Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Graph compression—save information by exploiting redundancy

Jie Sun[1], **Erik M Bollt**[1] **and Daniel ben-Avraham**[2]

[1] Department of Mathematics and Computer Science, Clarkson University, Potsdam, NY 13699-5815, USA
[2] Department of Physics, Clarkson University, Potsdam, NY 13699-5820, USA
E-mail: sunj@clarkson.edu, bolltem@clarkson.edu and qd00@clarkson.edu

Received 21 December 2007
Accepted 6 May 2008
Published 3 June 2008

Online at stacks.iop.org/JSTAT/2008/P06001
doi:10.1088/1742-5468/2008/06/P06001

Abstract. In this paper we raise the question of how to compress sparse graphs. By introducing the idea of redundancy, we find a way to measure the overlap of neighbors between nodes in networks. We exploit symmetry and information by making use of the overlap in neighbors and analyzing how information is reduced by shrinking the network and, using the specific data structure we created, we generalize the problem of compression as an optimization problem on the possible choices of orbits. To find a reasonably good solution to this problem we use a greedy algorithm to determine the orbit of symmetry identifications, to achieve compression. Some example implementations of our algorithm are illustrated and analyzed.

Keywords: random graphs, networks

ArXiv ePrint: 0712.3312

J. Stat. Mech. (2008) P06001

Contents

1. Introduction

Complex networks have been studied extensively in recent years in the fields of mathematics, physics, computer science, biology, sociology, etc [1]–[4]. Various networks are used to model and analyze real-world objects and their interactions with each other. For example, in sociology, airports and airflights that connect them can be represented by a network [12]; in biology, yeast reactions are also modeled by networks [9], etc. The mathematical terminology for a network is conveniently described in the language of graph theory. A common encoding of graphs uses an adjacency matrix, or an edge list, when the adjacency matrix is sparse [5]. However, even for a large network the edge list contains a large information storage. In the case that some important network is transferred frequently between computers, it will save time and cost if there is a scheme to efficiently encode, and therefore compress, the network first. Fundamentally we find it a relevant issue to ask how much information is necessary to present a given network, and how symmetry can be exploited to this end.

In this paper we will demonstrate one way to reduce the information storage of a network by using the idea that habitually graphs have many nodes that share many common neighbors. So, instead of recording all the links we could rather just store some of them and the difference between neighbors. The ideal compression ratio using this scheme will be $\eta = 2/\langle k \rangle$, where $\langle k \rangle$ is the average degree of the network, compared to the standard compression using the Yale Sparse Matrix Format [6, 7] which gives $\eta_Y = 1/2 + 1/\langle k \rangle$. In practice this ratio is not attainable but the real compression ratio is still better than using YSMF, as shown by our results.

J. Stat. Mech. (2008) P06001

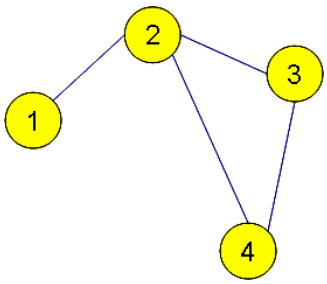

Figure 1. A drawing of a planar embedding of an example graph.

A graph $G = (V, E)$ is a set of vertices (or nodes) $V = \{v_1, v_2, \ldots, v_N\}$ together with edges (or links) $E = \{(v_i, v_j)\}$ which are the connected pairs. Graphs are often used to model networks. It is sometimes convenient to call the vertices that connect to a vertex i in a graph to be the neighbors of i. We will only consider undirected and unweighted graphs in this paper.

A drawing as in figure 1 allows us to directly visualize the graph (i.e. the nodes and the connections between them), but a truism that anyone who works with real-world graphs from real data knows is that commonly those graphs are so large that even a drawing will not give any insight. Visualizing structure in graphs of such sizes ($N > 100$–1000) begs for some computer assistance.

An *adjacency matrix* is a common, although inefficient, data representation of a graph. The adjacency matrix A_{G} of a graph $G = (V, E)$ is a $N \times N$ square matrix where N is the number of vertices of the graph and the entries $a_{i,j}$ of A_{G} are defined by

$$\begin{aligned} a_{i,j} &= 1 \qquad \text{if node i and node j are connected} \\ a_{i,j} &= 0 \qquad \text{else.} \end{aligned} \tag{1}$$

For example, the adjacency matrix A_{G} for the graph in figure 1 is

$$A_{\mathrm{G}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}. \tag{2}$$

However, in the case that the number of edges in a graph are so few that the corresponding adjacency matrix is sparse, the *edge list* will be used instead. The edge list is a list of all the pairs of nodes that form edges in a graph. It is essentially the same as the edge set E for a graph $G = (V, E)$. Using the edge list E_{G} to represent the same graph as above we will have

$$E_{\mathrm{G}} = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{2, 4\}\}. \tag{3}$$

Note here that in the edge list we actually record the label of nodes for each edge in the graph, so for an undirected graph, we can exchange the order for each pair of nodes.

We will only consider sparse simple graphs, whose adjacency matrices will thus be binary sparse matrices, and the standard information storage for such graphs or matrices will be the information units that are needed for the corresponding edge list (or two-dimensional arrays).

J. Stat. Mech. (2008) P06001

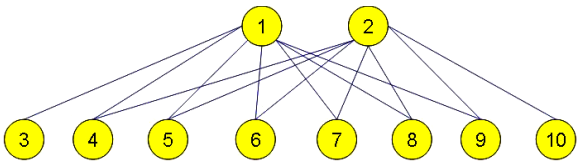

Figure 2. An extreme example which shows similarity between vertices.

We now sharpen the definition for the *unit of information* in our context. From the perspective of information theory, a message which contains N different symbols will require $\log_2 N$ bits for each symbol, without any further coding scheme. The edge list representation is one example of a text file which contains N different symbols (often represented by natural numbers from 1 to N) for a graph containing N vertices. Note that the unit of information depends only on the number of symbols that appear in the message, i.e. the number of vertices in a graph, so for any given graph this will be a fixed number. Thus, when we restrict the discussion to any particular graph, it is convenient to assume that each pair of labels in the edge list requires one information unit without making explicit what is the size of that unit. For example, the above graph requires four information units. In this paper we will focus on how to represent the same graph using fewer information units than its original representation.

2. A motivating example and the idea of redundancy

As a motivating example, let us consider the following graph and its edge list.

Note that here the neighbors of node 1 are almost the same as those of node 2. The edge list E_{G} for this graph will be

$$E = \{\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{1,7\},\{1,8\},\{1,9\},\\ \{2,4\},\{2,5\},\{2,6\},\{2,7\},\{2,8\},\{2,9\},\{2,10\}\}. \tag{4}$$

This requires 14 information units for the edge list. However, if we look back to the graph, we note that in this graph there are many common neighbors between node 1 and node 2, so there is a great deal of information redundancy. Considering the subgraphs, the neighbors of node 1 are almost the same as the neighbors of node 2, except that node 3 links to 1, but not 2, while node 10 links to 2, but not 1.

Taking the redundancy into account, we generate a new way to describe the same graph, exploiting the graphs. In the graph of figure 2, we see that the subgraph including vertices 1, 3, 4, 5, 6, 7, 8, 9 is very similar to the subgraph including vertices 2, 4, 5, 6, 7, 8, 9, 10, see figure 3. We exploit this redundancy in our coding.

We store the subgraph which only consists of node 1, and all its neighbors. Then, we add just two more parameters:

$$\alpha = (1,2) \tag{5}$$

and

$$\beta = \{-3,10\} \tag{6}$$

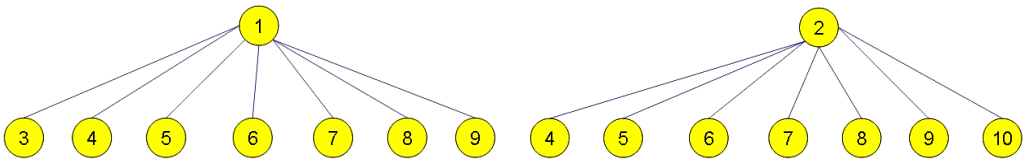

Figure 3. Similar subgraphs of the original graph. Here the subgraph containing node 1 (on the left) is very similar to the one dominated by node 2 (on the right).

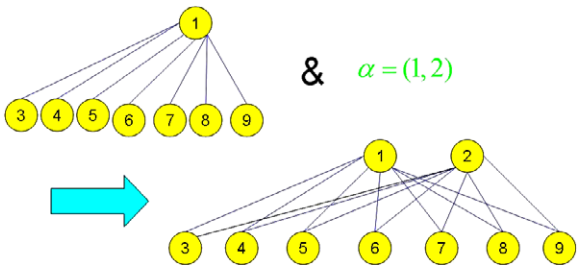

Figure 4. Construct from the subgraph and parameter $\alpha = (1, 2)$. 'Copy' from node 1 to node 2.

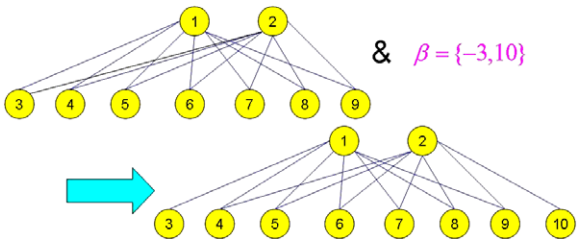

Figure 5. Add and delete links according to $\beta = \{-3, 10\}$.

that allow us to reconstruct the original graph. Here the ordered pair $\alpha = (1, 2)$ tells us that, in order to reconstruct the original graph, we need to first copy node 1 to node 2. By copy, we mean the addition of a new node into the existing graph with label 2, and then linking all the neighbors of node 1 to the new node 2. See figure 4.

The set $\beta = \{-3, 10\}$ tells us that we should then delete the link that connects the new nodes 2 and 3 and add a new link between 2 and 10. See figure 5.

After all these operations we see that we successfully reconstruct the graph with fewer information units: in this case, nearly half as many as the original edge list. So, instead of equation (4), we may use the edge list of the subgraph

$$E_{\mathrm{SG}} = \{\{1, 3\}, \{1, 4\}, \{1, 5\}, \{1, 6\}, \{1, 7\}, \{1, 8\}, \{1, 9\}\} \tag{7}$$

as well as two sets

$$\begin{aligned} \alpha &= (1, 2) \\ \beta &= \{-3, 10\} \end{aligned} \tag{8}$$

to represent the same graph (figure 6).

J. Stat. Mech. (2008) P06001

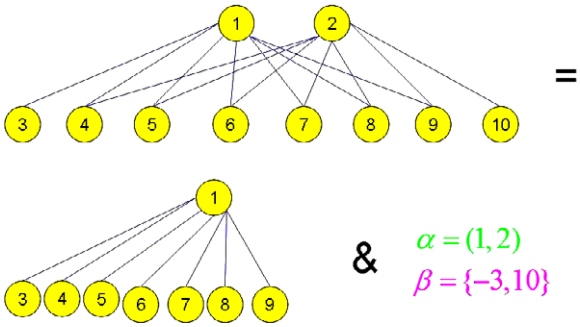

Figure 6. Reconstruction of the original graph using a subgraph and the parameters α and β.

The above example suggests that, by exploiting the symmetry of the graph, we might be able to reduce the information storage for certain graphs by using a small subgraph as well as α and β as defined above.

However, there remains the question of how to choose the pair of vertices so that we actually reduce the information, and which is the best possible pair? It is important to answer these questions since most of the graphs are so large that we will never be able to see the symmetry just by inspection as we did for the above toy example.

In the following we answer the first question, and partly the second, by using a greedy algorithm. In section 3 we will define information redundancy for a binary sparse matrix and show that it reveals the neighbor similarity between vertices in a graph which is represented by its corresponding adjacency matrix. Then in section 4 we will give a detailed description of our algorithm which allows us to implement our main idea. Then in section 5 we will show some examples of these applications followed by a discussion in section 6.

3. Information redundancy and compression of sparse matrices

Throughout this paper, we describe our methods in terms of manipulations of adjacency matrices to describe corresponding manipulations to the graphs. We choose this approach for pedagogical reasons, particularly regarding presentation of the analysis of information redundancy and algorithmic complexity. However, we emphasize that all of the necessary manipulations can and should be done in practice in terms of the more efficient edge list representation.

3.1. How to choose pairs of vertices to reduce information

The graphs we seek to compress are typically represented by large sparse adjacency matrices. An edge list is a specific data structure for representing such matrices to reduce information storage. We will consider the edge list form to be the standard way of storing sparse matrices, which requires M units of information for a graph with M edges. There are approaches for compressing sparse matrices, among which the most general is the Yale Sparse Matrix Format [6, 7], which does not make any assumptions on the structure of the matrix and only requires $\frac{1}{2}(M+N)$ units of information. There are other approaches, such

J. Stat. Mech. (2008) P06001

as [8], which emphasize not only the storage but also the cost for data access time. We will focus on the data storage, so the Yale Format will be considered as a basic benchmark approach for compression of a sparse matrix, to which we will compare our results. The Yale format yields the compression ratio (see the appendix)

$$\eta_Y = \frac{M+N}{2M} = \frac{1}{2} + \frac{1}{\langle k \rangle} \tag{9}$$

where $\langle k \rangle = 2M/N$ is the average degree of the graph.

We will show our approach of compressing the sparse matrices by first illustrating how the redundancy of a binary sparse matrix will be defined regarding our specific operation on the matrix.

Generally, the adjacency matrix is a binary sparse matrix, $A = \{a_{ij}\}$, where a_{ij} equals 0 or 1 indicating the connectivity between node i and j. For a simple graph consisting of M edges this matrix has $2M$ non-zero entries, but since it is symmetric only half of them are necessary to represent the graph, which yields M units of information for the edge list.

Now, if two nodes i and j in the graph share a lot of similar neighbors, in the adjacency matrix row i and row j will have a lot of common column entries, and likewise for column i and column j (due to the symmetry of the matrix).

Supposing that we apply the operation to the graph, mentioned in the last section, by choosing $\alpha = (i, j)$ and the corresponding β, we will not need row j and column j in the matrix to represent the graph. The number of non-zero entries in row j and column j is $2k_j$, where k_j is the degree of node j in the graph. By doing that, the number of non-zero entries in the new adjacency matrix becomes $2M - 2k_j$, which requires $M - k_j$ units of information. However, the extra information we have to record is encoded in α and β. α always has two entries, which requires 1 unit of information, and the units of information for β depend on the number of different neighbors between node i and node j. If i and j have Δ_{ij} different neighbors, the size of β will be

$$|\beta| = \Delta_{ij}, \tag{10}$$

and the units of information for β will thus be $\frac{1}{2}\Delta_{ij}$. Taking both the reduction of the matrix and the extra information into account, the actual information it requires after the operation is

$$M - k_j + 1 + \tfrac{1}{2}\Delta_{ij} = M - (k_j - 1 - \tfrac{1}{2}\Delta_{ij}). \tag{11}$$

This is true for i different from j. We could extend the operation to allow

$$\alpha = (i, i), \tag{12}$$

meaning a self-match. Then we will put all the neighbors of i into the corresponding set β and then delete these links associated with i. Then by a similar argument we find that after this operation we need

$$M - k_i + 1 + \tfrac{1}{2}k_i = M - (\tfrac{1}{2}k_i - 1) \tag{13}$$

units of information using the new format.

Note that here we need to clarify exactly the meaning of different neighbors since in the case that i and j are connected i is a neighbor of j but j is not, and likewise for j.

J. Stat. Mech. (2008) P06001

However, this extra information can be simply encoded in α by making the following rule: $\alpha = (i, j)$ means when we reconstruct we do not connect i and j and $\alpha = (i, -j)$ means we connect i and j when we reconstruct. Then we can write $\Delta_{ij} = \|A(i,:) - A(j,:)\|_1 - 2a_{ij}$.

From the above discussion we see that if we define

$$\begin{aligned} r_{ij} &= k_j - 1 - \tfrac{1}{2}\Delta_{ij}, \qquad i \neq j \\ r_{ii} &= \tfrac{1}{2}k_i - 1 \end{aligned} \tag{14}$$

then by choosing $\alpha = (i, j)$, r_{ij} measures exactly the amount of information it reduces. We call r_{ij} the information redundancy between node i and j. Note here that in general this redundancy is not symmetric in i and j, since for any pair of nodes Δ_{ij} is symmetric but the degree of these two nodes can be different, and deleting the node with higher degree will always reduce more units of information compared to deleting the lower degree node.

We form the redundancy matrix R by setting the entry in row i and column j to be r_{ij}. We perform the shrinking operation for the pair with maximum r_{ij}, thus saving the maximum amount of information.

For example, again using the graph from section 2, the adjacency matrix is

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{15}$$

and the corresponding redundancy matrix is

$$R = \begin{bmatrix} 2.5 & 5 & -3 & -2.5 & -2.5 & -2.5 & -2.5 & -2.5 & -2.5 & -4 \\ 5 & 2.5 & -4 & -2.5 & -2.5 & -2.5 & -2.5 & -2.5 & -2.5 & -3 \\ 3 & 2 & -0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & -1 \\ 2.5 & 2.5 & -0.5 & 0 & 1 & 1 & 1 & 1 & 1 & -0.5 \\ 2.5 & 2.5 & -0.5 & 1 & 0 & 1 & 1 & 1 & 1 & -0.5 \\ 2.5 & 2.5 & -0.5 & 1 & 1 & 0 & 1 & 1 & 1 & -0.5 \\ 2.5 & 2.5 & -0.5 & 1 & 1 & 1 & 0 & 1 & 1 & -0.5 \\ 2.5 & 2.5 & -0.5 & 1 & 1 & 1 & 1 & 0 & 1 & -0.5 \\ 2.5 & 2.5 & -0.5 & 1 & 1 & 1 & 1 & 1 & 0 & -0.5 \\ 2 & 3 & -1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & -0.5 \end{bmatrix}. \tag{16}$$

The maximum entry in R is $r_{12} = r_{21} = 5$, indicating that either choice of $\alpha = (1, 2)$ or $\alpha = (2, 1)$ will give the maximum information reduction, and the corresponding β can be obtained by recording the column entries in row 1 and row 2 according to our rule.

In the above discussion we only consider a one-step shrinking operation on the graph and find out the direct relationship between the maximum information reduction and the redundancy matrix. But we know that, after deleting one node, the resulting graph is

J. Stat. Mech. (2008) P06001

still sparse and so could be compressed further by our scheme. The question is then how to successively choose α and β to obtain the best overall compression.

3.2. On greedy optimization of the α, β, orbit

Let $\alpha_t = (i_t, j_t)$ denote the operation at step t, $t = 1, 2, \ldots, T$ (here the sign for j_t would not affect our analysis so, for convenience, we just write j_t). In order to analyze the multi-step effect, we first consider how the adjacency matrix A is affected by the orbit $\{\alpha_t\}$. Let $A_0 = A$ be the original adjacency matrix. Let A_t be the corresponding adjacency matrix after applying α_t and the entries in it be $A_t(i, j)$. On deleting node j_t we actually set row and column j_t to be zero in A_{t-1} and all the other entries are unchanged, to obtain the new matrix A_t, i.e.

$$\begin{aligned} A_t(i,j) &= A_{t-1}(i,j) \qquad \text{if } i, j \neq j_t \\ A_t(i,j) &= 0 \qquad \text{if } i = j_t \text{ or } j = j_t. \end{aligned} \tag{17}$$

So by induction we see that

$$\begin{aligned} A_t(i,j) &= A_0(i,j) \qquad \text{if } i, j \notin \{j_1, \ldots, j_t\} \\ A_t(i,j) &= 0 \qquad \text{if } i \in \{j_1, \ldots, j_t\} \text{ or } j \in \{j_1, \ldots, j_t\}. \end{aligned} \tag{18}$$

Then we analyze how the redundancy matrix R changes. Use R_t to represent the redundancy matrix, $k_t(i)$ the degree of node i and $\Delta_t(i, j)$ the number of different neighbors of node i and j, associated with the graph of A_t. Since our goal is to achieve compression, once a node is deleted in the graph it is useless for future operations. So we will set $R_t(i, j) = 0$ if i or j has been deleted before, i.e.

$$R_t(i,j) = 0 \qquad \text{if } i \in \{j_1, \ldots, j_t\} \text{ or } j \in \{j_1, \ldots, j_t\}. \tag{19}$$

Now, for those i and j that have not been deleted, i.e. $i, j \notin \{j_1, \ldots, j_t\}$, by equation (14) we see that $R_t(i, j) = k_t(j) - 1 - \frac{1}{2}\Delta_t(i, j)$ for $i \neq j$ and $R_t(i, i) = \frac{1}{2}k_t(i) - 1$. Since A_t is obtained by deleting row and column j_t in A_{t-1}, the degree of each node changes according to

$$k_t(i) = k_{t-1}(i) - A_{t-1}(i, j_t) \tag{20}$$

and Δ_{ij} changes according to

$$\Delta_t(i,j) = \Delta_{t-1}(i,j) - |A_{t-1}(i,j_t) - A_{t-1}(j,j_t)|. \tag{21}$$

Thus, we conclude that for $i \neq j$

$$\begin{aligned} R_t(i,j) &= k_{t-1}(j) - A_{t-1}(j,j_t) - 1 - \tfrac{1}{2}[\Delta_{t-1}(i,j) - |A_{t-1}(i,j_t) - A_{t-1}(j,j_t)|] \\ &= k_{t-1}(j) - 1 - \tfrac{1}{2}\Delta_{t-1}(i,j) - A_{t-1}(j,j_t) + \tfrac{1}{2}|A_{t-1}(i,j_t) - A_{t-1}(j,j_t)| \\ &= R_{t-1}(i,j) + [\tfrac{1}{2}|A_{t-1}(i,j_t) - A_{t-1}(j,j_t)| - A_{t-1}(j,j_t)] \end{aligned} \tag{22}$$

and for $i = j$

$$\begin{aligned} R_t(i,i) &= \tfrac{1}{2}k_t(i) - 1 \\ &= \tfrac{1}{2}(k_{t-1}(i) - A_{t-1}(i,j_t)) - 1 \\ &= R_{t-1}(i,i) - \tfrac{1}{2}A_{t-1}(i,j_t). \end{aligned} \tag{23}$$

J. Stat. Mech. (2008) P06001

By induction, we obtain that for $i \neq j$

$$R_t(i,j) = R_0(i,j) + \sum_{\tau=1}^{t} [\tfrac{1}{2}|A_{\tau-1}(i,j_\tau) - A_{\tau-1}(j,j_\tau)| - A_{\tau-1}(j,j_\tau)] \tag{24}$$

and for $i = j$

$$R_t(i,i) = R_0(i,i) + \sum_{\tau=1}^{t} [-\tfrac{1}{2}A_{\tau-1}(i,j_\tau)]. \tag{25}$$

By use of the fact that $i, j \notin \{j_1, \ldots, j_t\}$, by equation (17), we can simplify the above two expressions to yield

$$\begin{aligned} R_t(i,j) &= R_0(i,j) + \sum_{\tau=1}^{t} [\tfrac{1}{2}|A_0(i,j_\tau) - A_0(j,j_\tau)| - A_0(j,j_\tau)] \qquad \text{if } i \neq j \\ R_t(i,i) &= R_0(i,i) + \sum_{\tau=1}^{t} [-\tfrac{1}{2}A_0(i,j_\tau)]. \end{aligned} \tag{26}$$

Note that, if we choose a pair (i_t, j_t) at step t, the information we save is measured by $R_{t-1}(i_t, j_t)$. Thus, for any orbit $\{\alpha_t = (i_t, j_t)\}_{t=1}^{T}$ satisfying $i_t, j_t \notin \{j_1, \ldots, j_{t-1}\}$ for $t = 2, 3, \ldots, T$ (we call such an orbit a *natural orbit*), the total information reduction (or information saving) will be

$$\begin{aligned} s(\{\alpha_t\}_{t=1}^{T}) &= \sum_{t=1}^{T} R_{t-1}(i_t, j_t) \\ &= \sum_{t=1}^{T} [R_0(i_t, j_t) + c(i_t, j_t, t)] \end{aligned} \tag{27}$$

where c is defined by

$$\begin{aligned} c(i,j,t) &= \sum_{\tau=1}^{t} [\tfrac{1}{2}|A_0(i,j_\tau) - A_0(j,j_\tau)| - A_0(j,j_\tau)] \qquad \text{if } i \neq j \\ c(i,i,t) &= \sum_{\tau=1}^{t} [-\tfrac{1}{2}A_0(i,j_\tau)]. \end{aligned} \tag{28}$$

So the compression problem can be stated as

$$\text{Find} \max_{\{\alpha_t\}_{t=1}^{T}} s(\{\alpha_t\}_{t=1}^{T}). \tag{29}$$

One more thing to mention is that the length of the orbit, T, is also a variable, which could not be larger than N since there are only N nodes in the graph and it is meaningless to delete an 'empty' node which does not even exist.

J. Stat. Mech. (2008) P06001

4. Greedy algorithm for compression

From the previous section we see that, for a given adjacency matrix, the final compression ratio depends on the orbit $\{\alpha_t\}_{t=1}^{T}$ we choose, and the compression problem becomes an optimization problem. However, to find the maximum of s and the corresponding best orbit is not trivial. One reason is that the number of natural orbits is of the order of $N!$, which makes it impractical to test and try for all possible orbits. Another reason which is crucial here is that, for any given orbit of length T, evaluating s costs $O(T^2)$ operations, making it hard to find an appropriate scheme to search for the true maximum or even the approximate maximum. Instead, we use a greedy algorithm to find an orbit which gives a reasonable compression ratio, and which is easy to apply.

The idea of the greedy algorithm is that, at each iteration step, we choose the pair of nodes i_t and j_t which maximizes $R_{t-1}(i, j)$ over all possible pairs, and we stop if the maximum value is non-positive. Also we need to record α and β according to the graph.

Here we summarize the greedy algorithm as pseudocode.

Given the adjacency matrix A of a graph (N nodes and M edges).

Begin:

Set $A_0 = A$;

Calculate $R_0(i, j)$ for all $i, j = 1, \ldots, N$. This forms the redundancy matrix $R_0 = R$.

Set $t = 1$.

1. Let $R_{t-1}(i_t, j_t)$ be the largest element in R_{t-1}.
   If $R_{t-1}(i_t, j_t) > 0$
   record $\alpha_t = (i_t, j_t)$,
   then go to step 2.
   Else,
   End.
2. Set β_t according to the difference between the two rows of$\alpha_t$in A_{t-1},
   Update A_{t-1} to A_t according to (17);
   Update R_{t-1} to R_t according to (22) and (23) for $i, j \neq j_t$;
   Set $R_t(i, j) = 0$ for i or $j = j_t$.
3. Set $t = t + 1$ and go to step 1.

The compressed version of the matrix will consist of: the final matrix A_T, the orbit $(\alpha_1, \ldots, \alpha_T)$ and the vectors $\{\beta_1, \ldots, \beta_T\}$, which will allow us to reconstruct $A = A_0$ and any intermediate matrix A_t during the compression process.

The computational complexity of this greedy algorithm is dominated by the initialization of the redundancy matrix R, which requires $O(MN)$ operations[3]. The subsequent operations will each require only an update of R according to formula (22) and (23), resulting in $O(N)$ operations per step. Thus the overall cost of the greedy algorithm will be $O(MN)$ and the average cost per step is $O(MN/T)$, where T is the number of shrinking steps ($T < N$).

[3] Here $O(MN)$ comes from the fact that we are comparing N^2 pair of vertices, and each comparison requires $O(M/N)$ operations (in an efficient format like the edge list) since the matrix is sparse. Thus the cost for initializing R is $O(N^2 \cdot M/N) = O(MN)$.

J. Stat. Mech. (2008) P06001

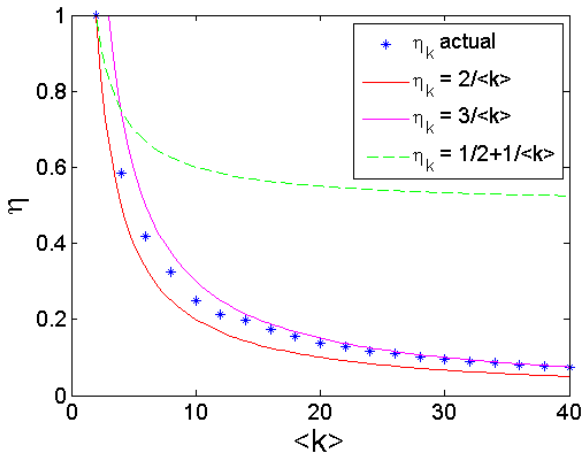

Figure 7. Compression results for lattice graphs. Stars indicate the final compression ratios for the lattice graphs with $\langle k\rangle$ 2–40. The compression limit is indicated by the bottom curve given by $\eta_k = 2/\langle k\rangle$, and we find that for $\langle k\rangle$ large the compression ratio is close to the empirical formula: $\eta_k = 3/\langle k\rangle$ (upper curve). For comparison, we plot the result using YSMF (broken line): $\eta_k = 1/2 + 1/\langle k\rangle$. For $\langle k\rangle > 2$, our algorithm always achieves a better result than the YSMF and the advantage increases with increasing $\langle k\rangle$.

5. Examples of application to graphs

In this section we will show some examples of our compression scheme on several networks. We begin with the lattice graph, which is expected to be readily compressible due to the high degree of overlapping between neighbors of nodes. As a secondary example, we add some random alterations and apply our method to the corresponding Watts–Strogatz network. Finally we show some results for real-world networks.

5.1. A simple benchmark example: lattice graph

One of the most symmetric graphs is the lattice graph, a one-dimensional chain where each site is connected to $k/2$ nearest neighbors to its right and left. In this case $\langle k\rangle = k$ represents the degree of each vertex in the lattice graph. The total number of nodes is $N \gg \langle k\rangle$ and the corresponding adjacency matrix is sparse.

We implement our algorithm for a lattice graph with different $\langle k\rangle$. The results are shown in figure 7. Here we take $N = 500$.

5.2. Compressing a Watts–Strogatz small-world graph

It is not surprising that the lattice graphs are easy to compress since these graphs are highly symmetric and nodes have lots of overlaps in their neighbors. However, in the case that we do not have such perfect symmetry, we still hope to achieve compression. Here we apply our algorithm to the WS graphs. The WS graph comes from the famous Watts–Strogatz model for real-world networks by showing the so-called small-world phenomenon. The WS graph is generated from a lattice graph by the usual rewiring of each edge with some given probability p from the uniform distribution.

J. Stat. Mech. (2008) P06001

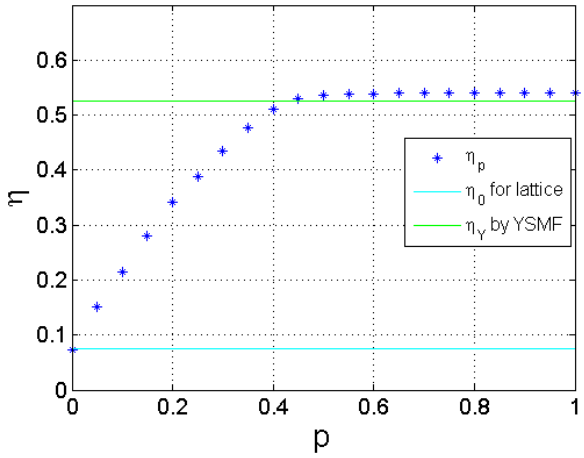

Figure 8. Compression results for WS graphs. Here the base lattice graph is with $N = 500$ and $\langle k \rangle = 40$. The stars show the compression results by our algorithm. The lower line is the compression ratio for the lattice $N = 500$ and $\langle k \rangle = 40$ and the upper line is the ratio from the YSMF. We see that, as p increases, there is less and less overlap between neighbors in the network and the compression ratio increases. For $p \sim 0.5$, we obtain a worse result than YSMF.

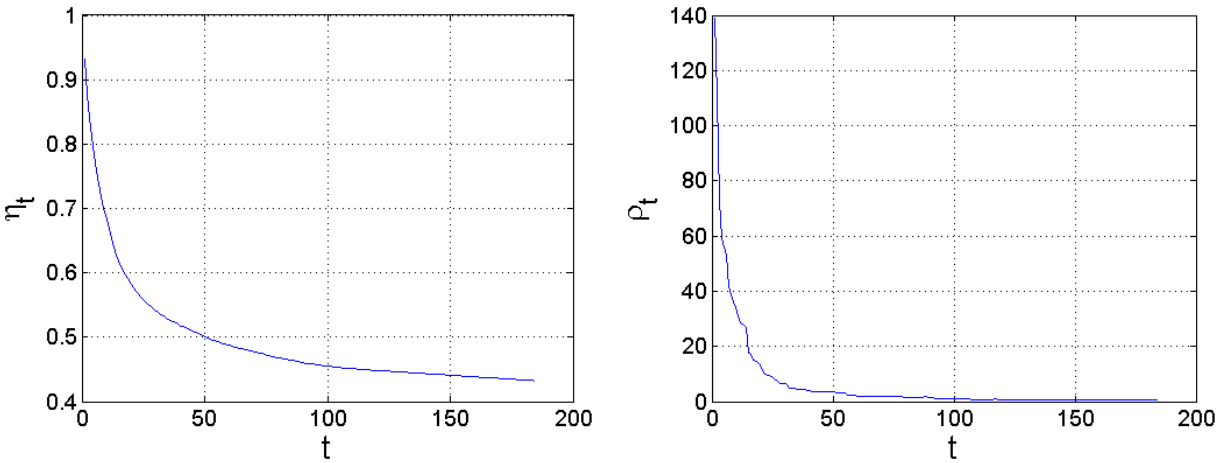

Figure 9. Compression process for metabolic network [10]: compression ratio η during each step (left) and information redundancy ρ for each step (right).

We apply our algorithm to WS graphs with different p to explore how p affects the compression behavior. Results are shown in figure 8.

5.3. Real-world graphs

In the following we show the compression results for some real-world graphs: a C.elegans metabolic network [10] (figure 9), a yeast network constructed from yeast reactions [9], an e-mail network [11] and an airline network of flight connections [12]. In table 1 we summarize the compression results for these real-world graphs.

Table 1. Compression results for some networks.

Network	N	$\langle k\rangle$	η_Y	η	η_*
Lattice	N	$\langle k\rangle$	$\frac{1}{2}+\frac{1}{\langle k\rangle}$	$\frac{3}{\langle k\rangle}$	$\frac{2}{\langle k\rangle}$
Yeast [9]	2361	6.08	0.66	0.50	0.33
Metabolic [10]	453	9.01	0.61	0.43	0.22
Email [11]	1133	9.62	0.60	0.49	0.21
Airline [12]	332	12.81	0.58	0.31	0.16

6. Discussion

From the previous section we see that our algorithm works for various kinds of graphs and gives a reasonable result. The ideal limit of our method for a graph with N nodes, M edges and average degree $\langle k\rangle = 2M/N$, which is relatively large, is $2/\langle k\rangle$. This is obtained when each β_t during the compression process is empty, meaning that most of the nodes share common neighbors, in which case we only need to record all the α_t, requiring $\frac{N}{2}$ units of information and yielding

$$\eta = \frac{N}{2M} = \frac{2}{\langle k\rangle}. \tag{30}$$

Notice that trees do not compress, since for trees $\langle k\rangle = 2$, so on average the overlap in neighbors will be even smaller (likely to be 0), and a possible way to achieve compression is by self-matching for large degree nodes, for example, the hubs in a star graph. For comparison, the YSMF always gives the compression ratio

$$\eta_Y = \frac{1}{2} + \frac{1}{\langle k\rangle} \tag{31}$$

which does not compress trees and has a lower bound $\frac{1}{2}$, while our method in principle approaches 0 as $\langle k\rangle \to \infty$. Actually the compression ratio using YSMF can be achieved by choosing a special orbit in our approach which only contains self-matches α, i.e.

$$\{\alpha_t\}_{t=1}^{T} = \{(i,i)\}_{i=1}^{N}. \tag{32}$$

In this case the neighbors of each node will be put into corresponding β sets and, since any α_i contains the same pair of numbers (i,i), we can just use one i to represent the pair, resulting in a total of $(N+M)/2$ information units. So our approach can be considered as a generalization of the YSMF.

However, as we observed in our compression results, the compression ratio given by (30) is, in general, not attainable since it is only achieved for the ideal case that nearly every node in the graph shares the same neighbors, and yet the graph needs to be sparse! However, for lattices we observe that the actual compression ratio achieved by our algorithm is about $3/\langle k\rangle$, which is of the same order as the ideal compression ratio. For WS graphs, when the noise p is small, our algorithm achieves a better compression ratio than YSMF, and the compression ratio is nearly linearly dependent on p for $p < 0.5$. For $p > 0.5$ the graph resembles Erdos–Renyi random graphs [13], there is no symmetry between nodes to be used and thus our approach does not give good results, as compared to the YSMF.

doi:10.1088/1742-5468/2008/06/P06001

J. Stat. Mech. (2008) P06001

For real-world graphs, the results from our algorithm are better than using YSMF, but not as good as we observed for lattice graphs. This suggests that in real-world graphs nodes, in general, share a certain amount of common neighbors even when the total number of links is small. This kind of overlap in neighbors is certainly not as common as we see in lattice graphs since real-world graphs, in general, have more complicated structures.

Acknowledgments

JS and EMB have been supported for this work by the Army Research Office grant 51950-MA. EMB has been further supported by the National Science Foundation under DMS-0708083 and DMS-0404778, and DBA is supported by the National Science Foundation under PHY-0555312. We thank Joseph D Skufca and James P Bagrow for discussions.

Appendix. Yale sparse matrix format for adjacency matrices of simple graphs

We illustrate first the standard form of YSMF by a simple example. Consider the matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \tag{A.1}$$

which can be encoded using YSMF by a collection of three arrays:

$$\begin{aligned} VA &= [1,1,1,1,1,1,1,1], \\ IA &= [1,2,5,7,9], \\ JA &= [2,1,3,4,2,4,2,3], \end{aligned} \tag{A.2}$$

where VA contains the values of all non-zero entries of A, listed in the order from left to right then top to bottom of the matrix A; IA encodes the number of non-zero entries of each row of A, $IA(1) := ||A(1,:)||_1$ and $IA(i+1) := IA(i) + ||A(i,:)||_1$ (for $i = 1, \ldots, N$); and JA contains the column indices of non-zero entries in A, listed in the same order as entries in VA.

However, for simple graphs, the corresponding adjacency matrices are symmetric and binary; thus it is sufficient to store only the upper (or lower) half of the matrix. Without loss of generality, we consider the upper half of A, so that $A(i,j) = 0$ whenever $i > j$. The array VA will consist of all 1's and thus is not necessary; also in IA we can let $IA(i) = ||A(i,:)||_1$ instead of the above definition so that IA contains exactly N entries, while JA is the same as the standard form of YSMF, which now contains M entries where M is the number of edges of the graph.

Thus, for an adjacency matrix of a simple graph, it requires $N + M$ entries to be stored using YSMF, which requires $(N + M)/2$ information units by the definition in the early sections in this paper. Compare to the original information storage M, we obtain the compression ratio $(M + N)/2M$.

References

[1] Watts D J and Strogatz S H, 1998 *Nature* **393** 440
[2] Watts D J, 1999 *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton, NJ: Princeton University Press)
[3] Barabasi A-L and Albert R, 1999 *Science* **286** 509
[4] Albert R and Barabasi A-L, *Statistical mechanics of complex networks*, 2002 *Rev. Mod. Phys.* **74** 47
[5] http://en.wikipedia.org/wiki/Adjacency_matrix and http://en.wikipedia.org/wiki/Adjacency_list
[6] Eisenstat S C, Elman H C, Schultz M H and Sherman A H, *The (new) Yale sparse matrix package*, 1984 *Elliptic Problem Solvers II* ed G Birkho and A Schoenstadt (New York: Academic) pp 45–52
[7] Eisenstat S C, Gursky M C, Schultz M H and Sherman A H, *Yale sparse matrix package I: the symmetric codes*, 1982 *Int. J. Numer. Methods Eng.* **18** 1145
[8] Tarjan R and Yao A, *Storing a sparse table*, 1971 *Commun. ACM* **22** 606
[9] Sun S, Ling L, Zhang N, Li G and Chen R, 2003 *Nucleic Acids Res.* **31** 2443
[10] Duch J and Arenas A, 2005 *Phys. Rev.* E **72** 027104
[11] Guimera R, Danon L, Diaz-Guilera A, Giralt F and Arenas A, 2003 *Phys. Rev.* E **68** 065103(R)
[12] Batagelj V and Mrvar A, 2006 *Pajek datasets* http://vlado.fmf.uni-lj.si/pub/networks/data/
[13] Bollobas B, 2001 *Random Graphs* 2nd edn (Cambridge: Cambridge University Press)

J. Stat. Mech. (2008) P06001