Measuring the Sensitivity of Graph Metrics to Missing Data

Anita Zakrzewska(✉) and David A. Bader

Georgia Institute of Technology, Atlanta, GA, USA
azakrzewska3@gatech.edu

Abstract. The increasing energy consumption of high performance computing has resulted in rising operational and environmental costs. Therefore, reducing the energy consumption of computation is an emerging area of interest. We study the approach of data sampling to reduce the energy costs of sparse graph algorithms. The resulting error levels for several graph metrics are measured to analyze the trade-off between energy consumption reduction and error. The three types of graphs studied, real graphs, synthetic random graphs, and synthetic small-world graphs, each show distinct behavior. Across all graphs, the error cost is initially relatively low. For example, four of the five real graphs studied needed less than a third of total energy to retain a degree centrality rank correlation coefficient of 0.85 when random vertices were removed. However, the error incurred for further energy reduction grows at an increasing rate, providing diminishing returns.

Keywords: Graphs · Graph algorithms · Sensitivity analysis · Missing data · Energy consumption · Power

1 Introduction

Power consumption has become a critical issue in computing. This is a concern both for supercomputers, where massive energy use poses a financial and an environmental cost, and for embedded in-the-field processing systems, which have a limited energy supply or battery lifetime. Achieving maximum computational capabilities on embedded systems while limiting power use is an important task.

We address energy reduction for irregular, sparse graph algorithms through data sampling or removal. Sparse networks often represent relationships, communication, or information flow. For example, a graph may represent an online social network, network traffic, biological networks, or financial transactions. Often such graphs are constructed from a massive, and constant, stream of data, which leads to large graphs and energy-expensive computations. However, in cases where an approximate solution suffices, it is not always necessary to store and use the entire graph. For example, when calculating distances, approximate results for shortest paths may be acceptable for a given application. If the goal

R. Wyrzykowski et al. (Eds.): PPAM 2013, Part I, LNCS 8384, pp. 783–792, 2014.
DOI: 10.1007/978-3-642-55224-3_73,

is to find the most influential, or important, vertices, it is only necessary to calculate top scores correctly since low-scoring vertices are of no interest. Since approximations are often satisfactory for real-world graph metrics, a certain degree of error in the underlying graph data, such as missing or incorrect edges and vertices, may be tolerated. Real-time streams of data may also amass too much information to be stored or lead to over-saturation, in which case certain vertices and edges may need to be removed over time.

Vertex and edge removal can also be performed intentionally with the goal of reducing energy consumption. Sampling results in a smaller graph, with fewer memory accesses, fewer compute operations, and a shorter overall running time, all of which contribute to less energy use. However, in order for this to be a feasible approach, it is necessary to determine the resulting level of error. We investigate the sensitivity of several graph metrics to missing vertices and edges, which can be used to set tolerable error level thresholds.

Previous work has compared the sensitivity of scale free and random networks to vertex removal [1]. Sampling and contraction methods have been used to reduce the size of internet topology graphs [13]. Graph analytic sensitivity to noisy data has been addressed by Borgatti *et al.* [5]. However, that work only considers vertex centrality measures on Erdős-Rényi random graphs [9], whose structure differs from that of real networks. Because the authors focus on errors in the data due to noise instead of conscious data sampling for power reduction, many of the errors analyzed, such as false positive edges, are not as applicable to the goal of energy reduction. Kossinets [12] studies the effects of missing data in social networks by analyzing a bipartite scientific collaboration network of authors and papers as well as bipartite random graphs. Our work differs because we focus on filtering methods for the purpose of decreasing the size of the graph and therefore the energy needed to compute various analytics.

2 Energy Model

The energy consumption of an algorithm can be modeled in terms of the energy per memory operation, energy per arithmetic operation, and constant energy that must be expended until the computation terminates, as given in Eq. (1), where W is the number of memory operations, ϵ_{flop} is the fixed energy cost of a compute operation, Q is the number of arithmetic operations, ϵ_{mem} is the fixed energy cost of a memory operation, T is the duration of the algorithm, and π_0 is the fixed constant energy cost, which may be idle energy or leakage [7,11].

$$E = W\epsilon_{flop} + Q\epsilon_{mem} + T\pi_0 \tag{1}$$

Because sparse graph algorithms tend to exhibit a low arithmetic intensity and are memory bound,we focus on the number of memory operations and energy per memory operation. Many real-world graphs have a low diameter and irregular structure with little or no locality in the data access pattern. Sparse graph algorithms tend to exhibit low data reuse and focus on traversing the graph

structure [15]. Therefore, focusing on memory cost is an appropriate proxy for the energy consumption of sparse graph algorithms.

Dynamic power management is a technique used to reduce power consumption in which system components are switched to a low-performance, or idle, state when load demands are low [4]. Memory power reduction can be achieved by dynamically adjusting memory voltage and frequency based on bandwidth utilization [8]. We describe three possible situations in which energy considerations can cause analytics to be run on incomplete graphs. From the algorithmic perspective, all have the same result. A subset of the vertices and edges of a graph are not used when calculating a graph analytic.

1. The system may choose not to access a subset of the graph in memory. This reduces the number of memory accesses.
2. Portions of memory may be turned to a low power mode to conserve energy, resulting in some data being unavailable. In-the-field embedded systems, for example, may do this after having detected low energy supplies.
3. The system may have insufficient storage for the entire graph and so a subset of the graph must be removed or never stored in the first place.

3 Methodology

Our approach to measuring sensitivity to missing data is as follows. We start with a true, base graph G and compute the value of a metric, called the true metric value. For each sampling level k, the graph is sampled in several ways and a subset of the vertices and edges is removed, creating the sampled graph $G_{k,sampled}$. The metric is recomputed on $G_{k,sampled}$, which gives the observed metric value. We then compare the true metric value to the observed metric value, resulting in a metric error. The energy required is calculated as the ratio of energy needed for $G_{k,sampled}$ to the energy needed for G. The relationship between the average metric error and energy required for each sampling level can then be examined. This process is repeated for all sampling methods described in Sect. 3.2.

3.1 Datasets

Testing is performed on both real and synthetic networks, listed in Table 1. The real graphs come from the 10^{th} DIMACS Implementation Challenge [2] and include citation networks, collaboration networks, a graph of users of the Pretty-Good-Privacy algorithm for secure information interchange, and a graph of the structure of the Internet from 2006. The synthetic graphs used were produced by an RMAT generator [6]. These include both Erdős-Rényi random graphs and small-world graphs that have many properties of real-world social networks, such as a power law degree distribution and low diameter [3,10,14]. We used parameters $\alpha = 0.25, \beta = 0.25, \gamma = 0.25, \delta = 0.25$ for the Erdős-Rényi random graph and $\alpha = 0.55, \beta = 0.1, \gamma = 0.1, \delta = 0.25$ for the small-world graph.

Table 1. Graph instances used in testing

Name	Vertices	Edges
citationCiteseer	268,495	1,156,647
coAuthorsCiteseer	227,320	814,134
coAuthorsDBLP	299,067	977,676
as-22july06	22,963	48,436
PGPgiantcompo	10,680	24,316
SmallWorld EF 8	32,768	237,523
SmallWorld EF 16	32,768	456,626
SmallWorld EF 32	32,768	861,878
Random EF 8	32,768	262,085
Random EF 16	32,768	524,031
Random EF 32	32,768	1,047,549

3.2 Graph Sampling Methods

The four approaches used to sample data are listed below for a graph with n vertices. For each one, we consider values of $p = 0.01, 0.05, 0.1, 0.15, \ldots, 0.8, 0.85$.

- **RandEdge:** Edges in the graph are chosen to be removed with equal probability p so that the error is distributed evenly across the network.
- **RandVertex:** Each vertex in the graph is chosen to be removed with equal probability p. When a vertex is removed, all of its incident edges are removed as well.
- **HighDegVertex:** The highest degree vertices and incident edges are removed. The top $p * n$ vertices are selected.
- **LowDegVertex:** The $p*n$ lowest degree vertices and their edges are removed.

3.3 Metrics Evaluated

We evaluate the graph connectivity, clustering coefficients, and degree centrality. The degree centrality of a vertex measures the number of edges incident on it and is the most basic centrality measure. We evaluate the sensitivity of degree centrality by measuring how much the rank of vertices' degree centrality changes when data is removed. For a given sampling level, the degree centrality rank is calculated for each vertex present in both the original and sampled graphs, resulting in two vectors. The Spearman correlation coefficient of these two rank vectors is then measured.

A connected component of a graph is a set of vertices linked by paths of edges. As vertices and edges are removed, the connected components of a graph may disconnect. While the number of components may increase, measuring the error in the number of connected components offers little information about how the structure of the graph has changed. A component splitting in half is a very different scenario from a few vertices disconnecting. We define the error in

connectivity as the proportion of pairs of vertices that were in the same component in the original graph and remain in the same component in the sampled graph. Only vertices with nonzero degree in the sampled graph are considered. Let $c(G, u, v)$ be an indicator function whose value is one when vertices u and v are in the same component in graph G and zero otherwise. The connectivity retained is then given by Eq. (2). Using the Shiloach-Vishkin algorithm [16], the estimated cost of computing connected components is $mlog(n)$ where m is the number of edges and n vertices.

$$ConnectivityRetained = \frac{\sum_{v,u \in G_{sampled}} c(G_{sampled}, u, v)}{\sum_{v,u \in G_{sampled}} c(G, u, v)} \tag{2}$$

The clustering coefficient measures the density of triangles in a graph and is one measure of the degree to which the graph is clustered. The local clustering coefficient of v is the ratio of closed triplets to open triplets of v and the global clustering coefficient is the ratio of total triangles to total triplets in the graph. High clustering coefficients suggest a small-world graph [17]. The global clustering coefficient can be used to characterize the entire graph, while local coefficients can reveal entities that engage in the most or least clustered activity. We measure the absolute and relative error in global clustering coefficient. To measure the sensitivity of local clustering coefficients, we calculate the Spearman correlation coefficient of the per-vertex rank in local clustering coefficient. Each vertex compares its list of adjacent vertices with the adjacency list of each of its neighbors, searching for intersections. Thus, the adjacency list of vertex v is accessed $d_v + 1$ times, once for itself, and once for each neighbor. Thus, the energy cost is given by $E = \epsilon_{mem} * \sum_v d_v^2$.

4 Results

The proportion of connectivity retained for each graph against energy required is plotted in Fig. 1. For each dataset, the energy value on the x-axis is the ratio of energy needed for $G_{sampled}$ to the energy needed for G. For all sampling types, the connected components of synthetic graphs are far more robust to missing data than those of real ones, which can be explained by the regular structure of RMAT graphs. Of the synthetic graphs, random graphs are the most robust with almost no error, while small-world graphs behave more similarly to real data. Removing low degree vertices causes the least amount of error across datasets. Removing high degree vertices, random vertices, and random edges provides diminishing returns as can be seen by the change in slope of the curves in Fig. 1.

The clustering coefficient of a graph also affects the sensitivity of its connected components. Among datasets studied, networks with a high global clustering coefficient require a higher proportion of energy to retain their connectivity structure. Figure 2 plots the clustering coefficient against the proportion of energy necessary to retain a connectivity of 0.85 and 0.95. For random edge, random vertex, and high degree vertex removal, the connected components of highly clustered graphs are least robust.

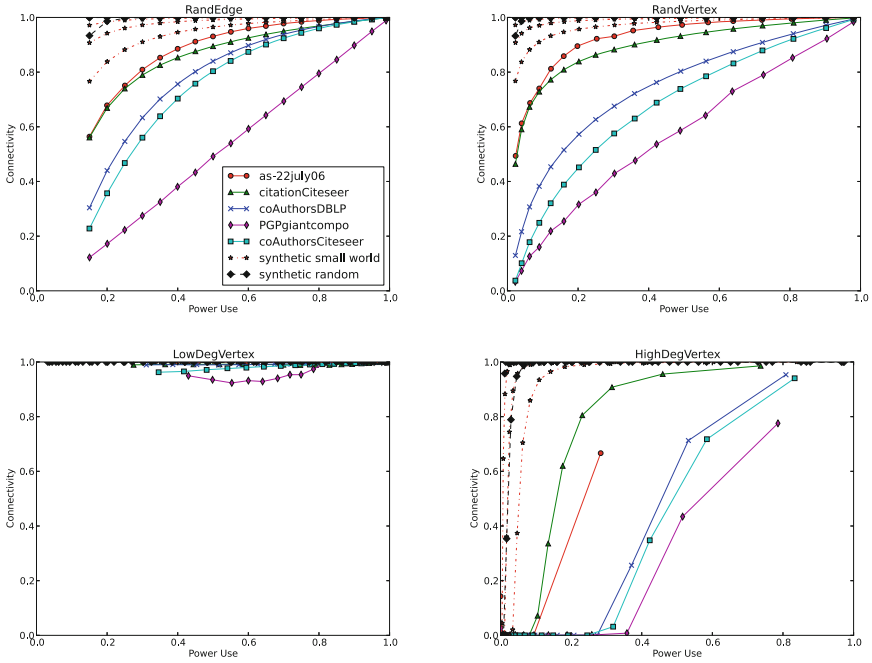

Fig. 1. The connectivity retained, or proportion of pairs of vertices that remain in the same connected component after sampling

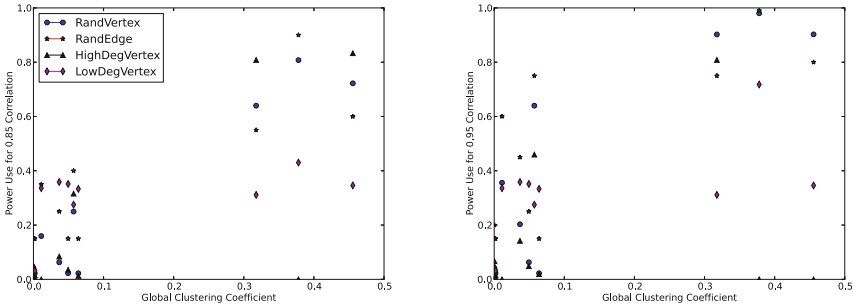

Fig. 2. Graph global clustering coefficient versus percentage of energy needed for 0.85 and 0.95 connectivity

Figure 3 plots the local clustering rank correlation coefficient against energy required. Real graphs are least sensitive to low and high degree vertex removal and most sensitive to random edge and vertex removal. In order to achieve a correlation coefficient of at least 0.85, the real graphs, in order listed in Table 1, need a 0.28, 0.47, 0.45, 1.0, and 0.13 proportion of energy with high degree vertex removal and 0.47, 0.48, 0.48, 0.2, and 0.86 with low degree vertex removal.

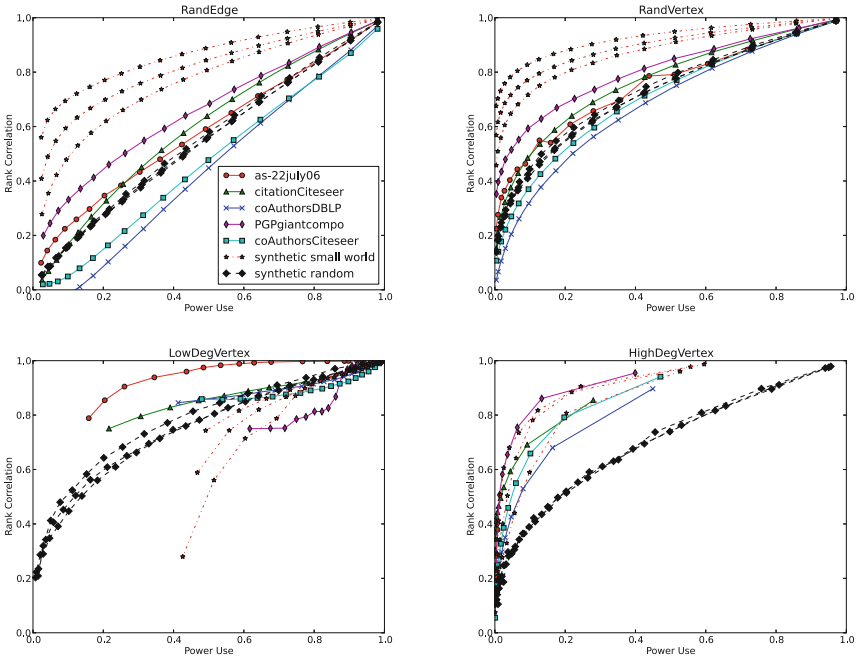

Fig. 3. Clustering coefficient rank correlation

With random edge and vertex removal, the real graphs need from 0.81 to 0.9 and from 0.62 to 0.73 energy, respectively. These relatively narrow bands show that random missing data may produce more consistent results, but at the cost of more energy usage. Random graphs are least sensitive to missing low degree vertices, requiring 0.61 to 0.7 energy for 0.85 correlation and most to missing random edges, requiring 0.81 to 0.9. Unlike random or real graphs, synthetic small-world graphs showed the most sensitivity when low degree vertices are removed. Despite these differences between the three network categories, Fig. 3 shows similar behavior for all types. As with connectivity error, data removal provides diminishing returns across the graphs studied. As the removal rate increases and the energy use decreases, the rate at which the clustering rank correlation coefficient falls increases. This suggests that significant energy savings could be achieved at relatively low error levels. It is interesting to note that for all graphs, this behavior is least prominent with random edge removal, where the curves are closer to linear.

Figure 4 plots the degree centrality rank correlation coefficient against energy used. A clear distinction can be seen between the sensitivity of real graphs, synthetic small-world graphs, and synthetic random graphs. Synthetic small-world graphs are most robust to all sampling methods, random graphs are the least robust, and the behavior of real graphs is in between the two. The robustness of small-world graphs compared to random ones can be explained by their skewed

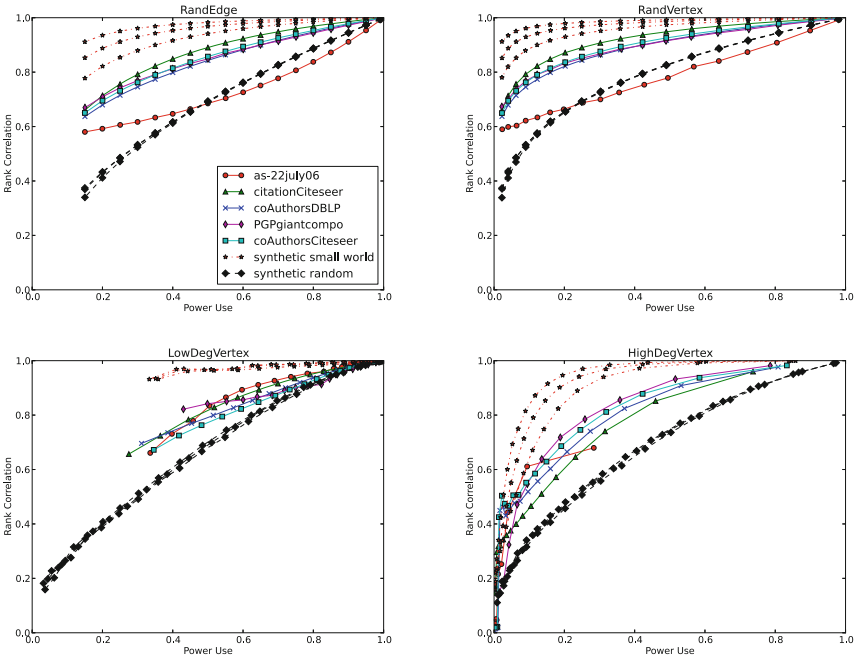

Fig. 4. Degree rank correlation

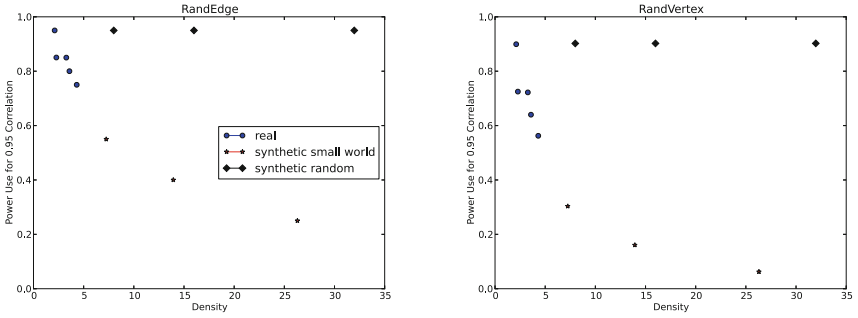

Fig. 5. Graph density versus energy needed to retain degree rank correlation of 0.95 using random edge removal (left) and random vertex removal (right)

degree distribution. Since there is little variation in vertex degree centrality in random graphs, this metric is very sensitive to missing data. Small-world graphs exhibit a large variation in degree centrality and so more data must be removed to change the metric. However, the results cannot be explained solely by a skewed degree distribution. The top 1 % of vertices contain a greater proportion of network edges in the real graphs than in the synthetic small-world graphs. Thus, the real graphs may have the most skewed degree distribution, but their degree

centrality is not most robust. Although the real datasets come from a variety of sources, their degree centrality values are affected similarly by missing data. As with connected components and local clustering coefficients, the gradient of the curves across datasets increases from right to left, suggesting that significant energy savings can be initially achieved with relatively low error, but that the error cost grows as more data is removed. Four of the five real graphs use less than a third of total energy to retain a degree centrality rank correlation coefficient of 0.85 with random vertex removal and less than 0.55 energy with random edge or high degree vertex removal.

The density of a graph, its ratio of edges to vertices, does affect its sensitivity to random edge and vertex sampling. In Fig. 5, density is plotted against the energy required to retain a degree centrality rank correlation coefficient of 0.95. Red stars denote synthetic small-world graphs and blue circles denote real graphs. For both of these categories, as the graph density increases, the proportion of energy needed decreases. This trend does not hold for random graphs, which are represented in the scatter plots with black diamonds.

5 Conclusion

We have investigated an approach for reducing the energy consumption of sparse graph algorithms with edge and vertex sampling. Such data removal will naturally result in errors which may or may not be tolerable, depending on the metric and application. We have examined the sensitivity of clustering coefficients, degree centrality, and connected components to various sampling strategies and analyzed the trade-off between energy reduction and error. Synthetic random graphs, synthetic small-world graphs, and real small-world graphs each tended to react distinctly. The structure of the graph is important in predicting the sensitivity to missing data and in choosing the best sampling technique and conclusions drawn from synthetic graphs may not be applicable to real data. Although the real networks came from a variety of sources, they tended to exhibit similar behavior that was distinct from that of either type of synthetic graph. Structural features such as the degree of clustering and density also have an effect on a network's robustness. It is interesting to note that in most cases, a similar pattern exists in the trade-off between energy savings and metric error. The gradient of the curve increases as energy use decreases, showing that the error cost of power saving is initially low, but grows at an increasing rate. This pattern suggests that significant energy savings might be achieved with relatively low error levels.

Acknowledgement. The work depicted in this paper was partially sponsored by Defense Advanced Research Projects Agency (DARPA) under agreement #HR0011-13-2-0001. The content, views and conclusions presented in this document do not necessarily reflect the position or the policy of DARPA or the U.S. Government, no official endorsement should be inferred. Distribution Statement A: "Approved for public release; distribution is unlimited."

References

1. Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex networks. Nature **406**(6794), 378–382 (2000)
2. Bader, D.A., Meyerhenke, H., Sanders, P., Wagner, D.: Graph partitioning and graph clustering. In: Proceedings of the 10th DIMACS Implementation Challenge Workshop. AMS (2013)
3. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
4. Benini, L., Bogliolo, A., De Micheli, G.: A survey of design techniques for system-level dynamic power management. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **8**(3), 299–316 (2000)
5. Borgatti, S.P., Carley, K.M., Krackhardt, D.: On the robustness of centrality measures under conditions of imperfect data. Soc. Netw. **28**(2), 124–136 (2006)
6. Chakrabarti, D., Zhan, Y., Faloutsos, C.: R-MAT: a recursive model for graph mining. In: SIAM International Conference on Data Mining (2004)
7. Choi, J., Bedard, D., Fowler, R., Vuduc, R.: A roofline model of energy. In: Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS) (2013)
8. David, H., Fallin, C., Gorbatov, E., Hanebutte, U.R., Mutlu, O.: Memory power management via dynamic voltage/frequency scaling. In: Proceedings of the 8th ACM International Conference on Autonomic Computing, pp. 31–40. ACM (2011)
9. Erdős, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hungar. Acad. Sci. **5**, 17–61 (1960)
10. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM '99, pp. 251–262. ACM (1999)
11. Korthikanti, V.A., Agha, G.: Towards optimizing energy costs of algorithms for shared memory architectures. In: Proceedings of the 22nd ACM Symposium on Parallelism in Algorithms and Architectures, SPAA '10, pp. 157–165. ACM (2010)
12. Kossinets, G.: Effects of missing data in social networks. Soc. Netw. **28**(3), 247–268 (2006)
13. Krishnamurthy, V., Faloutsos, M., Chrobak, M., Lao, L., Cui, J.-H., Percus, A.G.: Reducing large internet topologies for faster simulations. In: Boutaba, R., Almeroth, K.C., Puigjaner, R., Shen, S., Black, J.P. (eds.) NETWORKING 2005. LNCS, vol. 3462, pp. 328–341. Springer, Heidelberg (2005)
14. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 177–187. ACM (2005)
15. Lumsdaine, A., Gregor, D., Hendrickson, B., Berry, J.: Challenges in parallel graph processing. Parallel Process. Lett. **17**(1), 5–20 (2007)
16. Shiloach, Y., Vishkin, U.: An o(logn) parallel connectivity algorithm. J. Algorithms **3**, 57–67 (1982)
17. Watts, D., Strogatz, S.: Collective dynamics of small world networks. Nature **393**, 440–442 (1998)