a real, symmetric matrix; $\mathbf{v}$: starting vector

Output: entries of a symmetric tridiagonal matrix T_k: $\alpha_k = [T_k]_{k,k}$ and $\mu_k = [T_k]_{k,k-1}$

$\mathbf{v}_1 = \frac{1}{\|\mathbf{v}\|}\mathbf{v}$;

$\alpha_1 = \mathbf{v}_1^T A\mathbf{v}_1$;

$\tilde{\mathbf{v}_2} = A\mathbf{v}_1 - \alpha_1\mathbf{v}_1$;

for $k = 2, 3, \ldots$ **do**

$\mu_k = \|\tilde{\mathbf{v}_k}\|$;

$\mathbf{v}_k = \frac{1}{\mu_k}\tilde{\mathbf{v}_k}$;

$\mathbf{u}_k = A\mathbf{v}_k - \mu_k\mathbf{v}_{k-1}$;

$\alpha_k = \mathbf{v}_k^T\mathbf{u}_k$;

$\tilde{\mathbf{v}_{k+1}} = \mathbf{u}_k - \alpha_k\mathbf{v}_k$;

end

Algorithm 3: Lanczos Algorithm

B Additional experiments

Here, we present plots of of the intersection distances between the top 10 nodes, ranked by a variety of centrality measures, for a selection of real-world networks. More details can be found in Chapter 3.

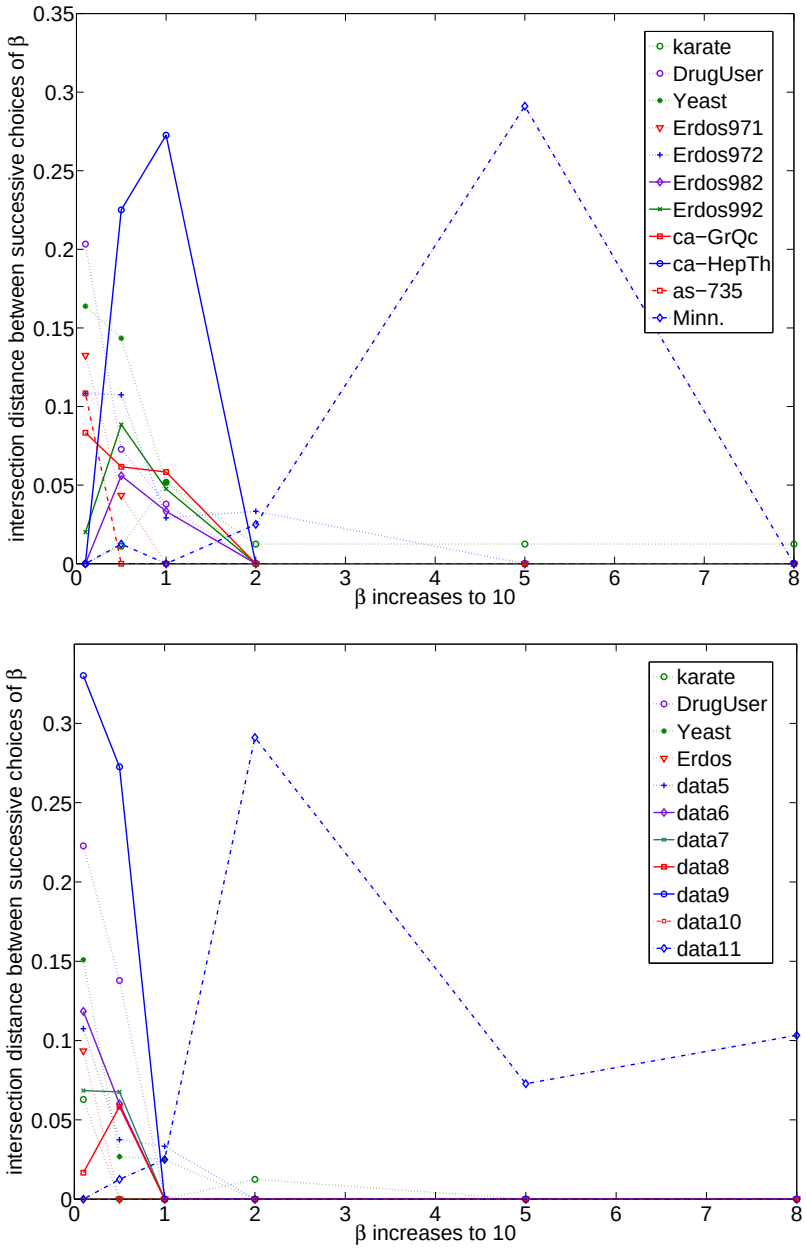

Figure B.1: The intersection distances between the exponential subgraph centrality (left) or total communicability (right) rankings produced by successive choices of β on the top 10 ranked nodes of each network. Each line corresponds to a network in Table 2.6.

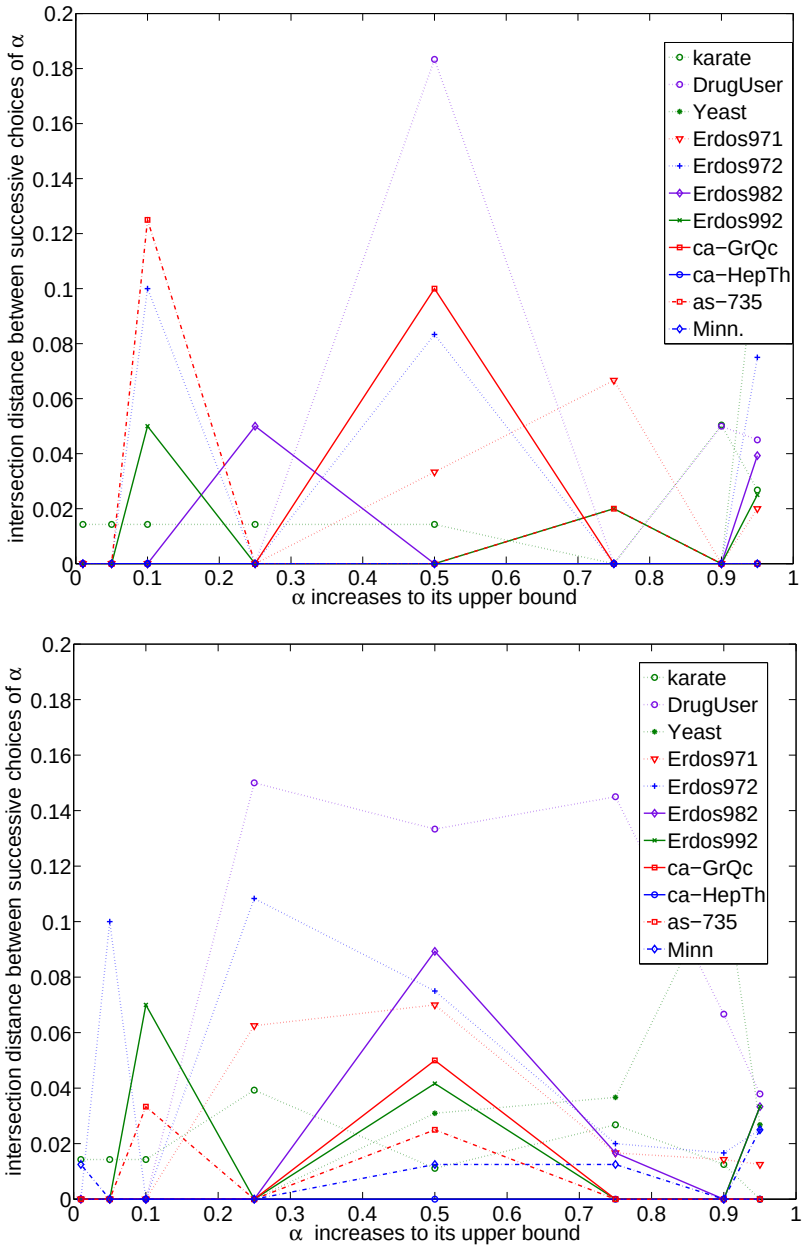

Figure B.2: The intersection distances between resolvent subgraph centrality (left) or Katz centrality (right) rankings produced by successive choices of α on the top 10 ranked nodes of each network. Each line corresponds to a network in Table 2.6.

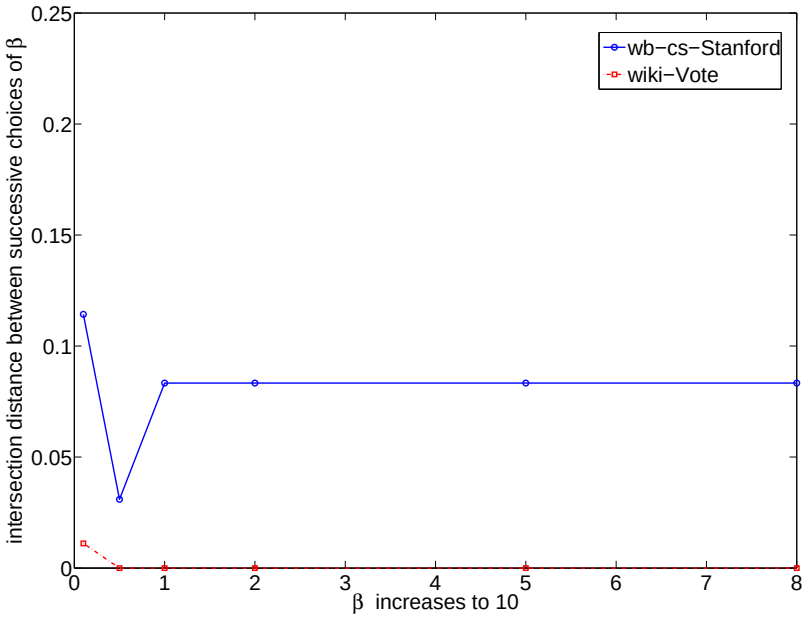

Figure B.3: The intersection distances between the total communicability rankings of the top 10 nodes of each network produced by successive choices of β. Each line corresponds to a network in Table 3.1.

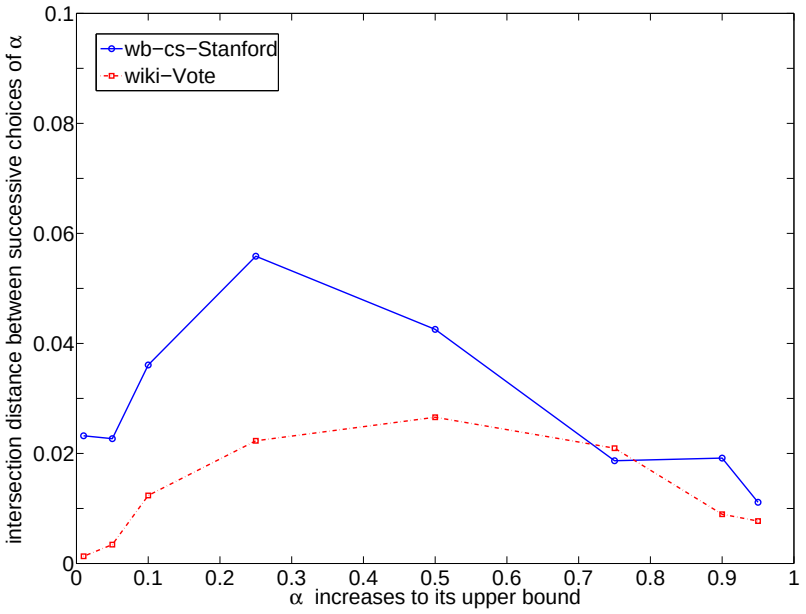

Figure B.4: The intersection distances between the Katz centrality rankings produced by successive choices of α. Each line corresponds to a network in Table 3.1.

References

[1] M. Afanasjew, M. Eiermann, O. G. Ernst, and S. Güttel, *Implementation of a restarted Krylov subspace method for the evaluation of matrix functions*, Lin. Algebra Appl., 429 (2008), pp. 2293–2314. 29

[2] M. A. M. de Aguiar and Y. Bar-Yam, *Spectral analysis and the dynamic response of complex networks*, Phys. Rev. E, 71 (2005), 016106. 3, 33

[3] R. Albert, H. Jeong, and A. L. Barabási, *Error and attack tolerance of complex networks*, Nature 406 (2000), pp. 378–382. 14

[4] Z. Bar-Yossef and L.-T. Mashiach, *Local approximation of PageRank and Reverse PageRank*, Proceedings CKIM'08, October 26–30, 2008, Napa Valley, California. 23, 110

[5] A. L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science, 286 (1999), pp. 509–512. 14, 15, 38

[6] M. Benzi, *A note on walk entropies in graphs*, Linear Algebra Appl., published online, 2013. DOI: 10.1016/j.laa.2013.12.014. 72

[7] M. Benzi and P. Boito, *Quadrature rule-based bounds for functions of adjacency matrices*, Lin. Algebra Appl., 433 (2010), pp. 637–652. 25, 29, 35, 98, 99

[8] M. Benzi, E. Estrada, C. Klymko, *Ranking hubs and authorities using matrix functions*, Lin. Algebra Appl., 438 (2013), pp. 2447–2474. 4, 5, 20, 28, 32, 54, 65, 66, 83

[9] M. Benzi and G. H. Golub, *Bounds on the entries of matrix functions with applications to preconditioning*, BIT, 39 (1999), pp. 417–438. 35, 126

[10] M. Benzi and C. Klymko, *Total communicability as a centrality measure*, Journal of Complex Networks 1 (2013), pp. 124-149. 5, 17, 65, 66, 131

[11] M. BENZI AND V. KUHLEMANN, *Chebyshev Acceleration of the GeneRank Algorithm*, Electronic Transactions on Numerical Analysis, 40 (2013), pp. 311–320. 14, 20, 62

[12] M. BIANCHINI, M. GORI, AND F. SCARSELLI, *Inside PageRank*, ACM Trans. on Internet Technology, 5 (2005), pp. 92–128. 20

[13] S. BOCCALETTI, V. LATORA, Y. MORENO, M. CHAVEZ, AND D.-U. HWANG, *Complex networks: structure and dynamics*, Phys. Rep., 424 (2006), pp. 175–308. 3, 4

[14] P. BOLDI, *TotalRank: Ranking Without Damping*, in Proceedings of WWW'05, Special interest tracks and posters of the 14th international conference on World Wide Web, Association for Computing Machinery, New York, NY (2005), pp. 898–899. 23

[15] B. BOLLOBÁS, *Modern Graph Theory*, Graduate Texts in Mathematics, Springer, Verlag, New York, 1998. 6

[16] P. BONACICH, *Power and centrality: a family of measures*, Amer. J. Sociology, 92 (1987), pp. 1170–1182. 4, 16, 32, 34

[17] P. BONACICH AND P. LLOYD, *Eigenvector-like measures of centrality for asymmetric relations*, Social Networks, 23 (2001), pp. 191–201. 16, 32

[18] F. BONCHI, P. ESFANDIAR, D. F. GLEICH, C. GREIF, AND L. V. S. LAKSHMANAN, *Fast matrix computations for pair-wise and column-wise commute times and Katz scores*, Internet Math., 8 (2012), pp. 73–112. 25, 98, 107, 130

[19] S. P. BORGATTI AND M. G. EVERETT, *A graph-theoretic perspective on centrality*, Social Networks, 28 (2006), pp. 466–484. 16, 32

[20] A. BORODIN, G. O. ROBERTS, J. S. ROSENTHAL, AND P. TSAPARAS, *Finding authorities and hubs from link structures on the World Wide Web*, Proceedings WWW10 (2001), ACM 1-58113-348-0/01/0005. 20, 117

[21] U. BRANDES AND T. ERLEBACH, eds., *Network Analysis: Methodological Foundations*, Lecture Notes in Computer Science Vol. 3418, Springer, New York, 2005. 3, 4, 107

[22] U. BRANDES, *On variants of shortest-path betweenness centrality and their generic computation*, Social Networks 30 (2008), pp. 136–145. 17

[23] S. BRIN AND L. PAGE, *The anatomy of a large-scale hypertextual Web search algorithm*, Comput. Net. ISDN Syst, 33 (1998), pp. 107–117. 20

[24] G. CALDARELLI, *Scale-free Networks*, Oxford University Press, Oxford, UK, 2007. 3, 14

[25] T. P. CHARTIER, E. KREUTZER, A. N. LANGVILLE, AND K. E. PEDINGS, *Sensitivity and stability of ranking vectors*, SIAM J. Sci. Comput., 33 (2011), pp. 1077–1102. 65

[26] A. CLAUSET, M. E. J. NEWMAN, AND C. MOORE, *Finding community structure in very large networks*, Physical Review E, 70 (2004), 066111. 14

[27] P. CONSTANTINE, D. GLEICH *Random Alpha PageRank*, Internet Mathematics, 6 (2009), pp. 199–236. 23

[28] M. C. CROFOOT, D. I. RUBENSTEIN, A. S. MAIYA, AND T. Y. BERGER-WOLF, *Aggression, grooming and group-level cooperation in white-faced capuchins (*Cebus capucinus*): Insights from social networks*, Amer. J. Primatology, 73 (2011), pp. 821–833. 3, 110

[29] J. J. CROFTS, E. ESTRADA, D. J. HIGHAM, AND A. TAYLOR, *Mapping directed networks*, Electr. Trans. Numer. Anal., 37 (2010), pp. 337–350. 101

[30] D. CVETKOVIĆ, P. ROWLINSON, S. SIMIĆ, *Eigenspaces of Graphs*, Cambridge University Press, Cambridge, 1997. 11, 16, 69

[31] T. A. DAVIS AND Y. HU, *The University of Florida Sparse Matrix Collection*, ACM Trans. Math. Soft., 38(1) (2011), Article 1. 44, 86, 117

[32] R. DIESTEL, *Graph Theory*, Springer-Verlag, Berlin, 2000. 6, 9

[33] C. H. Q. DING, H. ZHA, X. HE, P. HUSBANDS, AND H. D. SIMON, *Link analysis: hubs and authorities on the World Wide Web*, SIAM Rev., 46 (2004), pp. 256–268. 20

[34] M. EIERMANN AND O. G. ERNST, *A restarted Krylov subspace method for the evaluation of matrix functions*, SIAM J. Numer. Anal., 44 (2006), pp. 2481–2504. 29

[35] P. ERDÖS AND A. RÉNYI, *On Random Graphs. I*, Publicationes Mathematicae, 6 (1959), pp. 290–297. 14

[36] P. ERDÖS AND A. RÉNYI, *On the evolution of random graphs*, Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5 (1960), pp. 17–61. 14

[37] E. ESTRADA, Personal communication, February 2013. 51

[38] E. ESTRADA, *About the discriminant power of the subgraph centrality and other centrality measures (Working paper)*, arXiv:1305.6836. 72

[39] E. ESTRADA, *Communicability in temporal networks*, Phys. Rev. E 88 (2013), 042811. 83

[40] E. ESTRADA, *The Structure of Complex Networks*, Oxford University Press, Oxford, UK, 2011. 3, 4, 14, 16, 31, 32, 69

[41] E. ESTRADA, M. FOX, D. HIGHAM, AND G.-L. OPPO, eds., *Network Science. Complexity in Nature and Technology*, Springer, New York, 2010. 3

[42] E. ESTRADA, *Virtual identification of essential proteins within the protein interaction network of yeast*, Proteomics, 6 (2006), pp.35–40. 3, 14, 51, 131

[43] E. ESTRADA *Protein bipartivity and essentiality in the yeast Protein-Protein Interaction network*, J. Proteome Res., 5 (2006), pp. 2177–2184. 3, 14, 51, 131

[44] E. ESTRADA, J. A. DE LA PEÑA, AND N. HATANO, *Walk entropies in graphs*, Linear Algebra Appl., 443 (2014), pp. 235–244. 72

[45] E. ESTRADA AND N. HATANO, *Communicability in complex networks*, Phys. Rev. E, 77 (2008), article 036111. 17, 18, 95

[46] E. ESTRADA, N. HATANO, AND M. BENZI, *The physics of communicability in complex networks*, Phys. Reports, 514 (2012), pp. 89-119. 17, 18, 29, 31, 36, 107

[47] E. ESTRADA AND D. J. HIGHAM, *Network properties revealed through matrix functions*, SIAM Rev., 52 (2010), pp. 671–696. 9, 17, 18, 31, 95, 97, 100, 107

[48] E. ESTRADA AND J. A. RODRÍGUEZ-VELÁZQUEZ, *Subgraph centrality in complex networks*, Phys. Rev. E, 71 (2005), 056103. 17, 31, 32

[49] L. EULER, *Solutio problematis ad geometriam situs pertinentis*, Commentarii Academiae Scientiarum Imperialis Petropolitanae, 8 (1736) pp. 128–140. 6

[50] R. FAGIN, R. KUMAR, AND D. SIVAKUMAR, *Comparing top k lists*, SIAM J. Discr. Math., 17 (2003), pp. 134–160. 23

[51] M. FALOUTSOS, P. FALOUTSOS, AND C. FALOUTSOS, *On Power-Law Relationships of the Internet Topology*, In SIGCOMM (1999), pp. 252–262. 14

[52] A. FARAHAT, T. LOFARO, J. C. MILLER, G. RAE, AND L. A. WARD, *Authority rankings from HITS, PageRank, and SALSA: existence, uniqueness, and effect of initialization*, SIAM J. Sci. Comput., 27 (2006), pp. 1181–1201. 19, 97, 108, 109, 114

[53] C. FENU, D. MARTIN, L. REICHEL, AND G. RODRIGUEZ, *Network Analysis via Partial Spectral Factorization and Gauss Quadrature*, SIAM J. Sci. Comput., 35 (2013), pp. A2046–A2068. 28, 130

[54] D. FOGARAS, *Where to start browsing the Web?*, in T. Böhme, G. Heyer, and H. Unger (Eds.), *Proceedings IICS 2003*, Lecture Notes in Computer Science, 2877 (2003), pp. 65–79. 23, 109

[55] M. FRANCESCHET, *PageRank: Standing on the shoulders of giants*, Comm. ACM, 54 (2011), pp. 92–101. 20, 23, 107

[56] L. C. FREEMAN, *A set of measures of centrality based on betweenness*, Sociometry, 40 (1977), pp. 35–41. 16

[57] L. C. FREEMAN, *Centrality in networks : I. Conceptual clarification*, Social Networks, 1 (1979), pp. 215–239. 16

[58] D. GLEICH, P. GLYNN, G. GOLUB, AND C. GREIF, *Three results on the PageRank vector: Eigenstructure, sensitivity and the derivative*, in *Web Information Retrieval and Linear Algebra Algorithms*, eds. A. FROMMER, M. W. MAHONEY, AND D. B. SZYLD, Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2007. 23

[59] K. I. GOH, B. KAHNG, AND D. KIM, *Spectra and eigenvectors of scale-free networks*, Phys. Rev. E 64 (2001), 051903. 14

[60] G. H. GOLUB AND G. MEURANT, *Matrices, Moments and Quadrature with Applications*, Princeton University Press, Princeton, NJ, 2010. 25, 26, 27, 35

[61] G. H. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins University press, 3rd edition, Baltimore and London, 1996. 30

[62] P. GRINDROD AND D. HIGHAM, *A matrix iteration for dynamic network summaries*, SIAM Review, 55 (2013), pp. 118–128. 32, 65, 66

[63] S. GÜTTEL, *funm_kryl toolbox for MATLAB*, www.mathe.tu-freiberg.de/∼guettels/funm_kryl/. 29

[64] N. J. HIGHAM, *The scaling and squaring method for the matrix exponential revisited*, SIAM J. Matrix Anal. Applic., 26 (2006), pp. 1179–1193. 24

[65] N. J. HIGHAM, *Functions of Matrices. Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008. 12, 24, 28, 30, 109

[66] I. C. F. IPSEN AND R. S. WILLS *Mathematical Properties and Analysis of Google's PageRank* Bol. Soc. Esp. Mat. Apl., 34 (2006), pp. 191–196. 23

[67] L. KATZ, *A new status index derived from sociometric data analysis*, Psychometrika, 18 (1953), pp. 39–43. 3, 4, 17, 32, 107

[68] J. KLEINBERG, *Authoritative sources in a hyperlinked environment*, J. ACM, 46 (1999), pp. 604–632. 4, 18, 83, 95, 114

[69] T. G. KOLDA, A. PINAR, T. PLANTENGA, AND C. SESHADHRI *A Scalable Generative Graph Model with Community Structure*, arXiv:1302.6636 [cs.SI], March 2013. 15

[70] G. KOLLIAS AND E. GALLOPOULOS, *Functional Rankings with Multidamping: Generalizing PageRank with Inhomogeneous Matrix Products*, Tech. Rep. TR HPCLAB-SCG 01/09-11, University of Patras, Greece, 2011. 20

[71] P. L. KRAPIVSKY, G. J. RODGERS, AND S. REDNER, *Degree distributions of growing networks*, Phys. Rev. Lett., 86 (2001), pp. 5401–5404. 15

[72] A. N. LANGVILLE AND C. D. MEYER, *A survey of eigenvector methods for Web information retrieval*, SIAM Rev., 47 (2005), pp. 135–161. 4, 20

[73] A. N. LANGVILLE AND C. D. MEYER, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, NJ, 2006. 4, 23, 55, 96, 135, 136

[74] A. N. LANGVILLE AND C. D. MEYER, *Who's #1? The Science of Rating and Ranking*, Princeton University Press, Princeton, NJ, 2012. 4, 23, 37

[75] R. LEMPEL AND S. MORAN, *The stochastic approach for link-structure analysis (SALSA) and the TKC effect*, in Proceedings of the Ninth International Conference on the World Wide Web, 2000, pp. 387–401. 4

[76] G. MEURANT, *The Lanczos and Conjugate Gradient Algorithms: From Theory to Finite Precision Computations*, SIAM, Philadelphia, Pa, 2006. 27, 137

[77] G. MEURANT, *MMQ toolbox for MATLAB*, http://pagesperso-orange.fr/gerard.meurant/. 25, 128

[78] C. D. MEYER, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000. 9, 10, 12, 21, 30

[79] M. MIHAIL AND C. PAPADIMITRIOU, *On the eigenvalue power law*, in J. D. P. Rolim and S. Vadhan (Eds.), *Proceedings of RANDOM 2002*, Lectures Notes in Computer Science, 2483 (2002), pp. 254–262. 23, 110

[80] L. MIN AND Z. QINGGUI, *Degree Distribution of a Mixed Attachments Model for Evolving Networks*, Proceedings of WRCS 2009, pp. 815–822. 14

[81] J. L. MORRISON, R. BREITLING, D. J. HIGHAM, AND D. R. GILBERT, *GeneRank: Using search engine technology for the analysis of microarray experiments*, BMC Bioinformatics, 6:233 (2005). 20

[82] M. E.J. NEWMAN, *The structure of scientific collaboration networks*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 404-409. 15

[83] M. E. J. NEWMAN, *The structure and function of complex networks*, SIAM Rev., 45 (2003), pp. 167–256. 14

[84] M. E. J. NEWMAN, *Networks: An Introduction*, Cambridge University Press, Cambridge, UK, 2010. 3, 4, 14, 16, 29

[85] M. E. J. NEWMAN, A. L. BARABÁSI, AND D. J. WATTS, *The Structure and Dynamics of Networks*, Princeton University Press, Princeton, NJ, 2003. 3, 4

[86] G. POOLE AND T. BOULLION, *A survey on M-matrices*, SIAM Review, 16 (1974), pp. 419–427. 11

[87] M. PUCK ROMBACH AND M. PORTER, *Discriminating power of centrality measures*, arXiv:1305.3146. 72

[88] B. SAVAS AND I. DHILLON, *Clustered low rank approximation of graphs in information science applications*, Proceedings of the 2011 SIAM Conference on Data Mining, April 2011. 3, 130

[89] D. STEVANOVIĆ, *Comment on "Subgraph centrality in complex networks"*, Phys. Rev. E, 88 (2013), 026801. 72

[90] G. STRANG, *Introduction to Linear Algebra, 3rd Edition*, Wellesley-Cambridge Press, Cambridge, Ma, 2003. 9

[91] A. TAYLOR AND D. J. HIGHAM, *CONTEST: Toolbox files and documentation*, `http://www.mathstat.strath.ac.uk/research/groups/numerical_analysis/contest/toolbox`. 38

[92] A. TAYLOR AND D. J. HIGHAM, *CONTEST: A Controllable Test Matrix Toolbox for MATLAB*, ACM Trans. Math. Software, 35 (2009), pp. 26:1–26:17. 38

[93] P. TSAPARAS, *Datasets for Experiments on Link Analysis Ranking Algorithms*, available at http://www.cs.toronto.edu/∼tsap/experiments/datasets/ index.html (last accessed June 2012). 117

[94] S. VIGNA, *Spectral Ranking*, arXiv:0912.0238v9 [cs.IR], 2 March 2011. 16, 107

[95] D. J. WATTS AND S. H. STROGATZ, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440–442. 15, 40

[96] G. WU AND J. LI, *SWRank: an approach for ranking semantic web reversely and consistently*, in Third International Conference on Semantics, Knowledge and Grid, IEEE, 2007. DOI: 10.1109/SKG.2007.81 20, 110

[97] G. WU, J. LI, L. FENG, AND K. WANG, *Identifying potentially important concepts and relations in an ontology*, in A. Sheth et al. (Eds.), Proceedings of ISWC 2008, Lecture Notes in Computer Science, 5318 (2008), pp. 33–49. 110

[98] W. W. ZACHARY, *An information flow model for conflict and fission in small groups*, Journal of Anthropological Research 33 (1977), pp. 452–473. 44, 46

[99] L. ZOU, W. PEI, T. LI, Z. HE, AND Y. CHEUNG, *Topological fractal networks introduced by mixed degree distribution*, Physica A, 380 (2007), pp. 592–600. 14