A Letters Journal Exploring
the Frontiers of Physics

March 2008

EPL, **81** (2008) 68004
doi: 10.1209/0295-5075/81/68004

www.epljournal.org

Portraits of complex networks

J. P. Bagrow[1], E. M. Bollt[2,1], J. D. Skufca[2] and D. ben-Avraham[1]

[1] *Department of Physics, Clarkson University - Potsdam, NY 13699-5820, USA*
[2] *Department of Math and Computer Science, Clarkson University - Potsdam, NY 13699-5815, USA*

received 21 November 2007; accepted in final form 28 January 2008
published online 27 February 2008

PACS 89.75.Hc – Networks and genealogical trees
PACS 02.10.Ox – Combinatorics; graph theory
PACS 89.75.-k – Complex systems

Abstract – We propose a method for characterizing large complex networks by introducing a new matrix structure, unique for a given network, which encodes structural information; provides useful visualization, even for very large networks; and allows for rigorous statistical comparison between networks. Dynamic processes such as percolation can be visualized using animation.

Introduction. – Large, complex stochastic networks are conspicuous in science and everyday life and have attracted a great deal of interest [1–3]. A difficult problem when studying networks is that of comparison and identification. Given two networks, how similar are they? Could they have arisen from the same generating mechanism? Given a real-world network, such as a protein-protein interaction network, or an electric power grid, say, how can one determine which stochastic network model most accurately captures its relevant structure? Is there a reasonable way to illustrate what a particular network looks like?

A network, or graph, is characterized completely by its *adjacency matrix* —an $N \times N$ matrix whose nonzero entries denote the various links between the graph's N nodes. This representation, however, is not unique, in that it depends on the actual labeling of the nodes, and graph *isomorphs* (identical graphs with permuted labels) cannot be readily distinguished from one another [4]. The same is true of graphical representations, where node placement is arbitrary (fig. 1).

In this letter, we propose a new method for recognizing and characterizing large complex networks that is independent of labeling and circumvents the problem of graph isomorphism. For each network we compute its B-matrix: a signature that represents the network reliably and serves as its "portrait." We thus have a means for recognizing networks at a glance and judge their differences and similarities, for the first time, enormously increasing our understanding and intuition [5]. We also introduce a "distance," derived from the B-matrix, that quantifies network differences, rendering comparisons mathematically meaningful. One important application is to the comparison of phylogenetic trees representing various organisms [6].

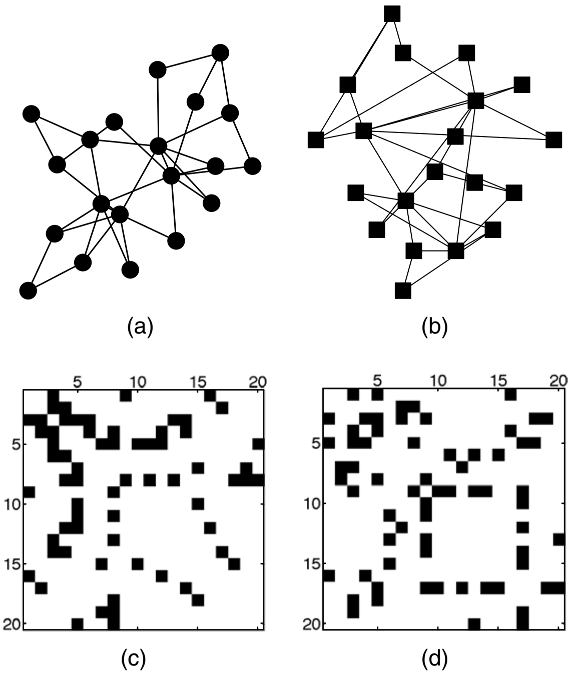


Fig. 1: Planar embeddings and adjacency matrices for a small network. It is difficult to tell visually that these represent the same network, even at such a small size.

Portraits. – A graph G consists of a finite set of nodes, or vertices, $V = \{v_1, v_2, ..., v_N\}$, and a set of edges, or links, between pairs of vertices, $E = \{(v_i, v_j)\}$. In applications, the vertices label elements of a network, and edges denote relationships between elements. The number of links, k_i, connected to a vertex v_i is the *degree* of the vertex. Much recent interest has focused on *scale-free* networks, which exhibit a power law degree distribution, $P(k) \sim k^{-\gamma}$. Despite its strong influence on various properties, the degree distribution is but one of many characteristics. Two large networks may possess similar degree distributions yet differ widely in *clustering* (the extent to which neighbors of a node connect to one another) [1], *assortativity* (the frequency of connections between nodes of like degrees) [7], and other important properties.

We now introduce the B-matrix. Define the *distance* between two nodes as the smallest number of links connecting them, found using Breadth-First Search (BFS) [8]. Thus, a node v_i is surrounded by ℓ-shells: the subsets of nodes at distance ℓ from v_i. Let

$$B_{\ell,k} = \text{number of nodes that have exactly } k \text{ members in their respective } \ell\text{-shells.} \qquad (1)$$

Note that B is independent of node labeling: all isomorphs of a graph have exactly the same B-matrix. Enumerating the shell members of a specific node requires $\mathcal{O}(N)$ steps for a sparse graph [8], thus construction of the B-matrix requires $\mathcal{O}(N^2)$ steps. Example B-matrices are shown in figs. 2–8 and discussed in the results section.

It is easy to see, from (1), that the degree distribution of a graph is encoded in the first row of its B-matrix,

$$B_{1,k} = NP(k), \qquad (2)$$

since the degree of a node equals the number of neighbors in its $\ell = 1$ shell. Generalizing this concept, we define the *degree of order* ℓ of a node as the number of members in its ℓ-th shell. Then, row ℓ of the B-matrix lists the graph's distribution of degrees of order ℓ:

$$B_{\ell,k} = NP_\ell(k). \qquad (3)$$

Consider a *maximally random* network, constructed by the Molloy-Reed algorithm [9]. Its structure is fully determined by its (first-order) degree distribution, or by the first row of its B-matrix. For example, the second row is

$$B_{2,k} = \sum_l B_{1,l} \sum_{\substack{j_1,j_2,\ldots,j_l \\ j_1+j_2+\cdots+j_l=k+l}} (p_{j_1}+p_{j_2}+\cdots+p_{j_l}), \qquad (4)$$

where $p_m \equiv mB_{1,m}/\sum_n nB_{1,n}$. Thus the B-matrix contains much additional information beyond the degree distribution, encoded in the difference between the actual $B_{2,k}$ and the expression (4) (and similarly for higher rows).

Results. – The intuition one gains simply by *looking* at these portraits is of great value [5]. Classification and comparison are immediate (figs. 5, 7). Dimensionality and regularity are encoded in the overall slope and row variances (fig. 6), while small-world behavior is displayed in the "aspect ratio" (fig. 4). Even correlation effects are discernable in the fine scale structure of the higher rows (fig. 8). Properties such as assortativity were previously impossible to visualize for even moderately sized networks.

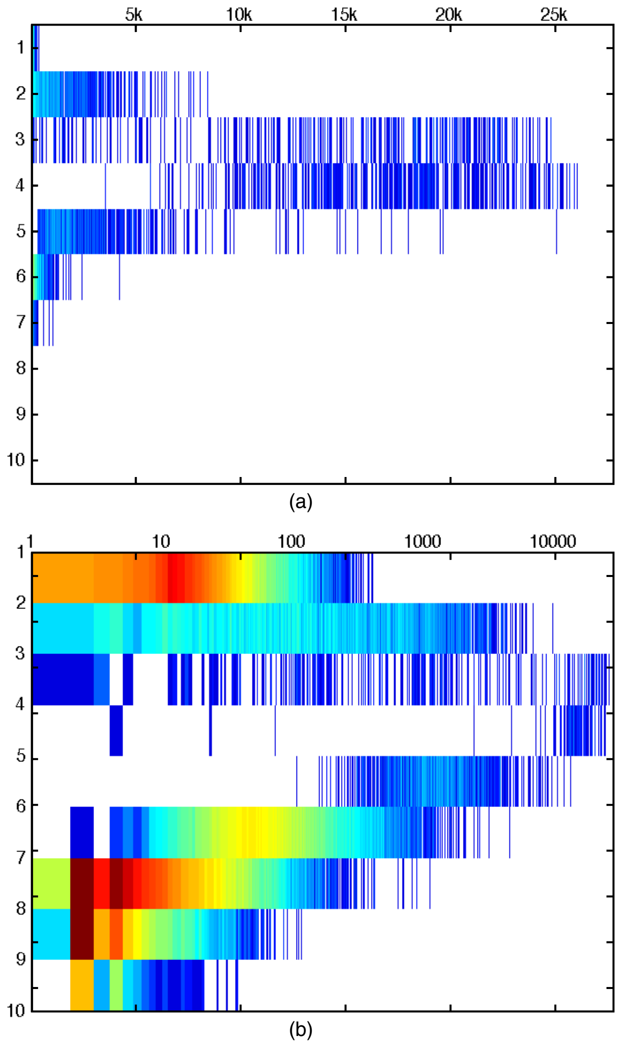


Fig. 2: (Color online) (a) A B-Matrix with a logarithmic color scale (the white background indicates zero elements of B). The degree distribution is slightly visible in the first row. The "turning point" about row 4 represents finite-size effects. Shown is the network of the 10% most connected actors on IMDB [2]. (b) The same matrix with a logarithmic horizontal axis. The degree distribution is now clearly visible.

Here is a list summarizing the contents and "moral of the story" for each panel, numbered by figure:

2. The algorithm is cheap enough to visualize very large matrices, as indicated by this example and its nearly 30000 columns. This also shows that a large amount of information is present in the matrix, far beyond the degree distribution encoded in the first row.

3. A large random network's B-matrix looks like the average of an ensemble of such networks (of the same size) (panels a,b). A phase transition such as percolation is immediately visible (c,d).

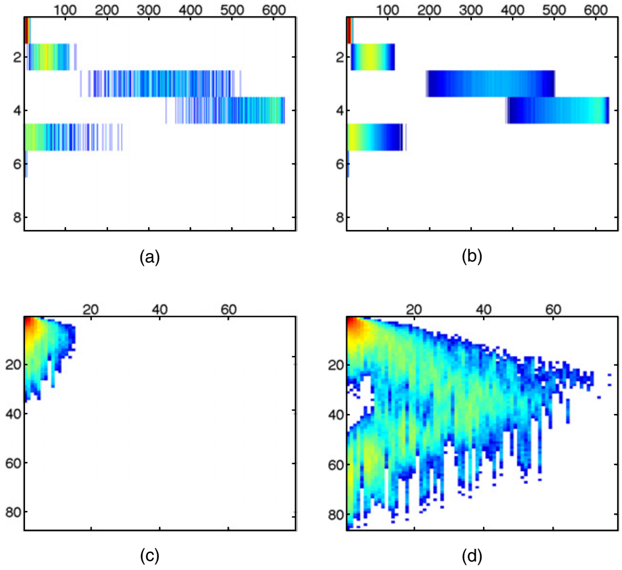


Fig. 3: (Color online) Erdős-Rényi (ER) graphs [10]. (a) One graph with $N = 1000$ nodes and $p = 0.008$. (b) The average of 100 graphs from (a). Visualizing percolation: $N = 10^4$ (c) below percolation, $p = (1.1N)^{-1}$; (d) at percolation, $p = 1/N$.

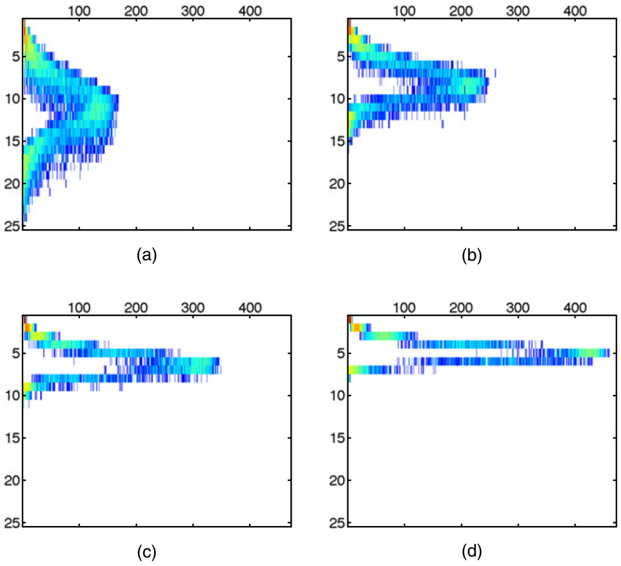


Fig. 4: (Color online) The emergence of small world. Shown are Newman-Watts-Strogatz graphs [11] with $N = 1000$; $k = 4$; and $p = 1/20, 1/10, 1/5$, and $2/5$; (a)–(d), respectively.

4. The transition to small world is visible in the changing "aspect ratio" of the portrait. These portraits have all been padded to the same dimensions.

5. Scale-free networks with identical numbers of nodes and power law exponents can still give radically different portraits. Thus the portrait can be used to infer a generating mechanism or scale-free model, by providing information beyond the degree distribution.

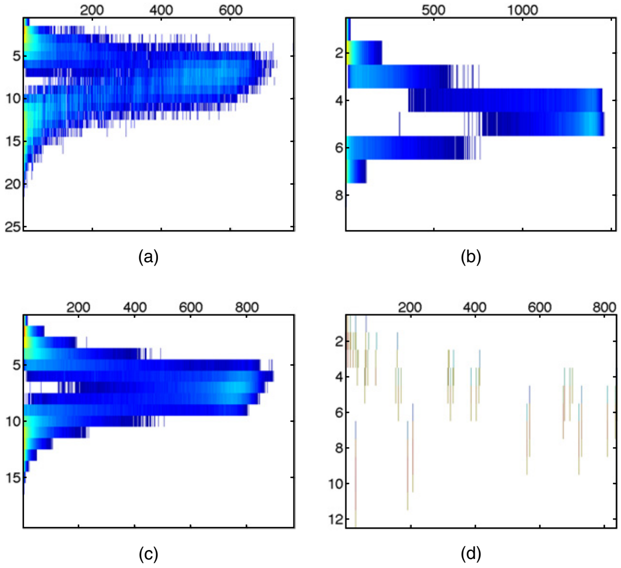


Fig. 5: (Color online) Scale-Free models. The average of 100 instances of the (undirected) Krapivsky-Redner ($r = 1/2$) [12]; Barabási-Albert (BA) ($m = 2$) [13]; and Molloy-Reed (MR) (drawn from $P(k) \sim k^{-3}$) [9] networks; as well as the (1,3)-Flower at generation 6 [14]; (a)–(d), respectively. All have $N = 2732$, $\gamma \approx 3$, but $\langle k \rangle$ varies. Note that (d) has been darkened slightly for clarity.

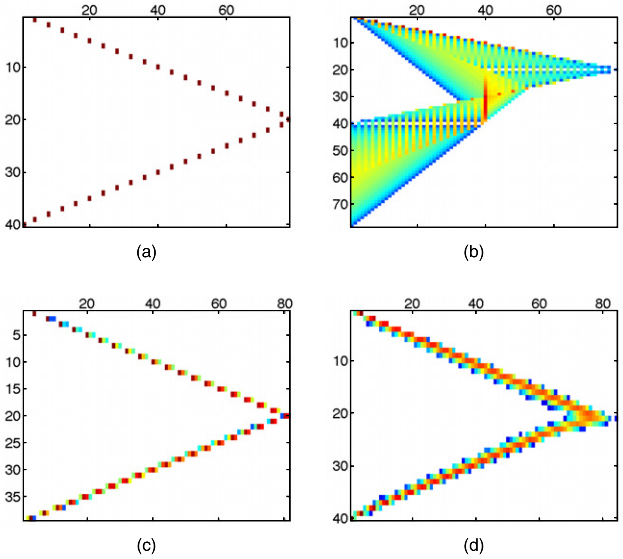
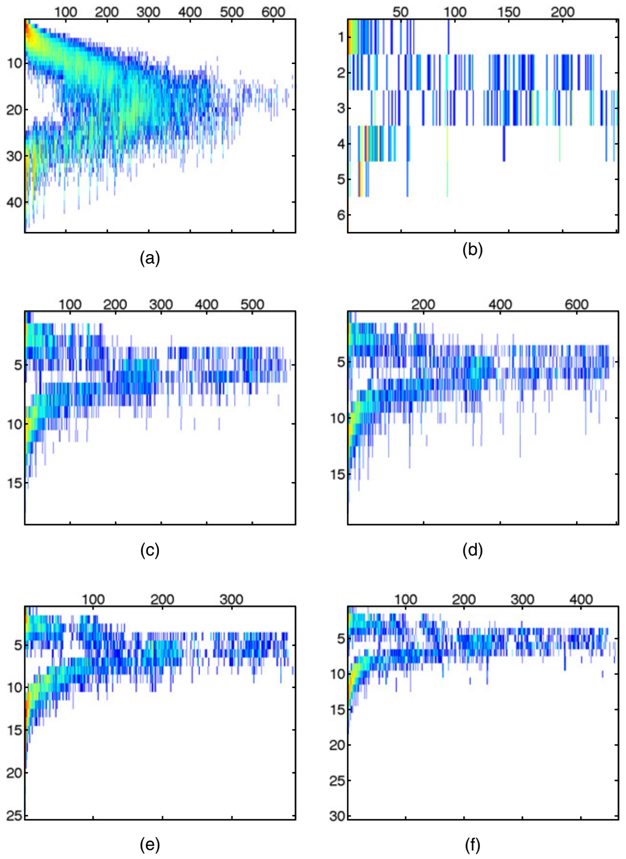


Fig. 6: (Color online) Regular 40×40 lattices with defects. (a) A periodic and (b) non-periodic lattice; (c) a lattice with skew-periodic boundaries; and (d) a periodic lattice with a random 5 percent of all nodes missing. Observe the strong linear slope, indicating the underlying two-dimensional lattice, as well as the narrowness of the distributions in (a), (c), and (d), due to the regularity of the periodic lattice. Similarly, 1D lattices show a constant (vertical) line and 3D lattices exhibit quadratic growth.

6. Lattice defects, dimensionality (since shells scale like the dimension −1), and "regularity" are all visible in the portrait. This is useful, since the change in edges

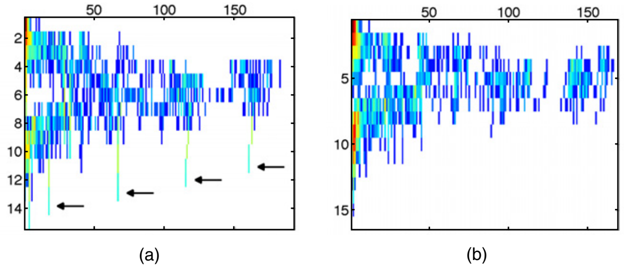


(a) (b)

(c) (d)

(e) (f)

Fig. 7: (Color online) Several real-world networks. (a) The western states power grid (unweighted) [1]; (b) US airlines network [15]; and directed metabolic networks for *H. influenzae*, *R. capsulatus*, *M. jannaschii*, and *C. elegans* [3], (c)–(f), respectively. The metabolic networks appear similar to one another yet unlike the power grid and airlines networks.



(a) (b)

Fig. 8: (Color online) (a) The original metabolic network of *M. genitalium* [3] with assortativity $A = -0.174216$ and (b) with $A = 0.000757$ after permuting random edge pairs while preserving the degree distribution. The fine-scale structure in the uppermost shells of (a) is no longer present in (b).

between a periodic and non-periodic lattice is small, though very specific, and this leads to massive change in the corresponding portraits.

7. Real world networks can give remarkably different portraits, but some classes of real-world networks can look similar (shown here with four metabolic networks in panels c–f). The four metabolic networks look quite similar despite widely varying scales in both axes. This suggests a simple scaling procedure; stretch one or both axes until the portraits overlap.

8. Correlation effects may still be visible in the higher rows of the portrait. Here is a highly disassortative metabolic network, note the vertical structures in the higher rows. Rewiring or perturbing this network to raise the assortativity destroys these structures.

Comparing portraits. – The portraits are useful for showing an intuitive picture of a network, but they can also be used quantitatively. A simple "distance" comparing networks G and G' may be defined, using their respective B-Matrices[1]. Motivated by the Kolmogorov-Smirnov (KS) test [16], we introduce the following statistic between corresponding pairs of rows B_ℓ and B'_ℓ:

$$K_\ell = \max_k \left| C_{\ell,k} - C'_{\ell,k} \right|, \tag{5}$$

where C is the matrix of cumulative distributions of B,

$$C_{\ell,k} = \left(\sum_{k' \leqslant k} B_{\ell,k'} \right) \Big/ \sum_{k'} B_{\ell,k'}. \tag{6}$$

The greater impact of lower shells on network properties (such as the average path length [17,18]) can be considered by assigning weights α_ℓ, based on shell "mass," for instance:

$$\alpha_\ell = \sum_k B_{\ell,k} + \sum_k B'_{\ell,k}. \tag{7}$$

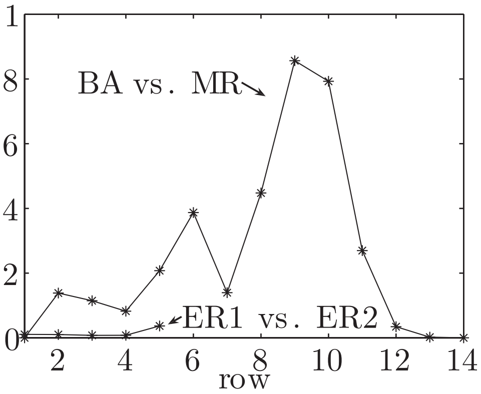
Finally, we choose a scalar distance Δ, generated by

$$\Delta(G, G') \equiv \Delta(B, B') = \left(\sum_\ell \alpha_\ell K_\ell \right) \Big/ \sum_\ell \alpha_\ell. \tag{8}$$

See fig. 9 for some concrete examples.

We apply this distance metric to four networks, summarized in fig. 9. Two Erdős-Rényi (ER) networks, with equal N and p, and a Barabási-Albert (BA) *vs.* a Molloy-Reed (MR) network built from the BA degree distribution. The plot indicates the value of the test statistic, eq. (5), while the table indicates the values of Δ, from eq. (8). The plot shows that the two ER networks agree very well with each other, while the BA and MR networks agree at first, but differences appear in higher rows (since BA has correlation effects missing in MR). The table values all agree with expectations: the ER graphs are very close to each other, the BA and MR graphs are farther apart from each other, and both BA and MR are very far from the ER networks.

[1] We assume that the networks are of comparable size. Empirically, the B-matrices may be scaled and normalized: $\{\ell, k, B\} \mapsto \{\ell/L, k/K, B/N\}$, where L and K are the largest shell number and largest degree (of any order), respectively. When the number of rows is small, one may first replace the B-matrix by a suitably smoothed surface (applying a spline procedure), then proceed with scaling.



	ER1	ER2	BA	MR
ER1	0	0.012	0.654	0.620
ER2	0.012	0	0.654	0.619
BA	0.654	0.654	0	0.232
MR	0.620	0.619	0.232	0

Fig. 9: Top: row-wise statistic K_ℓ: two ER graphs with $N = 10^4$ and $p = 0.002$; and a BA (diameter 10) *vs.* an MR network ($P(k) \sim k^{-3}$, diameter 14), both with $N = 5 \times 10^4$. Both the BA and MR networks have the same degree distribution, so the first rows agree. Differences in, *e.g.*, assortativity, soon become evident. Bottom: table containing the values of Δ, given by eq. (8), for the four networks. This table shows that the two ER graphs are very close to each, while the MR and BA graphs are somewhat far apart from each other and very far from the ER graphs, as expected.

Mathematically, it remains an open question if Δ is a metric or semi-metric (pseudometric). It is obvious from eqs. (5) and (7) that $\Delta(x, y) \geqslant 0$ and $\Delta(x, y) = \Delta(y, x)$. Furthermore, the numbers in fig. 9 satisfy the triangle inequality, but does this hold generally? The final issue at hand concerns indiscernibility, $\Delta(x, y) = 0 \iff x = y$. Discernibility in $\Delta(B, B')$ appears to hold, but there exist two non-isomorphic graphs, the dodecahedral and Desargues graphs, which have identical B's, disproving discernibility in $\Delta(G, G')$, if only because their B-matrices are indiscernible[2].

Conclusions and future work. – To summarize, B-matrices offer us an unambiguous way to visualize and discriminate between various complex networks. With little practice one can readily pick the patterns that distinguish one case from another: for example, the metabolic networks (fig. 7) have a distinctly similar appearance, with a prominent "knot" near the center of the portraits. Even small changes in structure induce visible changes in the B-matrix (figs. 3c,d, and 6d); the largest changes being induced by the removal or addition of links of highest *betweenness centrality* [19].

We have also introduced a distance, associated with the B-matrix, that quantifies the differences between complex networks. The distance between networks belonging to the same ensemble is small (figs. 3a,b, and 9), but it grows larger for networks in different ensembles (fig. 9).

Several generalizations come to mind. Equation (1) encompasses directed graphs and may be extended to weighted graphs: shells are defined by a set of weights $W = \{w_1, w_2, \ldots, w_d\}$ and could be found by Dijkstra's algorithm [20]. One may also generalize B to *edges* by defining the distance from a node v_i to an edge (v_j, v_k) as the mean of distances $d(v_i, v_j)$ and $d(v_i, v_k)$ (see footnote [3]). This "edges matrix" has half-integer rows with row 1/2 encoding the degree distribution, $B_{1/2,k} = NP(k)$, and so forth.

Among the most promising applications of B-matrices, besides identification and comparisons, is the question of the information content of complex graphs. The portraits can be compressed by applying conventional algorithms. The size of the compressed files could serve as a measure of information content (the difference in entropy of stochastic scale-free networks, *vs.* that of the highly ordered flower, in fig. 5, is apparent even visually).

Another interesting problem would be to use the "smoothness" of the matrices to create some quantitative measure of regularity, perhaps based on the variances of each row. This could also provide a useful measure of information content as well as symmetry and perhaps other characteristics. In several instances, we have discussed the "slope" of the matrix without giving specifics. While it is easy to identify dimensionality from the lattices of fig. 6, other networks are of higher dimension with broader row distributions and it is more difficult to pick out the slope visually. A specific fitting procedure or other technique may be useful.

Given the degree probability distribution (the first row of the B-matrix) there exist algorithms to construct complex networks that satisfy that degree distribution [9]. Perhaps the most important open question is the inverse of obtaining the B-matrices: Given a B-matrix, find a procedure to construct random complex nets belonging to the ensemble represented by it. This is related to the question of satisfiability constructing a random net that satisfies just the $P_1(k)$ degree distribution is already non-trivial [21], and this complicates as higher-order P_j's are added in. There exist already examples of procedures for obtaining maximally random nets with more than the $P_1(k)$ constraint, for example in [22] it is shown how to satisfy both the degree distribution and arbitrary degree-degree correlations.

Regarding the famous Graph Isomorphism problem, consider the non-isomorphic dodecahedral and Desargues graphs; both are cubic distance-regular with 20 nodes [23] and both have identical B-matrices[4], so B does not

[2] These graphs are both discussed in the final section.

[3] $B_{\ell,k}$ is now the number of nodes with k *edges* at distance ℓ.

[4] Distance-regular graphs will have exactly one nonzero element per row in B; in principle, this may be exploited to *search* for undiscovered distance-regular graphs by rewiring edges along some scheme to minimize the number of nonzero elements per row. This would likely be cost-prohibitive in practice.

uniquely encode a network. In practice, the probability of two large, non-isomorphic graphs chosen from a large ensemble having identical B-matrices appears to be vanishingly small, since the slightest difference will propagate throughout B. The dodecahedral and Desargues graphs are very similar in appearance, and the specific relationship between their edge sets that allows for identical B's is unlikely to arise at random. We propose that B is a "very good" answer to graph isomorphism. It is also worth noting that the Desargues and dodecahedral graphs have different edge matrices: we conjecture that graphs are uniquely identified with both matrices. The true power of B as a measure of graph isomorphism remains an open question and warrants further study.

Finally, it is worth noting that the construction of B requires an $\mathcal{O}(N^2)$ algorithm, which may preclude its use for extremely large networks. However, this algorithm is easily parallelized by spreading the starting nodes over multiple machines.

* * *

We thank H. Rozenfeld for discussions. We thank the NSF for support from a National Science Foundation Graduate Research Fellowship (JPB) and awards No. DMS-0404778 (EMB) and PHY0555312 (DbA).

REFERENCES

[1] Watts D. J. and Strogatz S. H., *Nature*, **393** (1998) 440.

[2] Amaral L. A. N., Scala A., Barthélémy M. and Stanley H. E., *Proc. Natl. Acad. Sci. U.S.A.*, **97** (2000) 11149.

[3] Jeong H., Tombor B., Albert R., Oltvai Z. N. and Barabasi A.-L., *Nature*, **407** (2000) 651.

[4] McKay B., *Congr. Numer.*, **30** (1981) 45.

[5] Many more B-matrices, and animations, are currently available at http://people.clarkson.edu/~qd00/.

[6] Billera L. J., Holmes S. P. and Vogtmann K., *Adv. Appl. Math.*, **27** (2001) 733 and references therein.

[7] Newman M. E. J., *Phys. Rev. Lett.*, **89** (2002) 208701.

[8] Cormen T. H., Leiserson C. E. and Rivest R. L., *Introduction to Algorithms* (The MIT Press, Cambridge, Mass.) 1990.

[9] Molloy M. and Reed B., *Comb. Probab. Comput.*, **7** (1998) 295.

[10] Erdős P. and Rényi A., *Publ. Math.*, **6** (1959) 290.

[11] Newman M. E. J. and Watts D. J., *Phys. Rev. E*, **60** (1999) 7332.

[12] Krapivsky P. L. and Redner S., *Phys. Rev. E*, **63** (2001) 066123.

[13] Barabasi A.-L. and Albert R., *Science*, **286** (1999) 509.

[14] Rozenfeld H. D., Havlin S. and ben-Avraham D., *New J. Phys.*, **9** (2007) 175.

[15] Batagelj V. and Mrvar A., *Pajek datasets* (2006), URL http://vlado.fmf.uni-lj.si/pub/networks/data/.

[16] Conover W. J., *Practical Nonparametric Statistics* (Wiley) 1998.

[17] Newman M. E. J., Strogatz S. H. and Watts D. J., *Phys. Rev. E*, **64** (2001) 026118.

[18] Dorogovtsev S. N., Mendes J. F. F. and Samukhin A. N., *Nucl. Phys. B*, **653** (2003) 307.

[19] Girvan M. and Newman M. E. J., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 7821.

[20] Dijkstra E. W., *Numer. Math.*, **1** (1959) 269.

[21] Molloy M. and Reed B., *Random Struct. Algorithms*, **6** (1995) 161179.

[22] Weber S. and Porto M., *Phys. Rev. E*, **76** (2007) 046111.

[23] Brouwer A. E., Cohen A. M. and Neumaier A., *Distance-Regular Graphs* (Springer-Verlag, New York) 1989.