Identification of Starting Points in Sampling of Complex Networks

Cesar Henrique Comin* and Luciano da Fontoura Costa†

Institute of Physics of São Carlos - University of São Paulo
Av. Trabalhador São Carlense 400, Caixa Postal 369,
CEP 13560-970 São Carlos, São Paulo, Brazil.

Abstract

Because of their large size, several real-world complex networks can only be represented and computationally handled in a sampled version. One of the most common sampling schemes is breadth-first where, after starting from a given node, its neighbors are taken, and then the neighbors of neighbors, and so on. Therefore, it becomes an important issue, given a specific network sampled by the breadth-first method, to try to identify the original node. In addition to providing a clue about how the sampling was performed, the identification of the original node also paves the way for the reconstruction of the sampling dynamics provided the original networks is available. In the current article, we propose and validate a new and effective methodology for the identification of the original nodes. The method is based on the calculation of the accessibility of the nodes in the sampled network. We show that the original node tends to have the highest values of accessibility. The potential of the methodology is illustrated with respect to three theoretical complex networks model as well as a real-world network, namely the United States patent network.

*Electronic address: chc@usp.br
†Electronic address: ldfcosta@gmail.com
I. INTRODUCTION

In complex networks research, sometimes we face the problem of having to analyze a very large system that can not be entirely known or stored and therefore requires sampling. Depending on the kind of information required, this sampling can proceed in many ways, although the most usual ones are depth first and breadth first (also known as dilation). The main difference between them is that in the former method the sampling proceeds through trails, i.e. departing from the initial (seed) node we visit one of its neighbors, then a neighbor of that neighbor, and so on. In the case of the breadth sampling scheme, we proceed from the seed node to all of its neighbors, then to their neighbors, and so on. In this paper we will only work with the breadth first sampling scheme because, besides corresponding to a frequent sampling approach, it is also directly related to several important processes taking place in complex systems, such as the dissemination of a biological or virtual virus, the spreading of a disease in an organism, the spread of the news about an event, as well as any other process underlain by diffusion dynamics.

Supposing that we have in hands a subnetwork that was breadth-first sampled from a bigger known or unknown network, a particularly important information that we can try to get about it is what was the respective starting node. Indeed, by knowing the seed node, it becomes possible to reproduce the process that created this subnetwork. Going back to the examples above, the seed node could give us the location of the first person or computer infected with a virus, the region of an organism where the disease started spreading, or the source of a notice or fad. Unfortunately, we usually do not have any information about the seed.

Little has been investigated in the literature about this subject, Costa et al. proposed a method of finding the origin of trails left by agents walking through a network, although the dilation process was done in a different manner. Kim and Jeong provide a wide view about sampling in scale-free networks, while Clauset and Moore analyse the bias created when sampling random graphs.

In this paper we propose and validate a method that allows to find the critically important seed node of a network that was extracted from a bigger one through a breadth sampling scheme. In order to do so, we will apply a powerful measurement called accessibility which has the characteristic of informing how effectively a node can reach other nodes. The
method will be applied to three network models, namely Erdös-Rényi [14], Barabási-Albert [15] and geographic [16], as well as a big real network that contains information about all the references between United States patents from 1975 to 1999 [3]. In the case of Erdös-Rényi model, we compare our model with two well known centrality measurements, namely closeness and betweenness.

Subsections II A and II B provide a brief review about the most important concepts and models of complex networks that will be used in our paper, while subsection II C 1 and II D present the accessibility measurement and shows how it can be applied to identify the seed node. Subsection III A and III B reports the results obtained for theoretical and real networks, respectively. Subsection III C and III D contains some insights about the method. Finally subsection III E compares the results obtained between accessibility and other two measurements.

II. MATERIALS AND METHODS

A. Representing complex networks

A Complex network is defined by a set of nodes, also known as vertices, with connections between them, called edges. A network can be of two types: directed or undirected. In this paper we will only work with undirected networks, but some results may apply to directed ones too. In order to represent a network of \( N \) nodes, we usually consider the adjacency matrix \( K \), that has \( N \) rows and columns. If node \( i \) exhibits a connection with node \( j \) then \( K(i,j) = K(j,i) = 1 \) (\( K(i,j) = K(j,i) = 0 \), otherwise). In many cases, like in this work, the adjacency matrix will have a lot of zeroes (sparse networks) or be prohibitively big (for too large \( N \)). In such cases, we can change the representation to an adjacency list, which is a vector of size \( N \), where each position \( i \) is the head of a dynamic list that stores the neighbors of node \( i \). In the case of undirected networks, the number of neighbors \( k \) of a node, i.e. \( k(i) = \sum_{j=1}^{N} K(i,j) \), is called the degree of that node. One of the most frequently adopted ways to characterize a network is in terms of its mean degree, which is defined as \( < k > = \frac{1}{N} \sum_{i=1}^{N} k(i) \). We observe that in this paper we will only consider the largest connected components, which is characterized by the property that it is possible to reach all the others nodes after departing from any node.
B. Description of the networks used

We will validate our proposed method through three theoretical network models and one real complex network. The network models are (a) Erdös-Rényi (ER) [14], (b) Barabási-Albert (BA) [15] and (c) geographic [16]. In the first model, we have \( N \) initial nodes and \( N(N-1)/2 \) possible connections. For each possible connection, we take it with probability \( p \) so that, in the end, we will have a network with \( < k > \approx p(N-1) \). In the second model we begin with a small initial number of nodes \( N_0 \) and sequentially add new nodes with \( m_0 \) edges, where each edge will connect to an existing node with probability proportional to the degree of that node. In the third model, we begin with a node in the \((0,0)\) position of a grid and randomly choose the position of new nodes that will join the network. These nodes will connect to the existing ones with a probability \( \beta e^{-\alpha d(i,j)} \), where \( d(i,j) \) is the euclidean distance between node \( i \) and \( j \) while \( \alpha \) (growth) and \( \beta \) (density) are adjustable parameters. The real network used in this work is the US patent database [3], which is an extensive compilation of connections (citations) between patents granted in the United States until 1999. The main characteristics of the largest undirected component of the patent database are shown in Table I.

C. Measuring centrality

1. Accessibility

The accessibility measurement [18] quantifies the potential of each node to access other nodes. Referring to Figure I if we were in the black node and wanted to visit every dark gray node using some random process, it is clear that in case Ia we would arrive at the top dark gray node half of the time, which is not an effective access means, while in case Ib all dark gray nodes are visited with equal probability, so the black node in Ib has higher accessibility than the one in Ia.

| Number of Nodes | Mean Degree | Degree Distribution                  | Clustering Coefficient |
|-----------------|-------------|--------------------------------------|------------------------|
| 3,764,117       | 8.8         | Linear, with angular coefficient −3.9| 0.076                  |

Table I: Measurements of the largest connected component of the US patent database
Figure 1: Accessibility example. The black node in 1b has higher accessibility than the one in 1a if we consider $h=2$ (see text for explanation).

In order to formally present this property, we need to define some concepts first. A walk of length $h$ over a network corresponds to the set of adjacent edges that an agent cross while passing by $h+1$ nodes. This walk can be random, that is, at each time step the agent goes, with equal probability, to one of the neighbors of the actual node, repeating the process again, until $h$ edges are crossed. In this paper we will utilize self-avoiding random walks, in which the agent never goes to an already visited node, this way we can reach more nodes in fewer steps.

Calling $P_h(i,j)$ the probability that an agent departing from node $i$ will end in node $j$ after $h$ steps of a self-avoiding random walk, we can define the diversity entropy of a node $i$ when going through $h$ steps as

$$E_h(i) = -rac{1}{\log(N-1)} \sum_{j=1,j\neq i}^N \begin{cases} P_h(i,j) \log(P_h(i,j)) & P_h(j,i) \neq 0 \\ 0 & P_h(j,i) = 0 \end{cases}$$

(1)

where $N$ is the number of nodes. The diversity entropy represents the accessibility of the node. It is important to say that an algorithm to find $E_h(i)$ can be deterministic, by taking every possible walk, or probabilistic, by letting $M$ agents walk over the network and counting how many arrived at each node. In this paper we only used deterministic algorithms.
2. Closeness and betweenness

Let $d_{ij}$ be the length of the shortest (geodesic) path between nodes $i$ and $j$, then the mean geodesic distance \cite{20} with respect to node $i$ is

$$l_i = \frac{1}{n-1} \sum_{j,j\neq i}^{} d_{ij}$$

(2)

where $n$ is the number of vertices on the network. Taking the inverse of $l_i$ we define the closeness centrality \cite{20} of the node $i$, that is

$$C_i = \frac{1}{l_i}.$$  

(3)

To define betweenness, let $n^i_{st}$ be the number of geodesic paths between nodes $s$ and $t$ that pass through $i$, and $g_{st}$ the total number of geodesic paths between $s$ and $t$, we define betweenness centrality \cite{20} as

$$\sum_{s,t,s\neq t}^{n^i_{st}} \frac{n^i_{st}}{g_{st}}.$$ 

(4)

D. Source identification

As suggested in \cite{18,19}, we can apply the accessibility measurement in order to try to identify the central nodes of the network. In the case of geographic networks the concept of centrality is intimately related to the geographic position of the nodes, while in other networks we can interpret it as providing information about the relative importance that a node has in the system, as far as the ability of reaching other nodes in a balanced way is concerned.

As noted above, when working with networks that are too large, it is usual to extract subnetworks and analyze only the properties of a subset of nodes \cite{12}. In this paper, we will use dilation in order to derive such networks. This procedure consists in beginning with an initial node (seed) $i$, and then visiting every neighbor of it. For each neighbor $j$ of $i$, visit every neighbor of $j$ (except $i$), and so on until a predetermined number of nodes have been visited, which will form the subnetwork (see Figure\cite{2}). After the extraction has been completed, it is expected that the node used for seed will be the most central to the subgraph.
Figure 2: Illustration of the dilation procedure, showing the network formed by one dilation (horizontal stripes) and two dilations (vertical stripes) of the seed node, represented by the black color.

generated, so we can think of the following idea for identification of the starting node: given a graph that was extracted by dilation we can try to find the central (seed) node that generated the subnetwork as that with the highest value of some centrality measurement. To verify this idea we will extract, by dilation, subnetworks of the four original networks described above and analyze the accessibility of the extracted nodes in order to try to recover the respective seed. As noted before, in subsection III E we will repeat the same process but use closeness and betweenness to find the seed.

It is important to emphasize that the accessibility has a dependence on the number of steps adopted, but it is expected that the correct number of steps used to identify the source will be the one that best covers all the nodes of the extracted network, i.e., if a subnetwork with 40 nodes was extracted from an original network with \( \langle k \rangle = 6 \), then we can say that almost two complete dilations were done to create it. Mathematically this can be expressed as

\[
    h \approx \text{int} \left( \frac{\log(N)}{\log(\langle k \rangle)} \right)
\]

where \( h \) is approximately the right step to choose, \( N \) and \( \langle k \rangle \) are, respectively, the size of the subnetwork and the mean degree of the original network and \( \text{int}(x) \) means the integer part of \( x \). The downside of this approach is that to estimate \( h \) we need some information from the original network, which is sometimes not available, so we will also try to use the
mean degree of the extracted network in equation 5. As a side note, $h$ could clearly be defined for each node as the largest geodesic path departing from it, but this approach would be much more susceptible to error.

Figure 3: Diversity entropy as a function of the distance from the seed, and the number of steps used for accessibility calculation, obtained for three networks models: (a) ER with $<k>=6$, (b) geographic with $<k>=6.7$ and (c) BA with $<k>=6$. All original networks had 10000 nodes. The error bars were rescaled by a factor of 10 for clearer visualization.
Figure 4: Seed finding rate as a function of the extracted network size and number of steps used for accessibility calculation for the same networks of Figure 3.

III. RESULTS AND DISCUSSION

A. Source identification of sampled theoretical networks

In order to investigate the viability of the idea proposed in Section II D, we considered the three theoretical models described before with $p = 0.0006$ for ER, $N_0 = 3$ and $m_0 = 3$ for BA and $\alpha = 1$ and $\beta = 0.001$ for the geographic model, all the original networks had $N = 10000$. We obtained the mean diversity entropy of the nodes of extracted subnetworks with 200 nodes as a function of the distance from the seed and the number of steps used for the accessibility calculation. So as to enhance statistical significance, the mean was estimated over 100 different subnetworks (100 distinct seeds), where each seed was chosen randomly. The results are shown in Figure 3.
We can see that, for some of the adopted steps, the seed node have higher mean accessibility than the others, so the idea was correct and we can now proceed to find the origin of extracted subnetworks. To do so we first considered the same networks as above and extracted subnetworks with sizes ranging from 10 to 5000 nodes using 400 different seeds for each size. After finding the nodes with the highest accessibility in each extracted network, we calculated the ratio between the number of seeds correctly found and the number of total seeds used (calling it success rate) as a function of the extracted size and the number of steps used for the accessibility calculation. The results are shown in Figure 4. It is important to note that in order to prevent border effects in the geographic model, we only picked seed nodes that were more than 0.1 distant from border, where 1 was the total spatial height and width of the network (in arbitrary units).

The method yielded very similar results between ER and geographic models, which can be explained by the fact that the latter model is very close to the ER network with the parameters chosen. The BA model did not perform so well because its original network tends to have very different accessibility for each node, e.g., in case there is a node near the seed with much higher degree (a cluster), the method will probably not work.

By analyzing Figure 4 it is clear that the method strongly depends on the number of steps used during accessibility calculation, each one having a maximum success rate for certain extracted sizes. Also, it appears to give bad results for certain extracted sizes. However, this is not entirely true, as we can see from Figures 5 and 6 where we plotted the rate of nodes with higher accessibility found as a function of the distance from the seed, taking into account extracted sizes that had the lowest success rates in Figure 4. It is clear that for the ER and geographic models, if the seed node was not found at least its neighbor was, which gives a good approximation about where the source is located. As for the BA model, it is evident that the method gives worse results.

B. Source identification in sampled real networks

The same calculations in III A were performed for the largest component of the US patent database with properties shown in table I with the only difference that the maximum extracted size was 1000 instead of 10000. The success rate of finding the seed and the histogram of highest accessibility versus distance from the seed (for 240 extracted nodes)
Figure 5: Finding rate of the node with the highest accessibility as a function of distance from the seed node. The sizes of the extracted networks, indicated below the figures, are the ones that obtained the lowest success rates of figure 4a and 4b.

We see that the success rate resembles the result obtained for the BA model, which was expected, and for some cases we could find the seed with certainty of almost 30 percent.

C. Choosing the right step

We saw that the success rate of the method strongly depends on the number of steps adopted. As we discussed in section III.A, equation 5 can give an estimate of the right value to choose, but it is somewhat unclear about using the mean degree of the original or the extracted network for the estimation. To compare the two methods, we extracted networks...
Figure 6: Finding rate of the node on the BA model with the highest accessibility as a function of distance from the seed node. The size of the extracted network is $N = 200$, the one with the lowest success rate between step two and three in Figure 4c.

Figure 7: (a) Success rate of the method and (b) finding rate of the node with the highest accessibility for the US patent database.

with varying number of dilations and tried to find this number using equation 5, the results are shown in Figure 8. It is clear that using the mean degree of the original network we found exactly every number of dilations used. It is also straightforward to note that extracting a network tends to lower its mean degree, which explains the higher estimations found for the other case. The difference in estimation is too large for the proper working of our method, and so we will only use the mean degree of the original network to obtain the following results.

Using the mechanism of finding the best step, the same calculations to find the success
Figure 8: Estimated number of dilations by equation 5 as a function of the real number of dilations done, using the mean degree of the original (red) and extracted (green) networks.

The used network was an ER with $N = 10000$ and $< k > = 6$.

rates shown in Figure 4 where performed, the results obtained are shown in Figure 9. We see that the concordance with the previous results are very good, especially for the ER and geographic models.

D. Dependence with network parameters

The results obtained so far used the same networks for each of the models. Now we aim at understanding how the success of the method varies with the network parameters. In order to do so, we considered the ER model and fixed the extracted size to that found to allow good success rates for 2 (40 nodes) and 3 (250 nodes) steps, and then calculated the same success rate as a function of the original network size and mean degree. The results are shown in Figure 10.

We clearly see that the success rate has no dependence on the size of the original network if $N$ is big compared to the size of the extracted one (for small $N$ the extracted network is approximately the original network) and that the method tends to give better results for larger $< k >$, which is somewhat expected, since networks with small $< k >$ tends to have low accessibility, and so extracting a subnetwork does not change it very much.
Figure 9: Seed finding rate as a function of the extracted network size for the same networks of Figure 4 but now trying to find the correct step for the calculation.

E. Comparison with other methods

To compare the accessibility with the other two widely used measurements described before, closeness and betweenness, we did the same process of section III A, but now for every extraction we try to guess the seed using, separately, the three measurements, obtaining the success rate of each method. The network used is the same as before, as indicated in the caption of figure 11.

We can see that for small extracted networks the closeness and betweenness measurements tend to give better results between steps of accessibility, but when the extracted network tends to the size of the original, the success rate of accessibility has a performance as far as three times better than the other two, which illustrates the power of measuring the balance of access in networks.
Figure 10: Success rate as a function of original network size and mean degree, for extracted sizes assuming (a) 40 and (b) 250 nodes. The small dots represent the simulation results. For better visualization the axes are rotated between (a) and (b).

Figure 11: Comparison of the success rate between accessibility (ACC), closeness (CLO) and betweenness (BET) as a function of extracted network size for an ER network with $N = 10000$ and $< k >= 6$

IV. CONCLUDING REMARKS

It is clear that the seed node has a great importance for the characterization of a network, we saw that by knowing it we can reproduce almost exactly the process of creation of the system, if it was done by dilation. Our purpose was to devise a method that could find this seed node with the highest success rate possible. We presented the accessibility measurement, which provides information about the ease of access to nodes, and applied it in diverse networks extracted from three known models (ER, geographic and BA) and the
US patent database. We found that the seed node has, in general, higher accessibility than the other nodes, so that finding the node with the highest potential of access allows the identification of the source of the network. The results obtained had success rates higher than 0.9 for ER and geographic models, which was a remarkable result, especially for the networks sampled with 3000 nodes. The problem of automatically choosing the right number of steps to evaluate the accessibility was easily solved by using the idea related to Equation 5, which again brought good results compared to the manual method. We then proceeded to evaluate the relation of the success rate with the ER network parameters, and found that the method works for networks of any size, as far as the extracted ones have smaller size than the original. Finally, we compared the success rate of the proposed method with two other measurements, finding that for extracted sizes approaching the original network size, the accessibility gives better results.

As said before, it is possible to obtain some different results using other dynamics to evaluate the accessibility. Also, we could devise a method of comparing the accessibility of the original network with that of the extracted one, which is an interesting idea, but would require us to know the original network, which is not always possible. Finally, the method could be applied to a network containing information about a real spreading process.

Acknowledgments

Luciano da Fontoura Costa thanks to CNPq (301303/06-1) and FAPESP (05/00587-5) for financial support.

[1] M. Barthelemy, Eur. Phys. Jour. B, vol 38, 163 (2004).
[2] Li W, Kurata H, "Visualizing Global Properties of Large Complex Networks", PLoS ONE 3(7): e2541. doi:10.1371/journal.pone.0002541 (2008).
[3] Hall, B. H., A. B. Jaffe, and M. Tratjenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.
[4] M. Latapy and C. Magnien, "Complex Network Measurements: Estimating the Relevance of Observed Properties", The 27th Conference on Computer Communications, pp. 1660-1668 (2008).
[5] YANG Bo, GAO Hai-xia and CHEN Zhong, "Efficient Sampling Strategies for Large-scale Complex Networks", 2008 International Conference on Management Science & Engineering, 334-339.

[6] R. Pastor-Satorras and A. Vespignani, "Epidemic Spreading in Scale-free Networks", Phys. Rev. Let., 86, (2001).

[7] A. Ganesh, L. Massoulie and D. Towsley. "The Effect of Network Topology on the Spread of Epidemics". In IEEE INFOCOM, 2005.

[8] D. S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A. L. Barabasi, Proceedings of the National Academy of Sciences of the United States of America, vol. 105, no. 29, pp. 9880–9885, (2008).

[9] Grãnlund, Andreas and Holme, Petter, Advances in Complex Systems (ACS), 08, issue 02, p. 261-273 (2005).

[10] L. da F. Costa, F. A. Rodrigues and G. Travieso, Phys. Rev. E 76, 046106 (2007).

[11] P.-J. Kim and H. Jeong, The European Physical Journal B, ISSN 1434-6036 (Online), p. 109-114, DOI 10.1140/epjb/e2007-00033-7.

[12] S.H. Lee, P.-J. Kim, H. Jeong, Phys. Rev. E 73, 016102 (2006).

[13] A. Clauset, C. Moore, Phys. Rev. Lett. 94, 018701 (2005).

[14] P. Erdos and A. Rényi, Publ. Math. 6, 290 (1959).

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

[16] M. Kaiser and C.C. Hilgetag, Phys. Rev. E 69, 036103 (2004).

[17] Francisco A. Rodrigues, Luciano da F. Costa, Phys. Rev. E 81, 036113 (2010).

[18] B.A.N. Travençolo and L. da F. Costa, Phys. Rev. A 373, 89-95 (2008).

[19] Bruno A N Travençolo, Matheus Palhares Viana and Luciano da Fontoura Costa, New J. Phys. 11, 063019 (2009).

[20] M. Newman, Networks: An Introduction, Oxford University Press, USA (May 20, 2010), p. 181-192.