Exploring complex networks through random walks

Luciano da Fontoura Costa* and Gonzalo Travieso†
Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, São Carlos, Sao Paulo, 13560-970, Brazil
(Received 15 July 2006; revised manuscript received 9 November 2006; published 11 January 2007)

Most real complex networks—such as protein interactions, social contacts, and the Internet—are only partially known and available to us. While the process of exploring such networks in many cases resembles a random walk, it becomes a key issue to investigate and characterize how effectively the nodes and edges of such networks can be covered by different strategies. At the same time, it is critically important to infer how well can topological measurements such as the average node degree and average clustering coefficient be estimated during such network explorations. The present article addresses these problems by considering random, Barabási-Albert (BA), and geographical network models with varying connectivity explored by three types of random walks: traditional, preferential to untracked edges, and preferential to unvisited nodes. A series of relevant results are obtained, including the fact that networks of the three studied models with the same size and average node degree allow similar node and edge coverage efficiency, the identification of linear scaling with the size of the network of the random walk step at which a given percentage of the nodes/edges is covered, and the critical result that the estimation of the averaged node degree and clustering coefficient by random walks on BA networks often leads to heavily biased results. Many are the theoretical and practical implications of such results.

DOI: 10.1103/PhysRevE.75.016102 PACS number(s): 89.75.Hc, 05.40.Fb, 07.05.Mh

I. INTRODUCTION

Despite its relatively young age, the area of investigation going by the name of *complex networks* [1–5] has established itself as a worthy relative—or perhaps inheritor—of graph theory and statistical physics. Such a success has been a direct consequence of the emphasis which has been given to structured interconnectivity, statistical formulations, interest in applications and, as in more recent developments (e.g., [3,4]), the all-important paradigm relating structure and dynamics. Yet, very frequently, the analyzed networks are assumed to be completely known and accessible to us. Indeed, while so many important problems involving completely described networks—such as community finding (e.g., [6])—remain as challenges in this area, why should one bother to consider incompletely specified networks?

Perhaps a good way to start making sense of this question is by considering our future. To what restaurant are we going tomorrow? What article will we read next? Which mirrors will ever see our faces again? Would not each such situation be describable as a node, while the flow of decisions among the possibilities would define a most extraordinary personal random walking a most complex network? Although such a dynamic network is undoubtedly out there (or in here), we are allowed to explore just a small portion of it at a time. And, with basis on whatever knowledge we can draw from such a small sample, we have to decide about critical next steps. However, the situations involving incomplete or sampled networks extend much further than this extreme example. For instance, the steps along any game or maze is but a sample of a much larger network of possibilities. Explorations of land, sea, and space also correspond to small samplings of a universe of possibilities, not to mention more "classical" large networks such as those obtained for protein interaction, social contacts, and the Internet. Last but not least, the own exploratory activities of science are but a most complex random walk on the intricate and infinite web of knowledge [7]. In all such cases, the success of the whole enterprise is critically connected to the quality and accuracy of the information we can infer about the properties of the whole network while judging from just a small sample of it. Little doubt can be raised about the importance of such a problem, which has received relatively little attention. Among the previously reported related works we have the investigation of random walks/diffusion in scale free networks with quenched disorder [8], the analysis of the effects of sampling the World Wide Web (WWW) through crawlers [9], and the investigation of tracerouter probes for sampling the Internet [10]. The problem of sampling networks has also been addressed from the sociological point of view (see [11–13]). The use of random walks for the self-organization of network growth has been considered in [14]. The literature about random walks in complex networks include [15–25]. The subject of general sampled networks has also been covered in [26,27]. Some systematic investigations of random walks as the means to sample complex networks have been reported recently [12,28–30]. Reference [28] considered three ways of sampling the networks (i.e., node, link, and snowball sampling), identified a bias in the estimation of several measurements and suggested means to avoid such problems.

The current paper is about incomplete and sampled networks and some related fundamental questions. We start with the basic mathematical concepts, identifying some of the most interesting related questions and perspectives, and proceed by illustrating what can be learned about random, Barabási-Albert (BA), and geographical networks while sampling them locally in random fashion or through three types of random walks—traditional (uniform decision prob-

*Electronic address: luciano@ifsc.usp.br
†Electronic address: gonzalo@ifsc.usp.br

ability), preferential to untracked edges, and preferential to untracked nodes. Particularly, the choice of these three complex networks models provide a reasonable diversity of connectivities, varying from completely indiscriminate in the case of the random models to the geographically structured and strongly regular connections found in the geographical model, with the Barabási-Albert networks representing structures involving hubs. In this way, the current investigation can be understood as an extension of previous works (especially [12,28]), in light of the sampling schemes and concepts of node/edge coverage suggested recently in [7].

II. BASIC CONCEPTS AND SOME FUNDAMENTAL ISSUES

An undirected [32] complex network $\Gamma=(V,E)$, involving a set of N nodes V and a set E of connections between such nodes.

An *incompletely specified complex network* is henceforth understood as any subnetwork G of Γ such that $G \neq \Gamma$. In this work we will restrict our attention to incomplete complex networks defined by sets of nodes and adjacent edges identified during random walks. Such networks can be represented as $((i_1,A_1);(i_2,A_2),\ldots,(i_M,A_M))$, where i_p are nodes sampled during the random walk through Γ, and A_p are sets containing the respective list of adjacent nodes. Note that necessarily $i_{p+1} \in A_p$ and that $(i_1,i_2,\ldots,i_M)$ corresponds to a *path* along Γ. It is also interesting to consider more substantial samples of Γ, for instance by considering not only the adjacent edges, but also the interconnections between the neighboring nodes of each node. Therefore, the case above becomes $((i_1,A_1,E_1);(i_2,A_2,E_2),\ldots,(i_M,A_M,E_M))$, where E_p is the set containing the edges between the nodes in A_p. Figure 1 illustrates a complex network (a), and respective examples of incompletely specified networks obtained by random walks considering neighboring nodes (b), and the latter plus the edges between neighboring nodes (c).

Given an incompletely specified complex network G, a natural question which arises is: to what accuracy the properties of the whole network Γ can be inferred from the available sampled information? Because the estimation of global properties of Γ such as shortest paths and diameter constitutes a bigger challenge to the moving agent, we concentrate our attention on *local* topological properties, more specifically node degree and clustering coefficient of visited nodes. The degree k_i of node i and the clustering coefficient C_i of that node are calculated as described in [3].

Three types of random walks are considered in the present work: (i) "*traditional*:" the next edge is chosen with uniform probability among the adjacent edges; (ii) *preferential to untracked edges*: the next edge is chosen among the untracked adjacent edges and, in case no such edges exist, uniformly among all the adjacent edges; and (iii) *preferential to unvisited nodes*: the next edge is chosen among those adjacent edges leading to unvisited nodes and, in case no such edges exist, uniformly among all the adjacent edges. Note that the plausibility of the preferential schemes depends on each modeled system. For instance, the preference to untracked nodes implies that the moving agent knows whether each

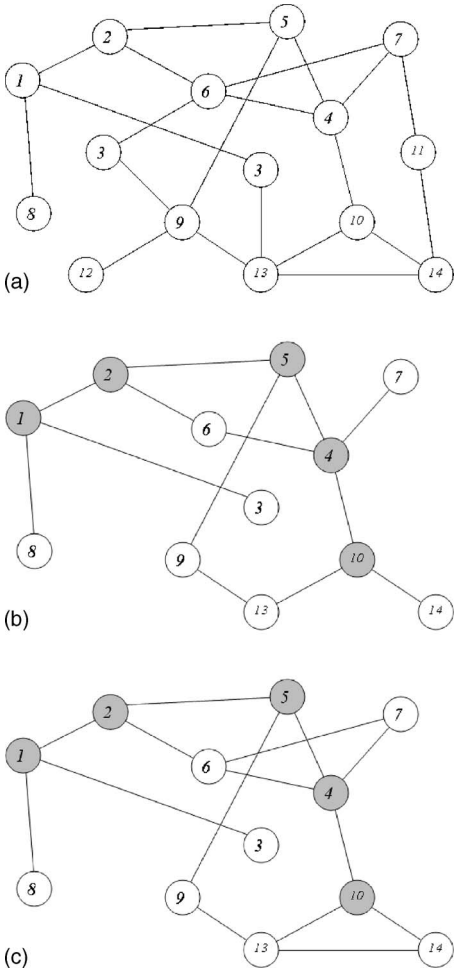

FIG. 1. A simple network (a) and two incompletely specified networks obtained by a random walk considering neighboring nodes and (b) the latter plus the edges between adjacent nodes (c). The gray nodes correspond to those sampled during the random walk.

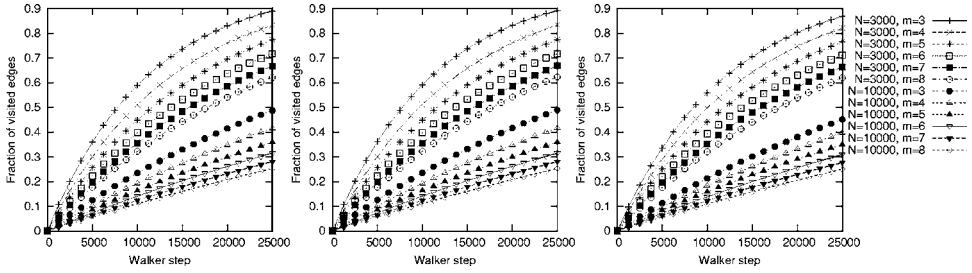

FIG. 2. The ratio of tracked edges in terms of the steps t for $N=3\,000$ and $N=10\,000$ considering the values of m as presented in the legend, for the random (left), BA (middle), and geographical (right) network models.

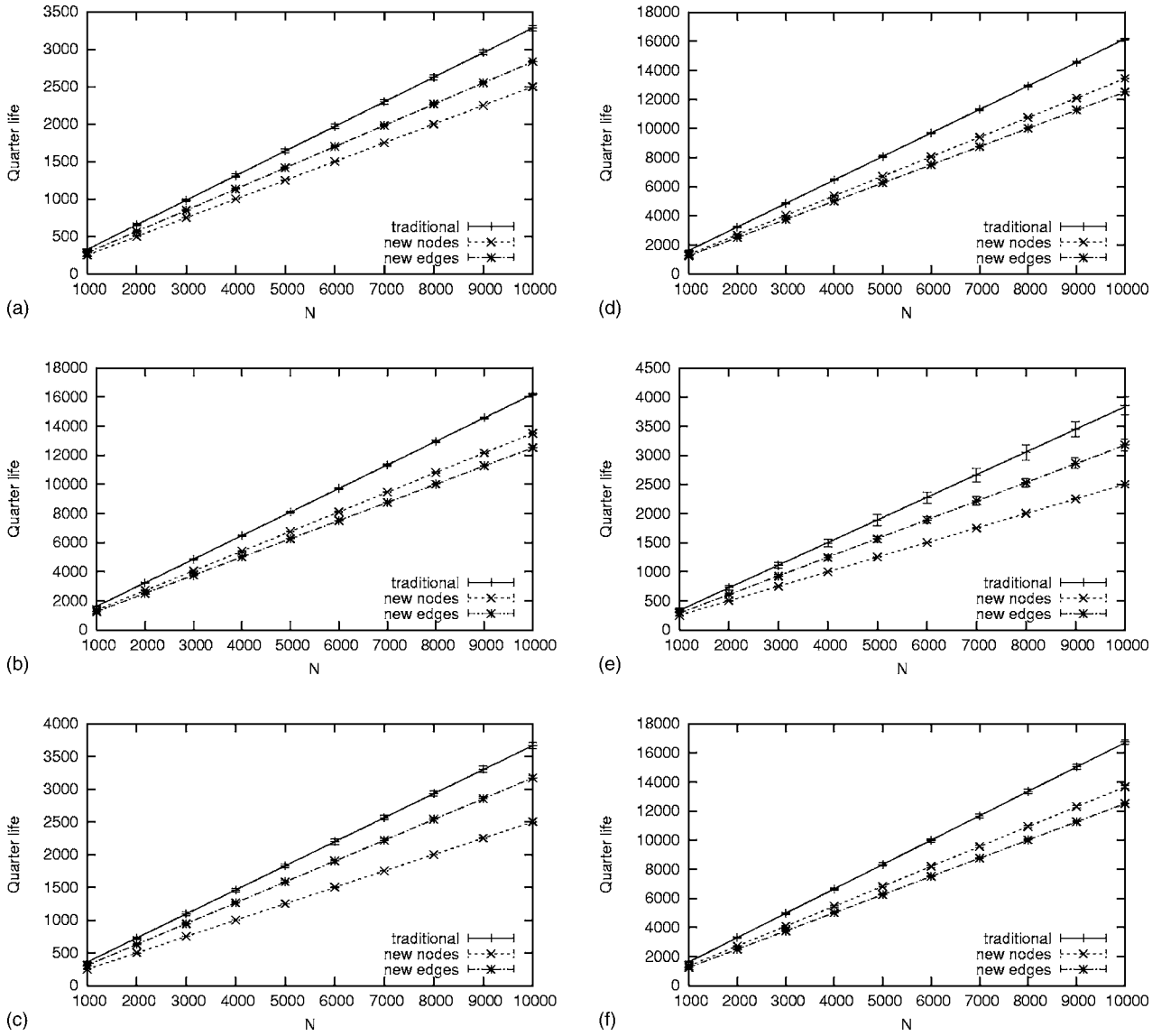

FIG. 3. The quarter-lives of the percentage of visited nodes (left column) and edges (right column) for the random (top), BA (middle), and geographical (bottom) models, for traditional, preferential to untracked nodes, and preferential to unvisited edges random walk strategies.

edge leads to a still unvisited node, though it may not know exactly which one. It is interesting to note that the process of sampling an existing network through a random walk can be interpreted as a mechanism for "growing" a network.

III. NODE AND EDGE COVERAGE

First we consider the following three complex network models: (a) *random (Erdős-Rényi) networks*; (b) *Barabási-Albert networks* (BA), built by using the preferential attachment scheme described in [1]; and (c) a geographical network model where nodes are distributed in a two-dimensional space and the connection probability decays with their distance. For all these network models, we fix the following two parameters: the number of nodes N and the number of edges for each node m, given thus an average degree of $\langle k\rangle = 2m$. The random networks are specified by the probability of connection of each pair of nodes, given by $\lambda = 2m/(N-1)$. In the BA networks, new nodes with m edges each are progressively incorporated into the network, with each of the m edges being attached to previous nodes with probability proportional to their respective node degrees; the network starts with $m_0 = m$ nodes. The geographical networks are constructed by uniformly distributing the N nodes in a two-dimensional square space of unitary length and linking each pair of nodes with probability $p = e^{-r/\rho}$, where r is the geographical distance between the nodes; the parameter ρ is adjusted for each pair of values N, m to achieve the desired average degree $\langle k\rangle = 2m$. Complex networks with number of nodes N equal to $1\,000, 2\,000, \ldots, 10\,000$ and $m = 3, 4, \ldots, 8$ have been considered. A total of 200 realizations of each configuration, for the three types of random walks, were simulated.

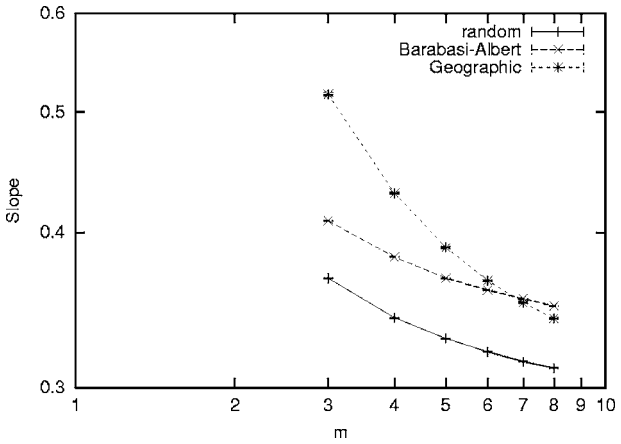

FIG. 4. Slopes of the ratios of visited nodes obtained for traditional random walks for $m=3,4,\ldots,8$ considering random, BA, and geographical network models (in logarithmic scale). Error bars show the asymptotic standard error of the regression.

Figure 2 illustrates the ratio of tracked edges in terms of the walker steps t for $N=3\,000$ and $N=10\,000$ considering $m=3,4,\ldots,8$. It is clear from the obtained results that, as expected, the higher the value of m, the smaller the ratio of visited edges. Note that the increase of N also contributes to less efficient coverage of the edges, as expressed by the respective smaller ratios of visited edges obtained for $N=10\,000$. For large enough total number of steps, all curves exhibited an almost linear initial region followed by saturation near the full ratio of visited edges (i.e., 1).

Figure 3(a) shows the "quarter-lives" h of the percentage of visited nodes in terms of the network size N with respect to the random network model with $m=5$, for the three types of random walk. This measurement corresponds to the average number of steps at which the random walk has covered a quarter of the total number of network nodes. Similar results have been obtained for other critical fractions (e.g., half-life). Results for the BA model are shown in Fig. 3(c) and for the geographical model in 3(e). Note that, as m is fixed at 5, the average degree $\langle k\rangle$ of all networks in this figure remains equal to 10, being therefore constant with N, while the average number of edges grows as $\langle E\rangle=N\langle k\rangle/2=5N$. Interestingly, linear dependence between the quarter-lives and N are obtained in all cases. It is also clear from these results that the most effective coverage of the nodes is obtained by the random walk preferential to unvisited nodes, with the random walk preferential to untracked edges presenting the next best performance. The quarter-lives for the percentage of tracked edges are shown in Figures 3(b), 3(d), and 3(f) respectively to random, BA and geographical network models. The best ratios of covered edges were obtained for the random walk preferential to untracked edges, with the random walk preferential to unvisited nodes presenting the next best performance. The traditional random walk resulted the least efficient strategy in all situations considered in this work. Note that the three types of random walks have similar edge coverage efficiencies on the three network models. For node coverage, there are slight differences from one network model to the other, with faster coverage in the random network model and slower coverage in the geographical model. The model has greater influence on node coverage for the traditional random walk, with almost no influence for the random walk preferential to new nodes.

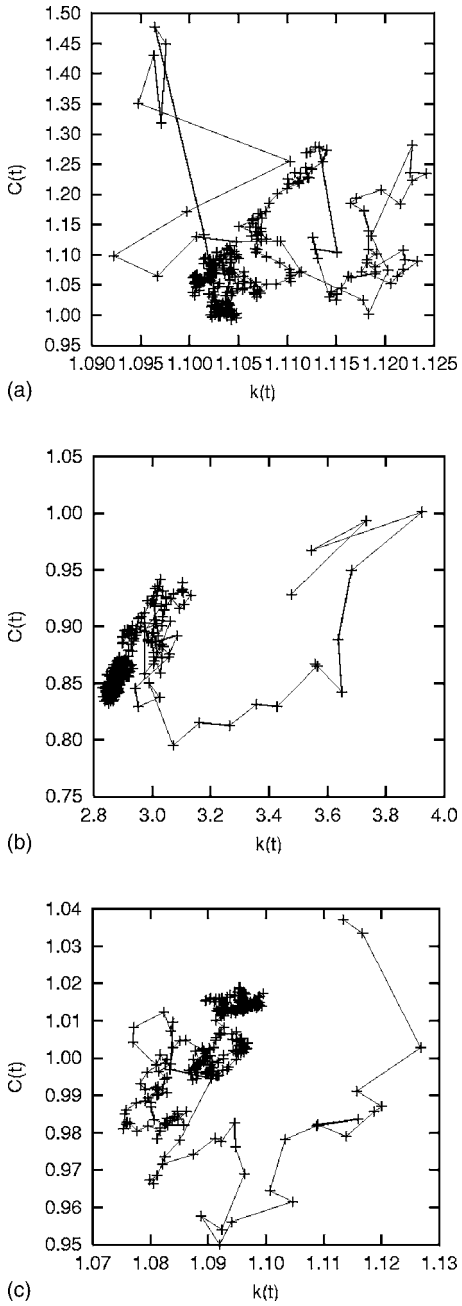

FIG. 5. Curve (actually a kind of random walk) defined by the estimations $(k(t),C(t))$, through traditional random walk, of the average node degree $k(t)$ and average clustering coefficient $C(t)$ in a random (a), BA (b), and geographical (c) network with $N=10\,000$ and $m=5$.

Further characterization of the dynamics of node coverage can be obtained by considering the scaling of the slopes of the curves of ratios of visited nodes in terms of several values of m. Remarkably, for random walks preferential to new nodes, the slopes obtained by least mean square fitting were verified not to vary significantly with m, being fixed at about

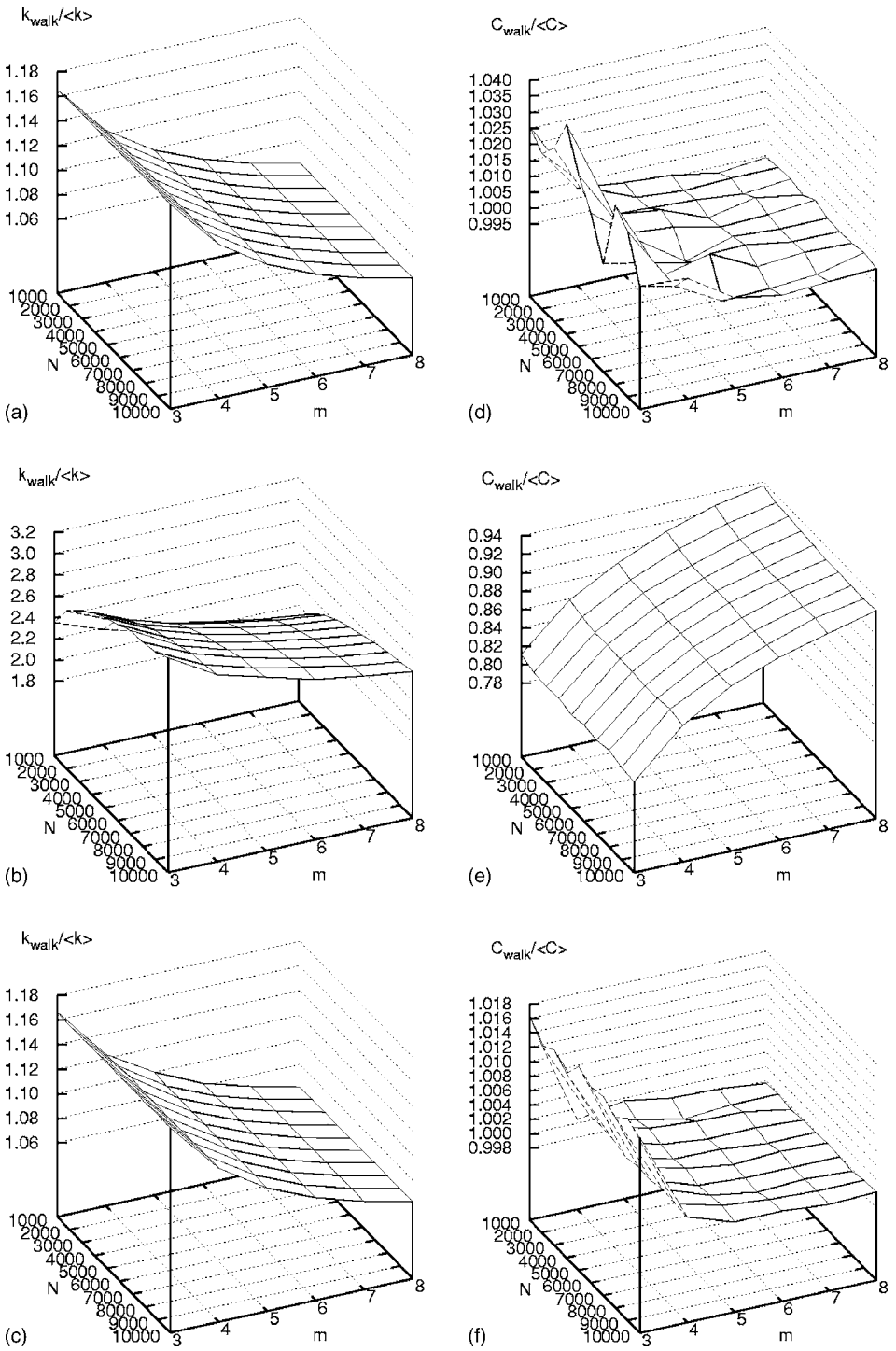

FIG. 6. Average degree (a), (b), and (c); and average clustering coefficient (d), (e), and (f); for the Erdős-Rényi (a), and (d); Barabási-Albert (b), and (e), and geographical (c), and (f) network models. The values shown are ratios from averages computed in the random walks (after 50 000 steps) to the real network values.

0.25 for the three network models. When preference is given for new links, the slopes are about 0.25 for the random model and 0.32 for the BA and geographical models, with slightly larger variation in the latter model. Figure 4 shows the log-log representation of the slopes in terms of m obtained for the traditional random walk for $m=3,4,\ldots,8$. The error bars (smaller than the size of the symbols) show the value found for the asymptotic standard error of the linear regression [31]. It is clear from this figure that, though the slopes tend to scale in similar fashion for the random and BA networks, node coverage is faster for the former. Geographical networks, on the other hand, have a markedly distinct behavior, with slower coverage and a faster decrease of the slopes with connectivity.

IV. ESTIMATION OF AVERAGE NODE DEGREE AND CLUSTERING COEFFICIENT

So far we have investigated the dynamics of node and edge coverage in random, BA and geographical models while considering the three types of random walks. In practice, as the size of the network being explored through the

random walks is typically unknown, the number of visited nodes or tracked edges by themselves provide little information about the topological properties or nature of the networks. The remainder of the present work addresses the estimation of measurements of the local connectivity of networks, namely the average node degree, average clustering coefficient, and degree distribution obtained along the random walks.

For generality's sake, the estimations for average degree and average clustering coefficient are henceforth presented in relative terms, i.e., as the ratio between the current estimation [e.g., $k(t)$, where t is the walker step] and the real value (e.g., $\langle k \rangle$). Figure 5 illustrates the curve defined by the estimations $(k(t), C(t))$ obtained by traditional random walks along network of the three models with $N=10\,000$ and $m=5$. Interestingly, these curves are indeed a kind of random walk with convergent limit. Such curves have been found to converge to limiting ratios (k_L, C_L) which can or not correspond to the ideal ratios (1, 1). In the case of the curve for the BA model in Fig. 5(b), we have $(k_L=2.87, C_L=0.84)$, i.e. the average node degree has been overestimated while the average clustering coefficient has been underestimated.

Through extensive simulations, reported in Fig. 6, we have observed that the estimations for the traditional random walk in random and geographical networks are independent of N, with slight overestimation of the average degree, but with the accuracy increasing with m, while better accuracies were achieved for the random model. For BA networks, the average degree is consistently overestimated, while the clustering coefficient is underestimated; the estimation accuracies decrease with N, but increase with m. The substantial biases implied by the random walk over BA networks is a direct consequence of the larger variability of node degree exhibited by this model (see also [12,28]). Therefore, nodes with higher degree will tend to be visited more frequently [33], implying overestimation of the average node degree and a slight bias on the clustering coefficient.

Further results on the estimation of the node degree distribution while proceeding along the path of a traditional random walk are shown in Fig. 7, which presents the degree distribution after 50 000 steps on networks of 10 000 nodes, as compared with the real degree distribution. It can be seen that the degree distributions estimated along the walk for the random and geographical networks are similar to the real distributions, with a slight bias toward higher degrees. On the other hand, for BA networks the slope of the power law changes.

Provided the moving agent can store all the information obtained from the network as it is being explored, yielding a partial map of the so far sampled structure, it is possible to obtain more accurate (i.e., unbiased) estimates of the average node degree and clustering coefficient during any of the considered random walks in any type of networks by performing the measurements without node repetition. However, an agent moving along a BA network without resource to such an up-to-date partial map will have to rely on averages of the measurements calculated at each step. This will cause the impression of inhabiting a network much more complex (in the sense of having higher average node degree) than it is indeed the case. Going back to the motivation at the beginning of this paper, it is difficult to avoid speculating whether our impression of living in a world with so many possibilities and complexities could not be in some way related to the above characterized effects.

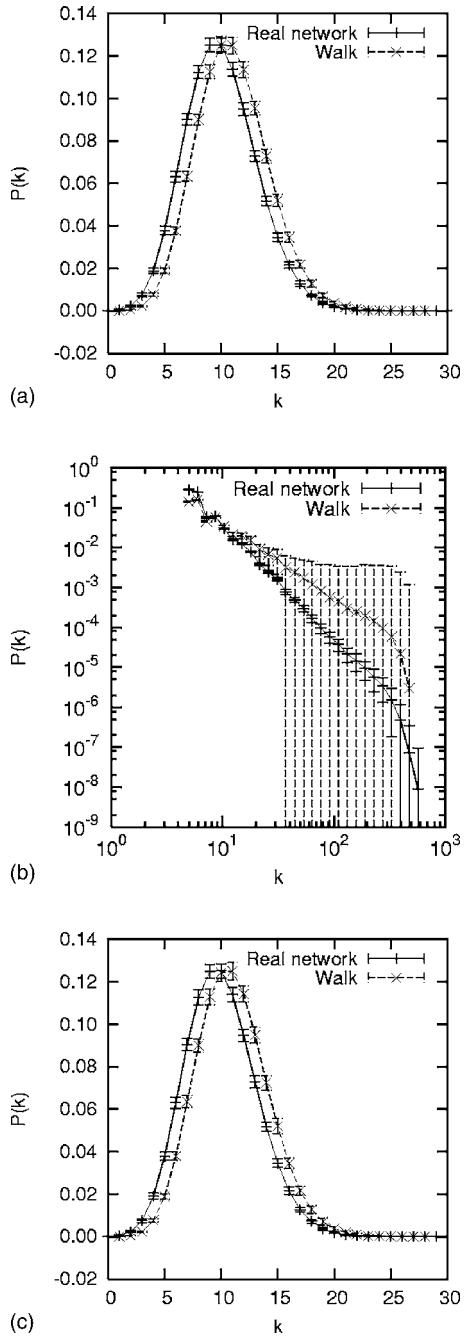

FIG. 7. Degree distributions for the random (a), BA (b), and geographical (c) networks, computed for the real network and for the nodes on the walk (after 50 000 steps). Note that the distribution of the BA case is shown in logarithmic axes.

V. CONCLUDING REMARKS

Due to the fact that most real complex networks are only partially available to us as a consequence of their sheer size

and complexity, it becomes of critical importance to understand how well these structures can be investigated by using sampling strategies such as different types of random walks. The present work has addressed this issue considering random, BA, and geographical network models with varying connectivity and sizes being sampled by three types of random walks. A series of results have been obtained which bear several theoretical and practical implications. Particularly surprising is the fact that all networks are similarly accessible as far as node and edge exploration is concerned. Actually, random networks tend to have their nodes and edges explored in a slightly more effective way, followed by the BA and geographical cases. Also important is the characterization of linear scaling with the network size of the quarter-life of the ratio of covered nodes and edges, and the identification of substantial biases in estimations of the average node degree and clustering coefficient in several situations. In particular, in the case of the BA model the average node degree tends to be estimated as being over twice as much as the real value, in accordance with previous related investigations [12,28]. In addition, our experiments allowed the identification of the fact that the node degree overestimation tends to increase with N and decrease with the average connectivity (m). Regarding the three different exploratory mechanisms, we found that the traditional approach (uniformly random selection of next node) resulted in the less efficient alternative. In the case of the node coverage, the strategy priorizing new nodes resulted in being the most effective. An analogue result was obtained for the case of edge coverage. The problem of recovering the degree distribution while performing the random walks was also considered and experimentally investigated. The estimated distribution was found to be similar to the real one in the cases of the random and geographical models. However, rather distinct slopes (considering log-log axes) were obtained for BA networks. Additional insights about the nontrivial dynamics of complex network exploration through random walks can be achieved by considering more global topological measurements such as shortest paths, diameters, hierarchical measurements, and betweenness centrality.

ACKNOWLEDGMENTS

This work was supported by FAPESP, under Grant No. 03/08269-7. Luciano da F. Costa is grateful to CNPq (308231/03-1) for financial support.

[1] R. Albert and A. L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[3] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chaves, and D. U. Hwang, Phys. Rep. **424**, 175 (2005).
[5] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, e-print cond-mat/0505185.
[6] M. E. J. Newman, Eur. Phys. J. B **38**, 321 (2004).
[7] L. da F. Costa, e-print physics/0601118.
[8] D. Stauffer and M. Sahimi, Phys. Rev. E **72**, 046128 (2001).
[9] M. A. Serrano, A. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani, cs.NI/0511035.
[10] L. Dall'Asta, E. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani, Theor. Comput. Sci. **355**, 6 (2006).
[11] M. Newman, Soc. Networks **25**, 83 (2001).
[12] M. J. Salganik and D. D. Heckathorn, Sociol. Methodol. **34**, 193 (2004).
[13] M. J. Salganik, P. D. Dodds, and D. J. Watts, Science **311**, 854 (2006).
[14] T. S. Evans and J. P. Saramäki, Phys. Rev. E **72**, 026138 (2005).
[15] B. Tadić, Eur. Phys. J. B **23**, 221 (2001).
[16] B. Tadić, in *Modeling of Complex Systems*, AIP Conf. Proc. No. 661 (AIP, New York, 2003), pp. 24–27.
[17] E. M. Bollt and D. ben Avraham, New J. Phys. **7**, 26 (2005).
[18] J. D. Noh and H. Rieger, Phys. Rev. E **69**, 036111 (2004).
[19] S. A. Pandit and R. E. Amritkar, Phys. Rev. E **63**, 041104 (2001).
[20] K. A. Eriksen, I. Simonsen, S. Maslov, and K. Sneppen, Phys. Rev. Lett. **90**, 148701 (2003).
[21] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).
[22] I. Simonsen, Physica A **357**, 317 (2005).
[23] D. Volchenkov and P. Blanchard, e-print physics/0608153.
[24] J. Candia, P. E. Parris, and V. M. Kenkre, e-print cond-mat/0608619.
[25] S. Fortunato and A. Flammini, e-print physics/0604203.
[26] M. P. H. Stumpf, C. Wiuf, and R. M. May, Proc. Natl. Acad. Sci. U.S.A. **102**, 4221 (2005).
[27] M. P. H. Stumpf and C. Wiuf, Phys. Rev. E **72**, 036118 (2005).
[28] S. H. Lee, P.-J. Kim, and H. Jeong, Phys. Rev. E **73**, 016102 (2006).
[29] S.-J. Yang, Phys. Rev. E **71**, 016107 (2005).
[30] A. Ramezanpour, e-print cond-mat/0607327.
[31] M. Ledvij, Ind. Phys. **9**, 24 (2003).
[32] The results in this article are immediately extended to more general networks, including directed and weighted models.
[33] Actually the rate of visits to the nodes of an undirected complex network, at thermodynamical equilibrium, can be verified to be proportional to the node degree.