Weighted Percolation on Directed Networks

Juan G. Restrepo,[1,*] Edward Ott,[2] and Brian R. Hunt[2]
[1]Northeastern University, Boston, Massachusetts 02115, USA
[2]University of Maryland, College Park, Maryland 20742, USA
(Received 4 April 2007; published 4 February 2008)

We present and numerically test an analysis of the percolation transition for general node removal strategies valid for locally treelike directed networks. On the basis of heuristic arguments we predict that, if the probability of removing node i is p_i, the network disintegrates if p_i is such that the largest eigenvalue of the matrix with entries $A_{ij}(1 - p_i)$ is less than 1, where A is the adjacency matrix of the network. The knowledge or applicability of a Markov network model is not required by our theory, thus making it applicable to situations not covered by previous works.

DOI: 10.1103/PhysRevLett.100.058701 PACS numbers: 89.75.Hc, 05.70.Ln, 87.23.Ge

There has been much recent interest in the structure and function of complex networks [1]. One aspect that has received considerable attention is the resilience of networks to the removal of some of their nodes [2–11]. This problem is related, for example, to the robustness of transportation and information networks to disturbances like random failures or targeted attacks, or to the resistance of biological networks to the action of drugs. Another related problem is determining the threshold for epidemic spreading [12]. An important objective is to determine what characteristics of the structure of the network determine the proportion of nodes that can be removed before the network disintegrates.

A model often considered is one in which nodes are removed from a network of size N with a uniform probability p. In the large N limit, for probabilities less than a critical value p_c, there is a connected component of the network of size of order N (the *giant component*), and for values of p larger than p_c, there is no connected component of size of order N [1,3–8,10]. The critical probability p_c at which this *percolation* transition occurs has been the subject of several theoretical works.

In what follows we define the in and out degrees of a node i by $d_i^{\text{out}} = \sum_{j=1}^{N} A_{ij}$ and $d_i^{\text{in}} = \sum_{j=1}^{N} A_{ji}$. Here A_{ij} is the network adjacency matrix; $A_{ij} = 1$ if there is a directed link from i to j and 0 otherwise. If $A = A^T$ the network is said to be undirected and $d_i^{\text{out}} = d_i^{\text{in}} = d_i$. For undirected, degree uncorrelated networks (the number of connections per node for neighboring nodes is not correlated), Cohen *et al.* have shown [4] that the critical probability is approximately given by $(1 - p_c)[(\langle d^2\rangle/\langle d\rangle) - 1] = 1$, where $\langle\cdot\rangle$ denotes an average over nodes. Reference [8] treats the case of undirected networks with correlations for *degree Markovian* networks, i.e., networks in which all nontrivial correlations are captured by the probability $P(d'|d)$ that a randomly chosen link from a node with degree d is connected to a node with degree d'. Other works on undirected networks have extended the Markovian approach to include the effect of clustering (e.g., often present in social networks); in particular, Ref. [11] presents an analysis for the case of weak clustering.

Reference [9] first studied the percolation transition in directed degree Markovian networks. The types of components studied are a *strongly connected component* (SCC), defined as a set of nodes such that every node in the SCC is reachable from any other node in the SCC by a directed path, its associated *in-component* (IN), defined as the set of nodes from which the SCC is reachable by a directed path, and *out-component* (OUT), defined as the set of nodes reachable from the SCC by a directed path. (There might be several such components.) Of interest is the largest strongly connected component which, if its size is of order N, is called the *giant strongly connected component*, GSCC. The out and in components of the GSCC are denoted GOUT and GIN.

It was found in Ref. [10] that as the probability of node removal p increases, GSCC, GOUT, and GIN disappear at the same critical value p_c. This value is determined by the largest eigenvalue of a matrix expressed in terms of $P_o(\mathbf{y}'|\mathbf{y})$ and $P_b(\mathbf{y}'|\mathbf{y})$ where $\mathbf{y} = (d_p^{\text{in}}, d_p^{\text{out}}, d_p^{\text{bi}})$, and d_p^{in}, d_p^{out} and d_p^{bi} are the number of incoming, outgoing and bidirectional edges for a given node. Here $P_o(\mathbf{y}'|\mathbf{y})$ and $P_b(\mathbf{y}'|\mathbf{y})$ are the probabilities of reaching a node of degree $\mathbf{y}'$ from a node of degree $\mathbf{y}$ by following an outgoing and a bidirectional edge, respectively.

One of our aims in this Letter is to remove the need for the applicability and knowledge of a Markov network model. In order to do so, we will focus on a class of directed networks that are locally treelike in the sense that they have few short loops [13]. More precisely, we assume that for each node i and not too large L, the number of *different* nodes reachable by paths of length L or less starting at node i is close to the *total* number of paths of length L or less starting from node i. In particular ($L = 2$) we assume that bidirectional edges are negligible in number. Under this assumption, $\mathbf{y} = (d^{\text{in}}, d^{\text{out}}, 0)$, and the matrix in Ref. [10] whose eigenvalue determines p_c reduces to

$$\hat{C}_{\mathbf{z}\mathbf{z}'} = (d^{\text{out}})' P(\mathbf{z}'|\mathbf{z}), \quad (1)$$

where $\mathbf{z} = (d^{\text{in}}, d^{\text{out}})$. We note that our locally treelike condition for directed networks is analogous to assuming negligible clustering.

In many situations the node removal probability is not a constant. For example, airports might have different security measures, or differ in their vulnerability to an attack or weather related shutdown due to their geographical location. Also, we have noted recently [14] that a measure of the dynamical importance of node i is proportional to $v_i u_i$, where u and v are the right and left eigenvectors corresponding to the largest eigenvalue λ of the adjacency matrix of the network, A: $Au = \lambda u$, $v^T A = \lambda v^T$. Thus, a potential removal strategy (to be used in an example later in this paper) is that in which node i is removed with a probability that depends on $v_i u_i$. More generally, we would like to study the effect of node removal strategies that assign a probability p_i to the removal of node i, and we refer to this problem as *weighted percolation*. (Other previously considered possibilities are that highly connected nodes are preferentially removed from the network, e.g., p_i is proportional to the degree of node i [5], or to a power of the degree of node i [6].)

Our objective is to present a simple heuristic method for treating *general* weighted percolation removal strategies (i.e., general p_i) on directed networks without the need for a Markovian network model. While we do not use the Markovian assumption nor a specific node removal probability, we assume a locally treelike network structure (we will discuss the validity of this assumption below), and we also require knowledge of the network adjacency matrix A. We find that the network disintegrates, as defined by the disappearance of the giant connected components, when the node removal strategy is such that the largest eigenvalue $\hat{\lambda}$ of the matrix $\hat{A}$ with entries $\hat{A}_{ij} = A_{ij}(1 - p_i)$ is less than 1.

To obtain the above result we adapt the mean field arguments given for example in [7,8] to our case. Consider first the disappearance of the giant in-component GIN. Let η_i be the probability that node i is not in the giant in-component GIN. Node i is not in GIN either if it has been removed (with probability p_i), or if it has not been removed and none of its out-links point to nodes in GIN. Consider two such nodes j_1 and j_2 that i points to. We argue that it is reasonable to make the approximation that whether j_1 belongs to GIN is independent of whether j_2 belongs to GIN. Whether j_1 is in GIN depends on whether the nodes it points to are in GIN, which depends on the nodes they point to, and so on. Our locally treelike assumption implies that the nodes that can be reached from j_1 by a short path are essentially independent of the nodes that can be reached from j_2 by a short path. Based on this independence assumption, η_i is given by (recall that $A_{ij} = 0$ or 1) $\eta_i = p_i + (1 - p_i)\prod_{j=1}^{N}(\eta_j)^{A_{ij}}$. This equation always has the trivial solution $\eta_i = 1$. The presence of a giant in-component requires a solution for which the expected size $s = N - \sum_{j=1}^{N}\eta_i$ is positive. Setting $\eta_i = e^{-z_i}$ and assuming $0 \le z_i$ and $\sum_{j=1}^{N} A_{ij} z_j \ll 1$, we obtain the approximation $z_i = \sum_{j=1}^{N} A_{ij}(1 - p_i) z_j$. When $p_i = 1$ (i.e., all nodes are removed), the only solution is the trivial solution $z_i = 0$. As we decrease the p_i's, a nontrivial solution (corresponding to a giant in-component) first appears when the largest eigenvalue $\hat{\lambda}$ of the matrix $\hat{A}$ with entries $\hat{A}_{ij} = A_{ij}(1 - p_i)$ is 1. Note that, as required, we can satisfy $\eta_i \le 1$ since the components z_i of the eigenvector corresponding to $\hat{\lambda}$ are, by the Frobenius theorem [15], nonnegative. Applying the same reasoning to the out-component GOUT we find that it appears when the largest eigenvalue of the matrix with entries $B_{ij} = A_{ji}(1 - p_j) = (\hat{A}^T)_{ij}$ is 1. Since the transpose of $\hat{A}$ and $\hat{A}$ have the same spectrum, the giant in and out-components appear simultaneously.

The above can also be understood by the following heuristic argument. Our previous discussion applies not only to GOUT (GIN), but more generally to sets generated by repeatedly following outgoing (incoming) links starting from a given node. Therefore, one can estimate the size of such sets and locate the transition as the point at which one of them has macroscopic size. In doing so, it is essential not to overcount the number of nodes, and it is here where our assumption of locally treelike network structure allows us to simplify the problem. The number of directed paths of length m starting from node i can be estimated using this assumption, for not too large m, as the sum of the components of the vector $\hat{A}^m e^i$, where e^i is the unit vector for coordinate i. If the largest eigenvalue of $\hat{A}$ is larger than 1, the number of paths of length m grows exponentially with m for some starting node i. Under our assumptions, these paths traverse different nodes, and thus the out component of i has large size, in agreement with our previous result.

In order to motivate the locally treelike assumption, we consider the relatively simple but illustrative case of uncorrelated networks. In particular, we estimate the fraction of bidirectional edges. The probability p_{ij} that nodes i and j share a bidirectional edge is given by $p_{ij} = d_i^{\text{out}} d_i^{\text{in}} d_j^{\text{in}} d_j^{\text{out}}/(N^2\langle d\rangle^2)$, where, since $\langle d^{\text{in}}\rangle = \langle d^{\text{out}}\rangle$, we use $\langle d\rangle$ to denote either of these averages. In the case that the degrees at nodes i and j are uncorrelated, on average, $\langle p_{ij}\rangle = (\langle d_i^{\text{out}} d_i^{\text{in}}\rangle/(N\langle d\rangle))^2 \approx (\lambda/N)^2$, where we have used the mean field approximation for the maximum eigenvalue λ of A (see below). Not too far above the percolation transition λ is of order 1, and we thus expect the locally treelike assumption to be valid.

We next discuss how our results compare to those of Ref. [10] in the case of negligibly few bidirectional edges [i.e., Eq. (1)]. If $p_i = p$, our result for the critical probability p_c reduces to $(1 - p_c)\lambda = 1$, where λ is the largest eigenvalue of the adjacency matrix A. If the network is degree Markovian, and we let $\psi_{\mathbf{z}}^{(m)}$ be the average number of directed paths of length m starting from nodes of degree $\mathbf{z}$, we have $\psi_{\mathbf{z}}^{(m+1)} = d^{\text{out}} \sum_{\mathbf{z}'} P(\mathbf{z}'|\mathbf{z}) \psi_{\mathbf{z}'}^{(m)}$. Since, for large

m, the number of paths of length m grows like λ^m, we associate to the previous equation the eigenvalue problem $\lambda_M \psi_{\mathbf{z}} = d^{\text{out}} \sum_{\mathbf{z}'} P(\mathbf{z}'|\mathbf{z}) \psi_{\mathbf{z}'}$, where λ_M is the *Markovian approximation* to λ. The previous result agrees with Eq. (1) [the matrices $d^{\text{out}} P(\mathbf{z}'|\mathbf{z})$ and $(d^{\text{out}})' P(\mathbf{z}'|\mathbf{z})$ have the same spectrum]. We note that, in the absence of degree-degree correlations, we have $P(\mathbf{z}'|\mathbf{z}) = d^{\text{in}} P(\mathbf{z}')/\langle d \rangle$, which yields the *mean field approximation* for the eigenvalue, $\lambda_{mf} = \langle d^{\text{in}} d^{\text{out}} \rangle / \langle d \rangle$. This agrees with the results in Ref. [10], where the effect of bidirectional edges is considered, and, for uncorrelated networks, a formula interpolating between λ_{mf} (when bidirectional edges are rare) and the undirected result $(\langle d^2 \rangle / \langle d \rangle) - 1$ is found. This supports our claim that, if short closed paths are rare, then λ should be a good approximation to the threshold.

We now illustrate our theory with two numerical examples and one real network. Example 1 illustrates the flexibility of our approach to address various weighted percolation node removal strategies, while example 2 illustrates the point that our approach does not require the knowledge or applicability of a Markov network model. We note that the networks in consideration are sparse, which allows us to use efficient techniques to compute the largest eigenvalue.

Example 1.—For simplicity, we consider uncorrelated random networks with degree distributions $P(d^{\text{in}}, d^{\text{out}})$ in which d^{in} and d^{out} are independent and have the same distribution $\tilde{P}(d)$, that is, $P(d^{\text{in}}, d^{\text{out}}) = \tilde{P}(d^{\text{in}}) \tilde{P}(d^{\text{out}})$. We use a generalization of the method in Ref. [16] in order to generate networks with a power law degree distribution, $\tilde{P}(d) \propto d^{-\gamma}$. We choose the sequence of expected degrees $\tilde{d}_i^{\text{in}} = c(i + i_0 - 1)^{-1/(\gamma-1)}$ for the in-degrees, and a random permutation of this sequence for the out-degrees, where $i = 1, \ldots, N$, and c and i_0 are chosen to obtain a desired maximum and average degree. Then, the adjacency matrix is constructed by setting $A_{ij} = 1$ for $i \neq j$ with probability $\tilde{d}_i^{\text{out}} \tilde{d}_j^{\text{in}} / (N \langle d \rangle)$ and zero otherwise ($A_{ii} = 0$). The ensemble expected value of the resulting network degree distribution is given by $P(d^{\text{in}}, d^{\text{out}})$. (Note that we assume $\tilde{d}_i^{\text{out}} \tilde{d}_j^{\text{in}} < N \langle d \rangle$.) In Fig. 1(a) we show, for a $N = 2000$ scale free network with exponent $\gamma = 2.5$ and $\langle d \rangle = 3$, the size of GIN as a function of the number of removed nodes R, when nodes are removed in order of decreasing $d_i^{\text{in}} d_i^{\text{out}}$ (thin solid line), decreasing dynamical importance [14] $(v_i u_i / v^T u)$ (dashed solid line), and randomly (thick line) [17]. The removal probabilities in the first two cases are given by $p_i = 1$ if $i \in S$ and 0 otherwise for a subset S of nodes. (For example, in the first case $S = \{i : d_i^{\text{in}} d_i^{\text{out}} > d_*^2\}$, and $\hat{A}$ reduces to the matrix obtained by removal of all nodes in A for which $d_i^{\text{in}} d_i^{\text{out}} > d_*^2$.) We also show in Fig. 1(b) the largest eigenvalue $\hat{\lambda}$ of $\hat{A}$ which in this case is equivalent to the adjacency matrix of the network resulting from the removal of the nodes. We observe for all three cases that the network disintegrates, as predicted, when $\hat{\lambda} = 1$ (indicated with the arrows in Fig. 1(a)].

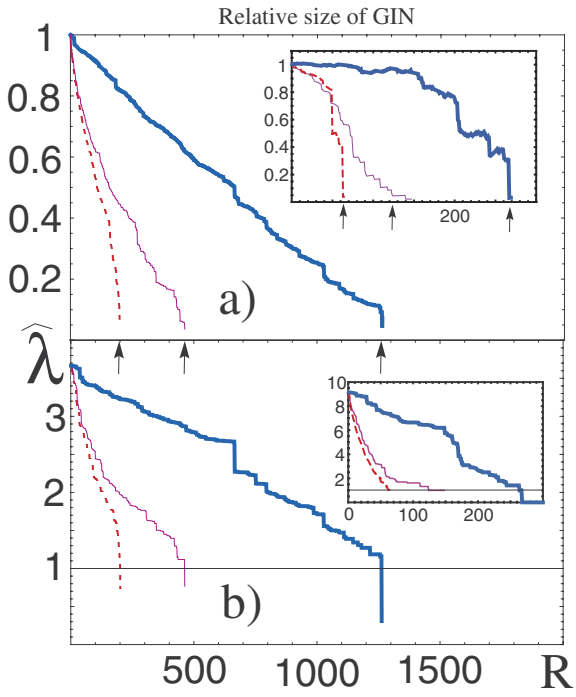

FIG. 1 (color online). (a) Ratio of the largest connected component to the number of remaining nodes, and (b) largest eigenvalue of the adjacency matrix of the network after node removal as a function of the number of removed nodes R, when nodes are removed randomly (thick solid line), and in order of decreasing $d_i^{\text{out}} d_i^{\text{in}}$ (thin solid line) and dynamical importance (dashed line) (see text). The percolation transition occurs when $\hat{\lambda} = 1$ [indicated by the vertical arrows in (a)]. The inset shows the same quantities for the neural network of *C. elegans*.

As an example of an application to a real network, we consider the neural network of *C. elegans* [18] ($N = 306$). The insets in Fig. 1(a) and 1(b) show the relative size of GIN and $\hat{\lambda}$ versus R for the three removal strategies, and we confirm that for this network the percolation transition occurs when $\hat{\lambda} = 1$ (vertical arrow). We remark that for some real networks, like the word association network in Ref. [19], the percolation transition occurs at a smaller fraction of removed nodes than the eigenvalue predicts, which suggests they are not locally treelike. In particular, word association networks are not expected to be locally treelike since one tends to associate certain groups of words densely with each other, establishing loops within the groups.

The simplest Markovian network assumption is that correlations between connected nodes depend solely on the degree of these nodes and not on any other property the nodes might have. While this is a useful analytical framework, probably applicable to many cases, it is also likely to fail in other cases (e.g., [20]). In our next example we will consider a network in which the node statistics depend on variables additional to the degrees and for which, therefore, a degree based Markovian approximation is not *a priori* expected to apply.

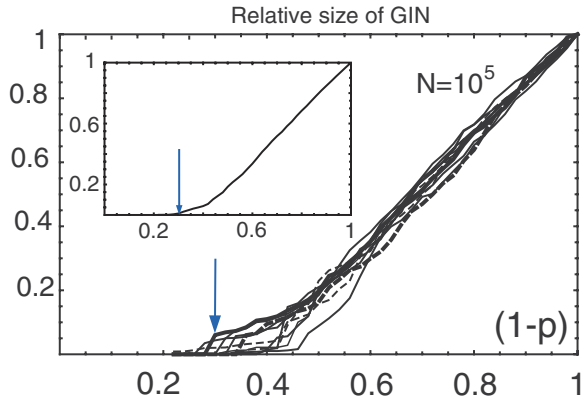

FIG. 2 (color online). Normalized size of the GIN versus $1 - p$, for example 2. The plot shows ten different realizations of the node removal process, and the inset shows their mean. The arrow on each plot shows the predicted percolation transition based on the eigenvalue approximation.

Example 2.—We start with a network generated as in example 1 with $N = 10^5$, $\gamma = 2.5$, and $\langle d \rangle = 3$. Then, we first specify a division of the nodes in the network into two groups of the same size, X and Y. We define a measure of the degree-degree correlations $\rho = \langle d_i^{\text{in}} d_j^{\text{out}} \rangle_e / \langle d_i \rangle_e^2$, where $\langle \ldots \rangle_e$ denotes an average over edges, $\langle Q_{ij} \rangle_e \equiv \sum_{i,j} A_{ij} Q_{ij} / \sum_{i,j} A_{ij}$. The following (an adaptation of the method in Ref. [21]) is repeated until the network has the desired amount of degree-degree correlations as evidenced in the value of ρ: Two edges are chosen at random, say connecting node i to node j and node n to node m. If i, j, n, m are all in X and $(d_n^{\text{in}} d_m^{\text{out}} + d_i^{\text{in}} d_j^{\text{out}} - d_n^{\text{in}} d_j^{\text{out}} - d_i^{\text{in}} d_m^{\text{out}}) < 0$, the edges are replaced with two edges connecting node i to node m and node n to node j. Otherwise the edges are unchanged. By repeating this process for several steps one creates two subnetworks, X and Y, with different degree-degree correlations (here, 2.29 and 0.98, respectively). Starting from such a network, we successively remove a randomly chosen node and compute the size of the GIN relative to its initial size. In Fig. 2 we plot this normalized size of the GIN as a function of the fraction of remaining nodes $(1 - p)$ for ten realizations of the node removal sequence. Although the transition points of individual realizations have some spread, the arrow (predicted from the eigenvalue) gives a good approximation of their mean (see inset). Similar results for both examples 1 and 2 were obtained for tests using other values of the network parameters γ, $\langle d \rangle$, and N.

We now discuss the advantages and disadvantages of the eigenvalue approach when compared to the Markov approximation. As opposed to the Markov approximation, the eigenvalue approximation allows the easy treatment of general node removal strategies (''weighted percolation''). Furthermore, it does not require the assumption that the node correlations depend only on their degree and are only to nearest neighbors. In addition, the construction of the matrix $d^{\text{out}} P(\mathbf{z}'|\mathbf{z})$ and the determination of its largest eigenvalue is in some cases harder than the direct determination of the largest eigenvalue of the adjacency matrix A. On the other hand, in many cases the adjacency matrix of the network is not known, and local sampling methods from which an approximation to the matrix $d^{\text{out}} P(\mathbf{z}'|\mathbf{z})$ can be constructed must be used. Additionally, the eigenvalue approach is valid only when the network has locally treelike structure. As such, our method should be viewed as complementary to the Markov approach.

This work was supported by ONR (Physics), by the NSF (No. PHY 0456240 and No. DMS 0104-087), and by USAFOSR No. FA95500410319.

*juanga@neu.edu

[1] M. E. J. Newman, SIAM Rev. **45**, 167 (2003); A.-L. Barabási and R. Albert, Rev. Mod. Phys. **74**, 47 (2002); S. Boccaletti *et al.*, Phys. Rep. **424**, 175 (2006).
[2] R. Albert *et al.*, Nature (London) **406**, 378 (2000).
[3] D. S. Callaway *et al.*, Phys. Rev. Lett. **85**, 5468 (2000).
[4] R. Cohen *et al.*, Phys. Rev. Lett. **85**, 4626 (2000).
[5] R. Cohen *et al.*, Phys. Rev. Lett. **91**, 247901 (2003).
[6] L. K. Gallos *et al.*, Phys. Rev. Lett. **94**, 188701 (2005).
[7] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[8] A. Vázquez and Y. Moreno, Phys. Rev. E **67**, 015101(R) (2003).
[9] N. Schwartz *et al.*, Phys. Rev. E **66**, 015104(R) (2002).
[10] M. Boguñá and M. A. Serrano, Phys. Rev. E **72**, 016106 (2005).
[11] M. A. Serrano and M. Boguñá, Phys. Rev. E **74**, 056115 (2006).
[12] P. Grassberger, Math. Biosci. **63**, 157 (1982). This Letter gives a mapping of the epidemic problem onto the percolation problem.
[13] In a recent independent work of B. Bollobás *et al.*the percolation threshhold was related to the largest eigenvalue of the adjacency matrix [arXiv.math/0701346]. However, our work and that of Bollobás *et al.* address very different circumstances. Bollobás *et al.* consider dense *undirected* networks for $N \gg 1$, $\lambda \gg 1$ and they show that $(1 - p)\lambda = 1 + o(1)$ asymptotically for large N. In this case the contribution of short closed paths can be neglected as it is $o(1)$. In contrast, for our problem, a sharper estimate is required, and we thus require short closed paths to be absent or sufficiently small in number.
[14] J. G. Restrepo *et al.*, Phys. Rev. Lett. **97**, 094102 (2006).
[15] C. R. MacCluer, SIAM Rev. **42**, 487 (2000).
[16] F. Chung *et al.*, Proc. Natl. Acad. Sci. U.S.A. **100**, 6313 (2003).
[17] Note that the first two removal criteria are related in that $d_i^{\text{out}} d_i^{\text{in}} / \sum (d_i^{\text{out}} d_i^{\text{in}})$ can be viewed as a mean field prediction of the dynamical importance.
[18] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[19] D. L. Nelson *et al.*, The University of South Florida word association, rhyme, and word fragment norms, http://www.usf.edu/FreeAssociation/, 1998.
[20] J. W. Kim *et al.*, Phys. Rev. E **66**, 046115 (2002).
[21] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).