CONTEST: A Controllable Test Matrix Toolbox for MATLAB

ALAN TAYLOR and DESMOND J. HIGHAM
University of Strathclyde

Large, sparse networks that describe complex interactions are a common feature across a number of disciplines, giving rise to many challenging matrix computational tasks. Several random graph models have been proposed that capture key properties of real-life networks. These models provide realistic, parametrized matrices for testing linear system and eigenvalue solvers. CONTEST (CONtrollable TEST matrices) is a random network toolbox for MATLAB that implements nine models. The models produce unweighted directed or undirected graphs; that is, symmetric or unsymmetric matrices with elements equal to zero or one. They have one or more parameters that affect features such as sparsity and characteristic pathlength and all can be of arbitrary dimension. Utility functions are supplied for rewiring, adding extra shortcuts and subsampling in order to create further classes of networks. Other utilities convert the adjacency matrices into real-valued coefficient matrices for naturally arising computational tasks that reduce to sparse linear system and eigenvalue problems.

Categories and Subject Descriptors: G.1.3 [**Numerical Analysis**]: Numerical Linear Algebra—*Sparse; structured; and very large systems (direct and iterative methods)*; G.4 [**Mathematics of Computing**]: Mathematical Software—*Certification and testing*

General Terms: Algorithms, Experimentation, Performance, Reliability

Additional Key Words and Phrases: clustering, matrix computation, preferential attachment, random graph, rewiring, sparse matrix, small-world

ACM Reference Format:
Taylor, A. and Higham, D. J. 2009. CONTEST: A controllable test matrix toolbox for MATLAB. ACM Trans. Math. Softw. 35, 4, Article 26 (February 2009), 17 pages.
DOI = 10.1145/1462173.1462175. http://doi.acm.org/10.1145/1462173.1462175.

D. J. Higham was supported by Engineering and Physical Sciences Research Council grants GR/S62383/01 and EP/E049370/1.
Authors' address: A. Taylor and D. J. Higham, Department of Mathematics, University of Strathclyde, Glasgow, G1 1XH, UK.

1. MOTIVATION

Networks describing connectivity structures arise across a vast range of application areas. Examples where it has proved useful to record data include interactions between genes [Kauffman 1969], proteins [de Silva and Stumpf 2005], cortical regions [Kamper et al. 2002; Sporns and Zwi 2004], Internet nodes [Faloutsos et al. 1999], Web pages [Broder et al. 2000; Page et al. 1998], countries [Fagiolo 2007], co-authors [Newman 2004], telephones [Abello et al. 1998], assets on the stock market [Boginski et al. 2003], and members of various populations [Conyon and Muldoon 2006; Kiss et al. 2006; Onody and de Castro 2004; Porter et al. 2005; Williams et al. 2002].

Typical data mining and visualization tasks reduce to linear system or eigenvalue computations with the large, sparse adjacency matrices that define the interactions. Several random graph models, that is, formulas for probabilistically inserting connections, have been derived that attempt to capture the key topological properties of real-life networks. Important goals for such work are to understand how a network has reached its current state and to predict how it will evolve. From a numerical analysis perspective, these random graph models are an extremely useful source of realistic, controllable test matrices for linear algebra software. This provides the motivation for the MATLAB toolbox CONTEST (CONtrollable TEST matrices),[1] which implements nine popular random network models, along with various utility functions for postprocessing the networks. The codes were developed and tested under MATLAB version 7.4.0.287 (R2007a). As supplementary material at the Web site, we record performance results for MATLAB's built-in iterative linear system solvers `pcg`, `qmr`, `symmlq`, `lsqr`, `minres`, `cgs`, `gmres`, `bicg`, and `bicgstab` using test matrices from the toolbox.

This article is arranged as follows. Section 2 gives a very brief overview of the historical development of random network models. In Section 3 we describe each of the nine models and the corresponding MATLAB code. Section 4 introduces the utility functions for altering existing networks, setting up coefficient matrices arising in common tasks and checking some basic topological properties. In Section 5 we give a very brief illustration of the toolbox in use, and we summarize the aims of this work in Section 6.

Our notation is as follows. We let n denote the number of nodes in a network, with $a_{ij} = a_{ji} = 1$ if nodes i and j are connected and $a_{ij} = a_{ji} = 0$ otherwise. So the adjacency matrix $A \in \mathbb{R}^{n\times n}$ is symmetric. We always have $a_{ii} = 0$; so nodes cannot be self-connected. The *degree* of node i is found by counting its neighbors, $\deg_i := \sum_{j=1}^{n} a_{ij}$. For $\deg_i > 1$ the *curvature* or *clustering coefficient* of node i is found by counting how many pairs of these neighbors are themselves connected, and dividing this number by the maximum possible number of connections, $\deg_i(\deg_i - 1)/2$. A definition in terms of MATLAB comands is given in Section 4.7.1.

A call to one of the random network functions in the toolbox will generate an $A \in \mathbb{R}^{n\times n}$ as an independent instance drawn from a random network model.

[1]CONTEST is available from the Web site
http://www.maths.strath.ac.uk/research/groups/numerical_analysis/contest.

The randomness is driven entirely by MATLAB's built in pseudorandom number generators, `rand` and `randn`, and our codes do not alter their states. So the user can get back the same matrix by resetting the states of these two random number generators. For consistency, we always generate adjacency matrices with the `sparse` attribute, even though for some parameter values a full matrix may arise (e.g., with the extreme choice of $p = 1$ in the Gilbert model of Section 3.1).

Although we produce only symmetric adjacency matrices, it is straightforward to create unsymmetric versions, corresponding to directed networks, by combining the upper and lower triangles from two independent samples from the same model. For example, calling `A = erdrey(n,m)` and `B = erdrey(n,m)`, where `erdrey` described in Section 3.1.1 implements the Erdős–Rényi model, we could set `C = triu(A) + tril(B)`.

2. BACKGROUND

It has been repeatedly observed that real connectivity networks are neither completely regular lattices nor classical random graphs. Following the landmark paper of Watts and Strogatz [1998], there has been a resurgence of interest in the idea of designing probabilistic models that capture important topological properties of real networks. Watts and Strogatz coined the phrase *small world network* to describe a regime where small pathlengths coexist with large clustering coefficients (nodes tend to live in cliquey, well-connected subgraphs and yet the network can be globally traversed with relatively few links). They also showed that this pair of properties arise when an appropriate amount of disorder is added to a regular lattice.

Another key property that is claimed to be common in real networks is a *scale-free degree distribution*,

$$\frac{\text{Number of nodes of degree } k}{n} \propto k^{-\gamma}, \tag{1}$$

where γ is a constant, typically in the range $2 \leq \gamma \leq 3$. The *preferential attachment* model of Barabási and Albert [1999] attempts to describe the way a network might grow when new nodes are added and new connections formed, and it produces scale-free degree distributions. More recently, however, the prevalence of the scale-free property has been questioned, at least in the context of biological networks [Khanin and Wit 2006; Pržulj et al. 2004; Stumpf et al. 2005].

In addition to small worlds and scale-freeness, a third dominant concept is that of *motifs* [Alon 2006; Milo et al. 2004]. A motif is a subgraph that is significantly overrepresented (relative to the occurrence of that subgraph in a "randomized" version of the network). These motifs may be regarded as the basic building blocks of the networks, and hence understanding their roles gives valuable insights into how the overall network operates [Mangan and Alon 2003; Mangan et al. 2003]. The closely related idea of *graphlet frequency* was introduced in Pržulj et al. [2004] as a means to compare networks and further developed in Pržulj et al. [2006]. Two networks are close if they are made up of building blocks in the same relative proportions. This gives a

powerful and comprehensive means to check whether a probabilistic model is capturing topological properties of real networks and to decide which models are most appropriate. Using these ideas, the software tool GraphCrunch for network comparison was developed in Milenkovic et al. [2008].

Overall, a recent and rapid expansion in theoretical and empirical research activity has produced several models for computing networks in a controlled manner that are "close" to real-life networks in a well-defined sense. It is our tenet that these computable networks are therefore excellent candidates for test matrices.

Although well-established sparse matrix test sets exist [Boisvert et al. 1997; Davis 2007; Duff et al. 1989], they have been built around fixed instances arising in particular application areas. Randomness is typically incorporated very simplistically. For example, Matrix Market[2] [Boisvert et al. 1997] makes available the random generators `DLATMR/ZLATMR` from LAPACK [Anderson et al. 1999], which independently assign random samples from a given distribution across the entries of an array and then randomly reset elements to zero in order to achieve a given level of sparsity. In Davis [2007], Davis argues that "random sparse matrices" are not appropriate for testing sparse matrix algorithms; however, those comments would appear to be aimed at different classes of matrices to those considered here. The models implemented in CONTEST use randomness to capture properties that are commonly observed in complex interaction networks.

The code in CONTEST was written to exploit vectorization and to use matrix-vector-level operations where possible, but ultimately our priority was to allow sparse matrices of the largest possible dimension to be computed. A secondary aim was produce short, readable, and maintainable programs. The importance of memory allocation and usage when generating sparse matrices in MATLAB is discussed in Gilbert et al. [1992] and in NA Digest.[3] Our justification for not focusing on execution time is that the tasks that will typically be performed with the matrices—eigensolves, linear systems solves, factorizations—will usually be more computationally expensive than the matrix generation phase.

3. MODELS

In this section, we give brief descriptions of the nine models implemented, and show how to use the corresponding MATLAB functions. In each case, the output argument `A` is a sparse, symmetric, zero-diagonal matrix of dimension `n`, with `n` being the first of the input arguments. The remaining input arguments take default values if not specified in the function call. Default parameters have been chosen to ensure that `A` corresponds to a connected (irreducible) graph with high probability, with the exception of `sticky` in Section 3.7.1, which, by construction, may produce many small disconnected subgraphs. In Figure 1 we show a spy plot for each of the nine models using `n=100`; this

[2]with Web site URL http://math.nist.gov/MatrixMarket/.
[3]at http://www.netlib.org/na-digest-html/07/v07n28.html#1.

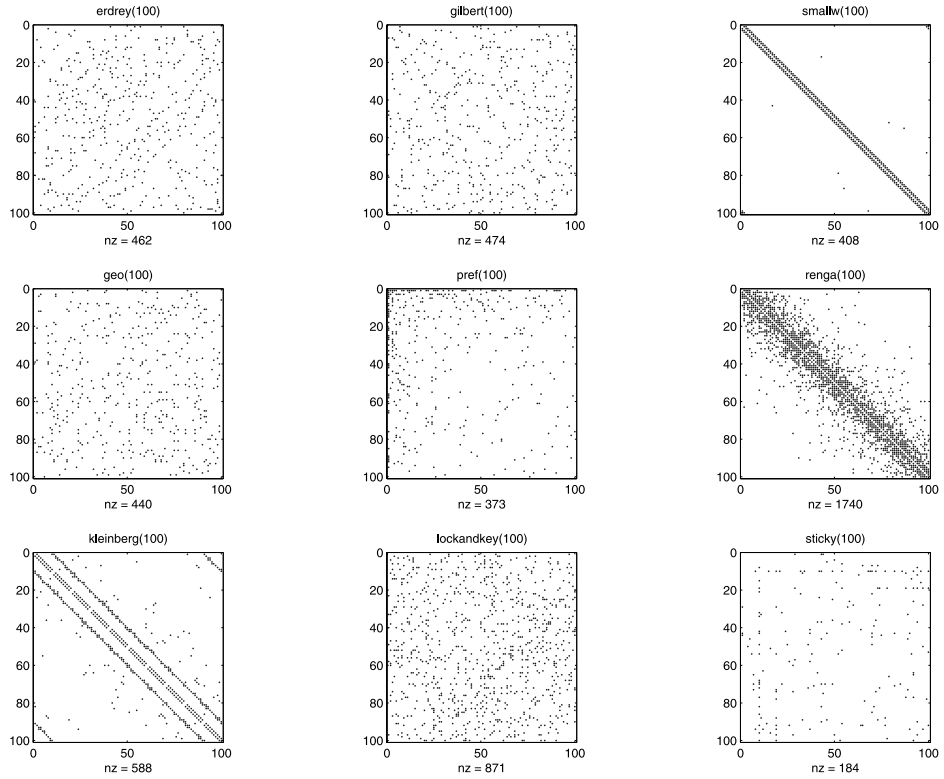

Fig. 1. Spy plots showing nonzero patterns for a 100×100 sample from each of the nine models.

dimension was chosen to make the visualization clearer; in practice values of n of the order 10^4 or higher would be more realistic.

3.1 Classical

Random graph theory began in earnest in the late 1950's, with the two classical models in Gilbert [1959] and Erdös and Rényi [1959]. These models are usually referred to as $\mathcal{G}(n, p)$ and $\mathcal{G}(n, m)$, but to help distinguish between them we will use the names Gilbert and Erdős-Rényi.

In Gilbert's model [Gilbert 1959] a fixed probability p is specified, and then each pair of nodes is, independently, connected with probability p. In the Erdős-Rényi model [Erdös and Rényi 1959] the number m of edges in the network is specified. (Of course, m must be no more than the maximum possible number of edges, $n(n-1)/2$.) We then select uniformly at random from the set of all graphs containing n nodes and m edges.

The properties of these classical random graphs have been well studied [Albert and Barabási 2002; Bollobás 1985], although in terms of currently adopted measures, such as pathlengths, clustering coefficients, and graphlet frequencies, they cannot be regarded as accurate models of realistic networks [de Silva and Stumpf 2005; Pržulj et al. 2004; Watts and Strogatz 1998]. Our

implementation for the Gilbert class is taken from Batagelj and Brandes [2005, Algorithm 1].

3.1.1 *Classical Codes: `gilbert` and `erdrey`.* The function `gilbert(n,p)` returns an instance from the Gilbert class. The optional second input argument defaults to $\log(n)/n$, so `A = gilbert(n)` is equivalent to `A = gilbert (n,log(n)/n)`. Similarly, `A = erdrey(n,m)` produces an Erdős-Rényi random graph, with m defaulting to the smallest integer bigger than $n\log(n)/2$.

3.2 Small World

Motivated by the small world concept of the experimental psychologist Stanley Milgram [1967], Watts and Strogatz [1998] proposed a random graph model that can be regarded as interpolating between a regular, periodic lattice and a classical random graph. Although the original work used rewiring, it is now more common to introduce randomness via the addition of shortcuts [Higham and Higham 2000; Newman et al. 2000]. Hence, in our Watts-Strogatz model we begin with a k-nearest-neighbor ring (nodes i and j are connected if and only if $|i-j| \le k$ or $|n-|i-j|| \le k$). Then, each node is considered independently in turn. With fixed probability p a node is given an extra link (i.e., a shortcut) connecting it to a node chosen uniformly at random across the network. (At the end of this process, self-links and repeated links between nodes are removed.)

3.2.1 *Small World Code: `smallw`.* The function `smallw` returns an instance of the Watts-Strogatz model, with syntax according to `A = smallw(n,k,p)`. The optional input arguments `k` and `p` default to 2 and 0.1, respectively. From a linear algebra perspective, the adjacency matrix has a symmetric, banded Toeplitz structure, with extra nonzeros added uniformly and symmetrically at random. We note that `smallw` makes use of the utility function `short` that is described in Section 4.2.1.

3.3 Geometric

A two-dimensional, nonperiodic, *geometric random graph* may be defined as follows. First, each of the n nodes is placed at random in the unit square: More precisely, the ith node is given coordinates (x_i, y_i), where $\{x_i, y_i\}_{i=1}^n$ are independent and identically distributed with uniform (0,1) distribution. Next, for some specified radius r, nodes i and j are connected if and only if $(x_i - x_j)^2 + (y_i - y_j)^2 \le r^2$. In words, an edge denotes that two nodes were placed no more than Euclidean distance r apart. Figure 2 illustrates the process with $n = 100$ and $r = 0.2$.

We emphasize that the resulting graph is simply the usual list of nodes and edges. Information about the precise locations $\{x_i, y_i\}_{i=1}^n$ is not part of the final mathematical object. Natural generalizations are possible.

—*Dimension.* The nodes can be randomly assigned to locations in the unit cube in $\mathbb{R}^m$, for some $m > 2$.

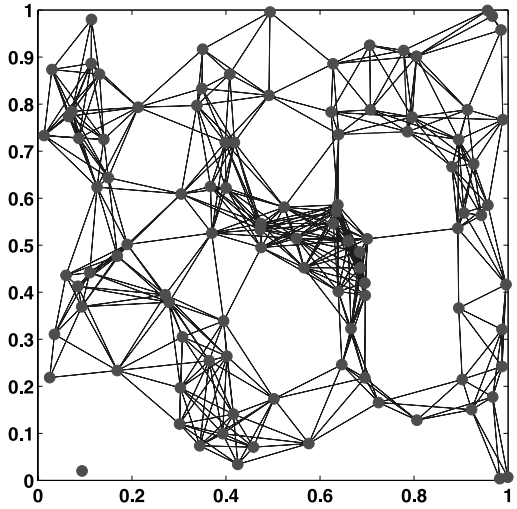

Fig. 2. Construction of a geometric random graph. Here, $n = 100$ and $r = 0.2$.

—*Periodicity.* Distance can be measured in a wrap-around fashion, so that, for example, in the unit square, $(x_i - x_j)^2 + (y_i - y_j)^2$ is replaced by

$$\left(\min\left(\left|x_i - x_j\right|, 1 - \left|x_i - x_j\right|\right)\right)^2 + \left(\min\left(\left|y_i - y_j\right|, 1 - \left|y_i - y_j\right|\right)\right)^2.$$

—*Norm.* The Euclidean norm can be replaced by any other vector norm.

Much theory is available concerning properties of geometric random graphs; see Penrose [2003] for a comprehensive treatment. Recently Pržulj et al. [2004] showed that two- and three-dimensional nonperiodic versions, using the Euclidean norm, give surprisingly accurate reproductions of many features of real biological networks, and an algorithm that tests for geometric structure is developed in Higham et al. [2008].

3.3.1 *Geometric Code: `geo`.* The call `A=geo(n,r,m,per,pnorm)` returns an instance of a geometric random graph. There are four optional input arguments.

—`r` specifies the radius, defaulting to $\sqrt{1.44/n}$, which is motivated by the asypmtotic ($n \to \infty$) level that guarantees connectivity in two dimensions [Penrose 2003];

—`m` specifies the dimension, defaulting to 2;

—`per` is a logical variable specifying whether periodic distance is to be used, defaulting to `per = 0`; not periodic;

—`pnorm` specifies the L_p-norm to be used, defaulting to 2.

3.4 Preferential Attachment

Barabási and Albert [1999] used the concept of preferential attachment to develop random graphs with scale-free degree distributions. In this model, the

network grows (new nodes are added and linked into the existing network) until n nodes have been created. For some fixed integer $d \geq 1$, each new node is given d links on arrival. These new connections are not chosen uniformly; the new node links to an existing node with a probability that is proportional to the current degree of that node. In this way, well-connected nodes tend to become even better connected (the rich get richer) as the network evolves. Our precise model is a translation into MATLAB of Batagelj and Brandes [2005, Algorithm 5], which uses the specification in Bollobás et al. [2001].

3.4.1 *Preferential Attachment Code: `pref`.* The call `A = pref(n,d)` returns an instance of a preferential attachment graph, using a single node as the initial network. The degree parameter `d` defaults to 2.

3.5 Range Dependent

3.5.1 *RENGA.* Yeast two hybrid Protein-Protein Interaction (PPI) networks have proteins as nodes. Two nodes share an undirected edge if they have been experimentally observed to interact [Xenarios et al. 2002]. Motivated by the structure of PPI networks, Grindrod [2002] proposed and analyzed a random graph model that, in a sense, generalizes Watts-Strogatz. In this model, the nodes have a natural linear ordering, $i = 1, 2, \ldots, n$. Independently over all pairs of nodes, we then insert a link between nodes i and j with probability $\alpha\lambda^{|j-i|-1}$, where $\alpha > 0$ and $\lambda \in (0, 1)$ are fixed parameters. The choice $\alpha = 1$ ensures that adjacently ordered nodes are always connected. The geometric factor $\lambda^{|j-i|-1}$ causes long-range edges to be less common than short-range edges.

Further analysis and generalizations of this model, now referred to as RENGA, appear in Grindrod et al. [2008] and Higham [2005; 2003]. Closely related models have also been used in percolation theory [Grimmett 1999].

3.5.2 *RENGA Code: `renga`.* The call `A = renga(n,lambda,alpha)` returns an instance of a RENGA, with `lambda` defaulting to 0.9 and `alpha` defaulting to 1.

3.5.3 *Kleinberg.* Kleinberg [2000] defined a variation of the Watts-Strogatz model, and used it to examine which types of navigation algorithm can exploit the existence of shortcuts. Kleinberg's model is based on a periodic, two-dimensional lattice: The $n = m^2$ nodes can be thought of as being equally spaced throughout a square, with each node having a location of the form $(i, j) \in \mathbb{R}^2$, where the integers i and j run from 1 to m. Every node is given short-range connections to its neighbors that are a lattice (Manhattan) distance of at most p away. Then each node is given q further long-range connections. For a given node u, the recipient v of each such long-range connection is chosen independently at random, with probability proportional to $r^{-\alpha}$. Here, r is the lattice distance between u and v and $\alpha \geq 0$ is a fixed parameter.

3.5.4 *Kleinberg Code: `kleinberg`.* The call `A = kleinberg(n,p,q,alpha)` generates an instance of the Kleinberg model. If the input dimension `n` is not

a perfect square then the output matrix has dimension `(round(sqrt(n)))^2`. Default values are `p = 1`, `q = 1`, and `alpha = 2`.

3.6 Lock and Key

Using some basic biological insights, Thomas et al. [2003] proposed a class of random graphs that model PPI networks. This class of models was further analysed in Morrison et al. [2006], where it was used to extract new biological information from real PPI datasets. The underlying modeling idea is that two proteins interact because they share physically matching parts, which, following Morrison et al. [2006], we refer to as *locks* and *keys*. There will be several different types of key, which we can think of as labeled by colors (red, green, blue, etc.) and for each type of key there is a matching lock (red, green, blue, etc.). In the model, each protein has the same chance of possessing each color of lock and each color of key. More precisely, for a given number of colors m, we take each node in turn and independently assign it each possible lock and key with some fixed probability p. The graph is then generated according to the rule that two nodes share an edge if and only if one possesses a key and the other possesses a lock of the same color. Self-links are removed.

3.6.1 *Lock and Key Code: `lockandkey`.* The call `A = lockandkey(n,m,p)` returns an instance of a lock-and-key graph where there are `m` different lock-and-key colors and each type of lock and key is handed out independently with fixed probability `p`. Default values are `m = ceil(n*log(n))` and `p = 1/n`.

3.7 Stickiness

The stickiness model was introduced in Pržulj and Higham [2006] to model PPI networks. It was motivated as a simplified version of the lock-and-key framework in which parameters could be fitted to real data. Here, a nonnegative vector $\widehat{d} \in \mathbb{R}^n$ is given, representing the scaled degree distribution of some target network; more precisely, $\widehat{d}_i = \deg_i / \sqrt{\sum_{j=1}^{n} \deg_i}$, where $\deg_i$ is the degree of the ith node in the target. Then a new random network is produced by connecting nodes i and j with probability $\widehat{d}_i \widehat{d}_j$. In this way the *expected degrees* in the random model match the target degrees. This model was found more accurate than previously proposed models at reproducing topological properties of PPI networks.

3.7.1 *Stickiness Code: `sticky`.* The call `A = sticky(deg)` generates an instance of a stickiness graph with expected degree distribution given by the one-dimensional array `deg`. To be consistent with our general philosophy that all models can be called with a single input argument, `n`, representing the dimension, we allow an exception where `sticky` is called as `A = sticky(n)`, with `n` a positive integer. In this case `A` will be an instance of a stickiness graph of dimension `n` with a scale-free expected degree distribution of the form (1) with $\gamma = 2.5$. It is also possible to specify two input parameters: A call `A = sticky(n,gamma)` specifies the value of γ to be used in (1).

4. UTILITY FUNCTIONS

4.1 Rewiring

The Watts-Strogatz model [Watts and Strogatz 1998] added randomness to a ring network by *rewiring* some edges. For a general undirected network, we define a rewiring process as follows, in terms of a fixed parameter p. Each entry in the lower triangle of the original adjacency matrix is examined in turn. If $a_{ij} \neq 0$ then, independently with probability p, we reset $a_{ij} = a_{ji} = 0$, choose a node k uniformly at random from all nonneighbors of node i, and set $a_{ik} = a_{ki} = 1$.

4.1.1 *Rewiring Code: `rewire`.* The call `R = rewire(A,p)` takes an adjacency matrix `A` and returns a rewired adjacency matrix `R`. The rewiring probability `p` defaults to `p = log(n)/n`.

4.2 Shortcuts

Rewiring has the theoretical drawback that it may cause a connected network to become unconnected. Adding *shortcuts* is an alternative procedure that gives very similar topological effects [Newman et al. 2000] but does not degrade connectivity. In this case the parameter p is a fixed probability that is used independently over all nodes. For each node, with probability p we add a new link from that node to a node chosen uniformly at random across the whole network. Self-links are then removed and repeated links treated as single links.

4.2.1 *Shortcut Code: `short`.* The call `S = short(A,p)` takes an adjacency matrix `A`, adds shortcuts, and returns the new adjacency matrix `S`. The shortcut probability `p` defaults to `log(n)/n`.

4.3 Subsampling

Information is often missing from real-life connectivity datasets [de Silva et al. 2006]. These omissions may be caused, for example, by errors in experimental observations (false negatives) or by an inherent restriction on the number or type of observations that can be made. In the case of yeast two hybrid PPI networks, it is widely accepted that the reported network is merely a noisy subset of the underlying "true" network, and we can think of the given network as being generated from a subsampling operation on the larger version [Titz et al. 2004]. Interestingly, it has been discovered that the subsampling operation may dramatically alter the topological properties of a network [de Silva et al. 2006; Han et al. 2005; Salathé et al. 2005].

We have implemented two subsampling algorithms. Given the adjacency matrix for a network, they return the adjacency matrix for a network consisting of a subset of those nodes and edges. The first algorithm does an unbiased, uniform node removal involving a fixed parameter p. Each node is considered in turn, and with independent probability $1 - p$ we remove that node and all edges that involve it, that is, we delete that row and column from the

adjacency matrix. The second algorithm uses a bait-and-prey approach, along the lines of Han et al. [2005], which models the generation of certain PPI datasets. Here, we use two fixed parameters, `bait` and `prey`. A proportion `bait` of the nodes are chosen as baits. Then, for each bait, a proportion `prey` of its edges are recorded, along with the prey nodes that are linked to the bait by these edges. The final subsampled network consists of the bait-prey edges and all the nodes that they involve.

4.3.1 *Supsampling Codes: `unisample` and `baitsample`.* The call `U = unisample(A,p)` takes an adjacency matrix `A` and returns a subnetwork `U` formed from an unbiased, uniform node removal. The probability `p` defaults to `0.5`.

The bait-and-prey algorithm can be called as `B = baitsample (A,bait, prey)`, with defaults `bait = 0.5` and `prey = 0.5`.

4.4 Laplacian Matrices

An undirected network can be characterized by its adjacency matrix, and basic linear algebra tells us that the eigenvectors and eigenvalues of this matrix carry relevant information. However, spectral graph theory [Chung 1997] has shown that it is generally more useful to look at the spectrum of the so-called Laplacian. There are two different matrices that take this name in the literature. We distinguish between them as follows.

—The *graph Laplacian* has the form $D - A$.
—The *normalized graph Laplacian* has the form $\widehat{D}^{-\frac{1}{2}}(D - A)\widehat{D}^{-\frac{1}{2}}$.

Here $D = \text{diag}(\deg_i)$ and $\widehat{D} = D$, with the exception that we take $\widehat{D}_{ii} = 1$ in the case where $\deg_i = 0$.

Clustering and partitioning tasks can be tackled by computing eigenvectors corresponding to small eigenvalues of these matrices. In particular, the *Fiedler vector* and *normalized Fiedler vector* of a connected network are defined to be the eigenvectors corresponding to the second smallest eigenvalues of the Laplacian and normalized Laplacian, respectively. Specific software exists for computing this type of information [Cour et al. 2005; Hendrickson and Leland 1994; Hu and Scott 2003].

4.4.1 *Laplacian Matrix Codes: `lap`.* The call `L = lap(A,nl)` takes a symmetric adjacency matrix `A` and returns a Laplacian; `nl=0` for unnormalized and `nl=1` for normalized. The default is `nl=1`.

4.5 PageRank Matrix

The PageRank algorithm returns a vector whose ith entry indicates the "importance" of the ith node in a network. The algorithm was invented by Page and Brin and forms the heart of the search engine Google [Langville and Meyer 2006; Page et al. 1998]. PageRank was originally designed for the directed network where nodes are Web pages and edges are hypertext links, but

it has also been used on networks in biology [Morrison et al. 2005]. Given an adjacency matrix A, the PageRank vector x solves the linear system

$$Px = \mathbf{1}, \qquad \text{where } P = I - dA^T\widehat{D}^{-1}. \tag{2}$$

Here, $d \in (0, 1)$ is a scalar parameter, the diagonal degree matrix $\widehat{D}$ is defined in Section 4.4, and $\mathbf{1}$ denotes the vector of 1s. More precisely, when A is unsymmetric we consider the *out degree*, so $D = \text{diag}\left(\sum_{j=1}^{N} a_{ij}\right)$ and $\widehat{D} = \text{diag}\left(\max(D_{ii}, 1)\right)$.

4.5.1 *PageRank Code: `pagerank`.* The call `P = pagerank(A,d)` takes an adjacency matrix `A` and returns the PageRank matrix `P`, with `d` defaulting to 0.85. The matrix `A` is not assumed to be symmetric; directed edges are allowed.

4.6 Mean Hitting Time Matrix

In many applications it is useful to consider the discrete time, finite state space, Markov chain that arises naturally from a network [Lovász 1996]. Here, if we are currently at node i then at the next time level we move to a node chosen uniformly among the neighbors of node i. The *transition matrix* for this Markov chain thus has the form $D^{-1}A$. Fixing a node, i, the the *mean hitting time* for node j is defined to be the average number of steps required for the Markov chain to reach state j, given that it starts at state i. The vector of mean hitting times can be found by solving the linear system $Mx = \mathbf{1}$, where $M \in \mathbb{R}^{n-1\times n-1}$ is the transition matrix with its ith row and column removed [Norris 1997].

4.6.1 *Mean Hitting Time Code: `mht`.* The call `M = mht(A,i)` takes an adjacency matrix `A` with nonzero out degrees and returns the mean hitting time matrix `M` for a chain that starts at node `i`, with `i` defaulting to `1`. The matrix `A` is not required to be symmetric.

4.7 Pathlength and Curvature

The *pathlength* between nodes i and j is the smallest number of edges that must be crossed to reach j starting from i. In terms of the adjacency matrix A, the pathlength between nodes i and j can be characterized as the smallest integer $k \geq 1$ such that $(A^k)_{ij} \neq 0$. If $(A^{n-1})_{ij} = 0$ then there is no suitable path and the pathlength may be regarded as infinite.

The *curvature*, or *clustering coefficient*, of a node was defined in Section 1. In MATLAB notation, the vector of clustering coefficients may be computed as follows.

```
diag(A^3)/(sum(A).*(sum(A) - 1))
```

4.7.1 *Pathlength and Curvature Codes: `pathlength` and `curvature`.* The call `Path = pathlength(A)` returns an array `Path` of the same dimension as the adjacency matrix `A`, such that `Path(i,j)` is the pathlength from node `i` to node `j`. We always set `Path(i,i)=0` and we use `Path(i,j)=inf` to denote that no path exists.

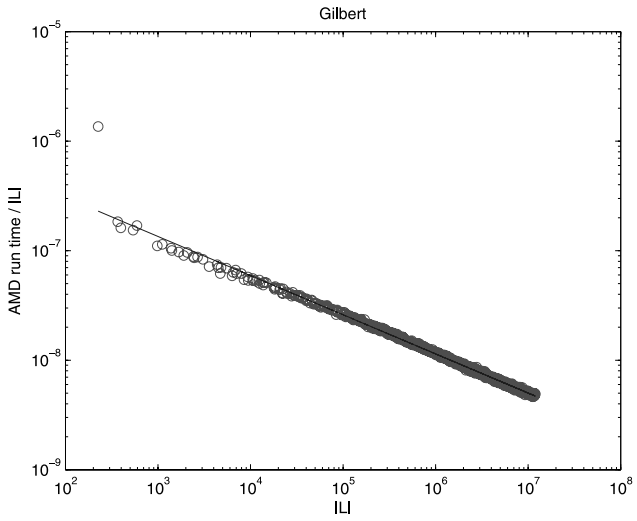

Fig. 3. `amd` runtimes for Gilbert model.

The call `curv = curvature(A)` takes an adjacency matrix `A` of dimension `n` and returns a one-dimensional array `curv` of length `n`, such that `curv(i)` records the curvature of node `i`. A second input argument is allowed. The call `curv = curvature(A,ind)` returns the maximum curvature if `ind` is the string `'max'`, the average curvature if `ind` is the string `'ave'`, and the curvature for the `i`th node if `ind` is the integer `i`. Undefined curvature evaluates to `NaN`.

5. COMPUTATIONAL EXPERIMENT

For a brief illustration of the toolbox in use, we follow Davis [2007] by examining the complexity of the minimum degree ordering algorithm, as implemented in MATLAB's `amd`. Letting L denote the Cholesky factor of the appropriate permuted version of A, we plot the runtime, scaled by $|L|$, against $|L|$, on a log-log scale. Davis [2007] distinguished between matrices from a deterministic test set coming from problems with and without inherent geometry. To mirror this, Figure 3 shows results for matrices arising from the Gilbert class, using `gilbert`, where there is no inherent structure, and Figure 4 shows results for matrices arising from the Kleinberg class, using `klein`, where there is an underlying lattice. The least-squares slope is indicated by a solid line. In each case the matrix dimension n was varied between 50 and 10,000. The test programs are available from the testing section of the toolbox Web site. The figures are consistent with the rule of thumb mentioned in Davis [2007] that the runtime is typically below $O(|L|)$.

6. SUMMARY: NETWORKS AS TEST MATRICES

The motivation for this work is the fact that recent random network models make excellent candidates for sparse test matrices. The models capture

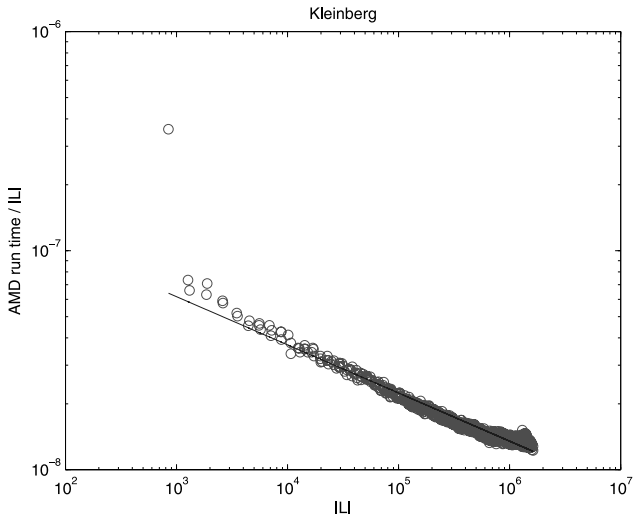

Fig. 4. `amd` runtimes for Kleinberg model.

features of interaction data observed across a wide range of application areas, and they incorporate parameters that allow the user to control topological features, including sparsity and the distribution of degree and clustering coefficients. Naturally arising computational tasks in network science present challenging test problems for general (symmetric and unsymmetric) linear system solvers and symmetric eigenvalue routines.

ACKNOWLEDGMENT

We thank T. Davis for useful advice about sparse matrix operations in MATLAB.

REFERENCES

ABELLO, J., BUCHSBAUM, A., AND WESTBROOK, J. 1998. A functional approach to external graph algorithms. Lecture Notes in Computer Science, vol. 1461, 332–343.

ALBERT, R. AND BARABÁSI, A.-L. 2002. Statistical mechanics of complex networks. *Rev. Modern Phys. 74*, 47–97.

ALON, U. 2006. *An Introduction to Systems Biology*. Chapman and Hall/CRC, London.

ANDERSON, E., BAI, Z., BISCHOF, C., BLACKFORD, S., DEMMEL, J., DONGARRA, J., CROZ, J. D., GREENBAUM, A., HAMMARLING, S., MCKENNEY, A., AND SORENSEN, D. 1999. *LAPACK Users' Guide*, 3rd ed. SIAM, PA.

BARABÁSI, A.-L. AND ALBERT, R. 1999. Emergence of scaling in random networks. *Sci. 286*, 5439, 509–12.

BATAGELJ, V. AND BRANDES, U. 2005. Efficient generation of large random networks. *Phys. Rev. E 71*, 036113.

BOGINSKI, V., BUTENKO, S., AND PARDALOS, P. M. 2003. On structural properties of the market graph. In *Innovations in Financial and Economic Networks*, A. Nagurney, Ed. Edward Elgar Publishers, 29–45.

BOISVERT, R., POZO, R., REMINGTON, K., BARRETT, R., AND DONGARRA, J. 1997. Matrix market: A Web resource for test matrix collections. In *The Quality of Numerical Software: Assessment and Enhancement*, R. Boisvert, Ed. Chapman and Hall, London, 125–137.

BOLLOBÁS, B. 1985. *Random Graphs*. Academic, London.

BOLLOBÁS, B., RIORDAN, O., SPENCER, J., AND TUSNÁDY, G. 2001. The degree sequence of a scale-free random graph process. *Random Structures Algor. 18*, 279–290.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., AND WIENER, J. 2000. Graph structure of the Web. *Comput. Netw. 33*, 309–320.

CHUNG, F. 1997. *Spectral Graph Theory*. American Mathematical Society, Providence, RI.

CONYON, M. J. AND MULDOON, M. R. 2006. The small world of corporate boards. *J. Business Finance Account. 33*, 1321–1343.

COUR, T., BENEZIT, F., AND SHI, J. 2005. Spectral segmentation with multiscale graph decomposition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 1124–1131.

DAVIS, T. 2007. The University of Florida sparse matrix collection. Tech. rep. CISE Department, REP-2007-298, University of Florida, USA.

DE SILVA, E. AND STUMPF, M. 2005. Complex networks and simple models in biology. *J. R. Soc. Interface 2*, 419–430.

DE SILVA, E., THORNE, T., INGRAM, P., AGRAFIOT, I., SWIRE, J., WIUF, C., AND STUMPF, M. P. H. 2006. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol. 4*, 39.

DUFF, I. S., GRIMES, R. G., AND LEWIS, J. G. 1989. Sparse matrix test problems. *ACM Trans. Math. Softw. 15*, 1–14.

ERDÖS, P. AND RÉNYI, A. 1959. On random graphs. *Publ. Math. Debrecen 6*, 290–297.

FAGIOLO, G. 2007. Clustering in complex directed networks. *Phys. Rev. 76*, 026107.

FALOUTSOS, M., FALOUTSOS, P., AND FALOUTSOS, C. 1999. On power-law relationships of the internet topology. *Comput. Commun. Rev. 29*, 251–262.

GILBERT, E. N. 1959. Random graphs. *Ann. Math. Statist. 30*, 1141–1144.

GILBERT, J. R., MOLER, C., AND SCHREIBER, R. 1992. Sparse matrices in MATLAB: Design and implementation. *SIAM J. Matrix Anal. Appl. 13*, 333–356.

GRIMMETT, G. 1999. *Percolation*, 2nd ed. Springer.

GRINDROD, P. 2002. Range-Dependent random graphs and their application to modeling large smal-world proteome datasets. *Phys. Rev. E 66*, 066702.

GRINDROD, P., HIGHAM, D. J., AND KALNA, G. 2008. Periodic reordering. Tech. rep. 6, University of Strathclyde, Department of Mathematics.

HAN, J. D. H., DUPUY, D., BERTIN, N., CUSICK, M. E., AND M., V. 2005. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnol. 23*, 839–844.

HENDRICKSON, B. AND LELAND, R. 1994. The Chaco user's guide: Version 2.0. Tech. rep. SAND94–2692, Sandia National Laboratories, Albuquerque, NM.

HIGHAM, D. J. 2003. Unravelling small world networks. *J. Comput. Appl. Math. 158*, 61–74.

HIGHAM, D. J. 2005. Spectral reordering of a range-dependent weighted random graph. *IMA J. Numer. Anal. 25*, 443–457.

HIGHAM, D. J. AND HIGHAM, N. J. 2000. *MATLAB Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

HIGHAM, D. J., PRŽULJ, N., AND RAŠAJSKI, M. 2008. Fitting a geometric graph to a protein-protein interaction network. *Bioinf. 24*, 1093–1099.

HU, Y. AND SCOTT, J. A. 2003. HSL_MC73: A fast multilevel Fiedler and profile reduction code. RAL-TR-2003-36, Numerical Analysis Group, Computational Science and Engineering Department, Rutherford Appleton Laboratory.

KAMPER, L., BOZKURT, A., RYBACKI, K., GEISSLER, A., GERKEN, I., STEPHAN, K. E., AND KÖTTER, R. 2002. An introduction to CoCoMac-Online. The online-interface of the primate connectivity database CoCoMac. In *Neuroscience Databases—A Practical Guide*, R. Kötter, Ed. Kluwer Academic, Norwell, MA, 155–169.

KAUFFMAN, S. A. 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol. 22*, 437–467.

KHANIN, R. AND WIT, E. 2006. How scale-free are gene networks? *J. Comput. Biol. 13*, 3, 810–818.

KISS, I. Z., GREEN, D. M., AND KAO, R. R. 2006. The network of sheep movements within Great Britain: Network properties and their implications for infectious disease spread. *J. Roy. Soc. Interface 3*, 669–677.

KLEINBERG, J. M. 2000. Navigation in a small world. *Nature 406*, 845.

LANGVILLE, A. N. AND MEYER, C. D. 2006. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, Princeton, NJ.

LOVÁSZ, L. 1996. Random walks on graphs: A survey. In *Paul Erdös is Eighty*, D. Miklós, V. T. Sós, and T. Szönyi, Eds. János Bolyai Mathematical Society, Budapest, 353–398.

MANGAN, S. AND ALON, U. 2003. Structure and function of the feed-forward loop network motif. *Proc. Nat. Acad. Sci. 100*, 11980–11985.

MANGAN, S., ZASLAVER, A., AND ALON, U. 2003. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Math. Biol. 334*, 2, 197–204.

MILENKOVIC, T., LAI, J., AND PRZULJ, N. 2008. GraphCrunch: A tool for large network analyses. *BMC Bioinf. 9*, 70.

MILGRAM, S. 1967. The small world problem. *Psychol. Today 2*, 60–67.

MILO, R., ITZKOVITZ, S., KASHTAN, N., LEVITT, R., SHEN-ORR, S., AYZENSHTAT, I., SHEFFER, M., AND ALON, U. 2004. Superfamilies of evolved and designed networks. *Sci. 303*, 1538–1542.

MORRISON, J. L., BREITLING, R., HIGHAM, D. J., AND GILBERT, D. R. 2005. Generank: Using search engine technology for the analysis of microarray experiments. *BMC Bioinf. 6*, 233.

MORRISON, J. L., BREITLING, R., HIGHAM, D. J., AND GILBERT, D. R. 2006. A lock-and-key model for protein-protein interactions. *Bioinf. 2*, 2012–2019.

NEWMAN, M. E. J. 2004. Who is the best connected scientist? A study of scientific coauthorship networks. In *Complex Networks*, E. Ben-Naim et al., Eds. Springer, 337–370.

NEWMAN, M. E. J., MOORE, C., AND WATTS, D. J. 2000. Mean-Field solution of the small-world network model. *Phys. Rev. Lett. 84*, 3201–3204.

NORRIS, J. R. 1997. *Markov Chains*. Cambridge University Press.

ONODY, R. N. AND DE CASTRO, P. A. 2004. Complex network study of Brazilian soccer players. *Phys. Rev. E 70*.

PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. 1998. The PageRank citation ranking: Bringing order to the Web. Tech. rep., Stanford Digital Library Technologies Project.

PENROSE, M. 2003. *Geometric Random Graphs*. Oxford Univeristy Press.

PORTER, M. A., MUCHA, P. J., NEWMAN, M. E. J., AND WARMBRAND, C. M. 2005. A network analysis of committees in the United States House of Representatives. *Proc. Nat. Acad. Sci. 102*, 7057–7062.

PRŽULJ, N., CORNEIL, D. G., AND JURISICA, I. 2004. Modeling interactome: Scale-Free or geometric? *Bioinf. 20*, 18, 3508–3515.

PRŽULJ, N., CORNEIL, D. G., AND JURISICA, I. 2006. Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinf. 22*, 974–980.

PRŽULJ, N. AND HIGHAM, D. J. 2006. Modeling protein-protein interaction networks via a stickiness index. *J. Roy. Soc. Interface 3*, 711–716.

SALATHÉ, M., MAY, R. M., AND BONHOEFFER, S. 2005. The evolution of network topology by selective removal. *J. Roy. Soc. Interface 2*, 533–536.

SPORNS, O. AND ZWI, J. D. 2004. The small world of the cerebral cortex. *Neuroinf. 2*, 145–162.

STUMPF, M. P. H., WIUF, C., AND MAY, R. M. 2005. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proc. Nat. Acad. Sci. 102*, 4221–4224.

THOMAS, A., CANNINGS, R., MONK, N. A. M., AND CANNINGS, C. 2003. On the structure of protein-protein interaction networks. *Biochem. Soc. Trans. 31*, 1491–1496.

TITZ, B., SCHLESNER, M., AND UETZ, P. 2004. What do we learn from high-throughput protein interaction data? *Expert Rev. Proteomics 1*, 111–121.

WATTS, D. J. AND STROGATZ, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature 393*, 440–442.

WILLIAMS, R. J., BERLOW, E. L., DUNNE, J. A., BARABÁSI, A.-L., AND MARTINEZ, N. D. 2002. Two degrees of separation in complex food webs. *Proc. Nat. Acad. Sci. 99*, 12913–12916.

XENARIOS, I., SALWINSKI, L., DUAN, X. J., HIGNEY, P., KIM, S. M., AND D., E. 2002. DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res. 30*, 1, 303–305.

Received June 2007; revised May 2008; accepted June 2008