Algorithms for graph and network analysis: Traversing/Searching/Sampling graphs

Paolo Lo Giudice[1] and Domenico Ursino[2*]

[1] DIIES, University "Mediterranea" of Reggio Calabria, Via Graziella, Località Feo di Vito, I-89122 Reggio Calabria, Italy

[2] DICEAM, University "Mediterranea" of Reggio Calabria, Via Graziella, Località Feo di Vito, I-89122 Reggio Calabria, Italy

Abstract

Traversing, searching and sampling approaches have been deeply investigated in graph theory and applied in a large variety of research fields. However, in bioinformatics and biomedicine, sampling complex networks is a new and little investigated task. For instance, it is used to classify knowledge for creating semantic maps, summarizations and multi-label classifications, or for search motifs. In this chapter, first we introduce a formalism to represent complex networks. Then, we provide both three taxonomies of network sampling approaches and a brief overview of each of them. After this, we present a comparison of these approaches. Finally, we draw some conclusions and take a look at the future.

Keywords: Node Sampling, Edge Sampling, Breadth First Search, Depth First Search, Random First Search, Snowball Sampling, Random Walk, Forest Fire Sampling, Respondent Driven Sampling

1 Introduction

Differently from other data analytics tasks, Network Analysis (NA) [15, 27, 28, 6] focuses on *relationships* existing between actors, instead of on actors.

In Network Analysis, one of the most important topics, representing the core for addressing several issues, is network search or network traversal. This problem refers to the task of visiting each node of the network. To carry out this activity, two main families of strategies are Breadth First Search (BFS, for short) and Depth First Search (DFS, for short). The former visits the siblings of the current node before visiting its children, whereas the latter visits the children of the current node before visiting its siblings.

When the networks being investigated are extremely large, and the computational effort necessary to perform the desired analyses is excessive, it could become impossible to operate on all its nodes and

*Corresponding Author – Phone Number: +39-347-7118918 – e-mail: ursino@unirc.it

edges. In this case, one of the most common approaches is sampling. Indeed, sampling approaches allow the extraction of knowledge about a network by investigating only a part of the network itself.

Clearly, the way the sample is chosen becomes crucial for extracting knowledge without errors or, at least, for minimizing the magnitude of errors. The problem of sampling from large graphs is discussed in [20]. Here, the authors investigate: *(i)* which sampling approaches must be used; *(ii)* how much the sampled graph can be reduced w.r.t. the original graph; *(iii)* how the measurements made on a sample can be scaled up to get estimates for the corresponding (generally much larger) graph. The problem of obtaining realistic samples with an as small as possible dimension is also described in [16]. Here, the authors show that some of the analyzed methods can maintain the key properties of the original graph, even if the sample dimension is about 30% smaller than the original graph.

Sampling has been extensively studied in the literature. For instance, in [10], the authors propose an approach that, given a communication network, determines the most important nodes and, then, links them each other. Instead, the authors of [26] propose a technique, based on both sampling and the randomized notion of *focus*, to allow the visualization of very large networks. An analysis of the statistical properties of a sampled network can be found in [18]. In [1], the authors use the social network Cyworld to analyze the main features of the snowball sampling approach.

Other approaches, such as [30, 7, 17, 11], focus mainly on sampling cost. Specifically, [30] analyzes how rapidly a crawler can reach nodes and links; [7] proposes a framework of parallel crawlers based on Breadth First Search (BFS); [17] investigates the impact of different sampling techniques on the computation of the average node degree of a network; [11] studies several crawling strategies and determines the sampling quality guaranteed by them and the computation effort they require. Finally, in [5, 4], the authors describe how the crawling problem and its solutions change when passing from a social networking to a social internetworking scenario (i.e., in a scenario where several social networks interact each other through *bridge* nodes).

In bioinformatics and biomedicine, sampling of complex networks is a new and little investigated task. One of the main issues faced in these two contexts is the rapid growth of the scientific knowledge presented in the literature. Sampling is mainly used to classify such knowledge. As a consequence, currently it is a supporting task for performing other activities, and is employed only rarely as the core approach to addressing issues in this context. For instance, in [8, 25, 13], the authors present some approaches that employ sampling on the existing literature to create: *(i)* semantic maps based on relationships [8]; *(ii)* summarizations [25]; *(iii)* multi-label classifications [13].

Sampling is also used to face a specific, yet extremely interesting, research problem, i.e., the search for motifs in a network. For instance, the authors of [14] propose a new algorithm allowing the estimation of the subgraph concentration at runtime; furthermore, in [3], the authors employ sampling to generate a probabilistic model of local protein structure; finally, in [2, 29], sampling is used to search motifs in biological networks.

In biomedical research, the most employed sampling approach is undoubtedly Random Walk (RW) and its variants. For instance, RW is adopted in [22, 24] to evaluate the relationships between proteins, genes and diseases. In [19], the authors employ RW to investigate and plot DNA sequences. Finally, in [23, 21, 9], RW is used to discover functional models and to infer pathway activity. In these latter cases, RW allows users to capture the information embedded in protein structure and to represent it in the resulting graph.

This chapter aims at providing an exhaustive overview of the existing algorithms for traversing, searching and sampling networks. It is organized as follows. In Section 2, we illustrate some preliminary concepts and introduce the formalism adopted throughout this chapter. In Section 3, first we propose three taxonomies for sampling approaches and, then, we provide a brief description of each approach. In Section 4, we present a comparison of sampling approaches based on property preservation and network property estimation. Finally, in Section 5, we draw our conclusions and have a look at future possible developments of this research issue.

2 Fundamentals

A network $\mathcal{N} = \langle V, E \rangle$ consists of a set V of nodes and a set E of edges. We use n and m to denote $|V|$ and $|E|$. Each edge, $e_{ij} = (v_i, v_j)$, connects the nodes v_i and v_j. Edges can be either directed (when they can be traversed only in one direction) or undirected (when they can be traversed in both directions). Furthermore, networks can be weighted or unweighted. If a network is weighted, the edge can be represented as (v_i, v_j, w_{ij}), where w_{ij} denotes the weight of the edges. On the basis of the reference context, this weight could represent strength, distance, similarity, etc.

Let v_i be a node of V. The *set of edges incident to* v_i is defined as $\iota(v_i) = \{(v_j, v_i, w_{ji}) | (v_j, v_i, w_{ji}) \in E\}$. The *neighborhood*, $\nu(v_i)$, is defined as $\nu(v_i) = \{v_j | (v_i, v_j, w_{ij}) \in E\}$.

A sampled network $\mathcal{N}_s = \langle V_s, E_s \rangle$ consists of a set $V_s \subseteq V$ of nodes and a set $E_s \subseteq E$ of edges such that $E_s \subseteq \{(v_i, v_j, w_{ij}) | v_i \in V_s, v_j \in V_s\}$. This last condition ensures that the sampled elements form a valid graph. We use the symbols n_s and m_s to denote $|V_s|$ and $|E_s|$, respectively. Clearly, $n_s \leq n$ and $m_s \leq m$. Each sampling activity has a cost and, often, a maximum budget B can be assigned to it. B has the same nature as the costs and can be used to cope with them.

3 Searching/traversing and sampling approaches

3.1 A look at searching/traversing approaches

The goal of a network search approach is to explore a network until a given desired target node has been reached. Instead, the goal of a network traversal approach is to find all the nodes of a network that can be reached from a given root node. The two main families of network searching/traversing strategies are Breadth First Search (BFS, for short) and Depth First Search (DFS, for short).

3.1.1 Breadth First Search/Traversal

If BFS is adopted as a network search approach, it begins at the root node and uses a queue as support data structure. First, it enqueues the root node.

At the generic iteration, it dequeues a node and examines it. If this node is the target one, it returns the corresponding value. Otherwise, it enqueues the direct child nodes that have not been visited.

If the queue is empty, it returns a negative value to indicate that the element was not found; otherwise, it performs the next iteration.

If BFS is adopted as a network traversal approach, it operates in an analogous way, but it terminates then the whole network has been visited.

The worst case time complexity of BFS is $O(|E| + |V|)$ as, in this case, every node and every edge will be explored.

3.1.2 Depth-First Search/Traversal

If DFS is adopted as a network search approach, it begins at the root node and uses a stack as support data structure. First, it pushes the root node.

At the generic iteration, it pops a node and examines it. If this node is the target one, it returns the corresponding value. Otherwise, it pushes a direct child node that has not been visited.

If the stack is empty, it returns a negative value to indicate that the element was not found; otherwise, it performs the next iteration.

If BFS is adopted as a network traversal approach, it operates in an analogous way, but it terminates when the whole network has been visited.

The worst case time complexity of DFS is $O(|E| + |V|)$ as, in this case, every node is visited only once and all the edges are crossed once.

3.2 Taxonomies of sampling approaches

There exist several taxonomies of sampling approaches. A first classification considers the sampling objective. In this case, we can distinguish approaches that: *(i)* get a representative subset of nodes; *(ii)* preserve certain properties of the original network; *(iii)* generate a random network. For this chapter, we will give more attention to the second type, i.e., property preservation.

A second taxonomy concerns the type of networks. In this case, we have: *(i)* Erdos-Renyi Network (ERN), also known as Random Graph, Exponential Random Graph, Poisson Random Graph, etc.; *(ii)* Power-Law Network (PLN), also called Scale-Free Network; *(iii)* Small-World Network (SMN); *(iv)* Fixed Degree Distribution Random Graph (FDDRG), also called "Configuration Model".

A third taxonomy is based on the adopted sampling techniques. In this case, we can consider:

- Node Sampling (NS).
- Edge Sampling (ES).
- Node Sampling with Neighborhood (NSN).
- Edge Sampling with Contraction (ESC).
- Node Sampling with Contraction (NSC).
- Traversal Based Sampling (TBS). This last is actually a family of techniques. In this case, the sampler starts with a set of initial nodes (and/or edges) and expands the sample on the basis of current observations. In this family, we can recognize:
  - Breadth First Search (BFS);

- Depth First Sampling (DFS);
- Random First Sampling (RFS);
- Snowball Sampling (SBS);
- Random Walk (RW);
- Metropolis-Hastings Random Walk (MHRW);
- Random Walk with Escaping (RWE);
- Multiple Independent Random Walkers (MIRW);
- Multi-Dimensional Random Walk (MDRW);
- Forest Fire Sampling (FFS);
- Respondent Driven Sampling (RDS) or Re-Weighted Random Walk (RWRW).

In the following, we use this last taxonomy and we give an overview to all the approaches mentioned above.

3.3 Description of sampling approaches

3.3.1 Node Sampling (NS)

This approach first selects V_s directly, i.e., uniformly or according to some distribution of V, determined on the basis of already known information about the nodes. Then, it selects the edges of E_s in such a way that $E_s = \{(v_i, v_j, w_{ij}) | (v_i, v_j, w_{ij}) \in E, v_i \in V_s, v_j \in V_s\}$.

3.3.2 Edge Sampling (ES)

This approach first selects $E_s \subseteq E$ uniformly at random, or according to some policy. Then, it selects V_s as $V_s = \{v_i, v_j | (v_i, v_j) \in E_s\}$. Alternatively, it can set $V_s = V$. In this last case, the edge sampling task reduces to a network sparsification task. As a matter of fact, network sparsification is a more general task than network sampling. Therefore, the latter can be considered as a specific case of the former.

3.3.3 Node Sampling with Neighborhood (NSN)

This approach first selects a set $\overline{V} \subseteq V$ directly, on the basis of available resources, without considering topological information. Then, it determines E_s as $E_s = \bigcup_{v_i \in \overline{V}} \iota(v_i)$ and $V_s = \{v_i, v_j | (v_i, v_j) \in E_s\}$. Finally, it returns $\mathcal{N}_s = \langle V_s, E_s \rangle$ as the sampled network.

3.3.4 Edge Sampling with Contraction (ESC)

This approach is based on the concept of contraction. We recall that the contraction of a pair of nodes v_i and v_j produces a network in which the two nodes v_i and v_j are replaced with a single node v_{ij} such that v_{ij} is adjacent to the union of the nodes to which v_i and v_j were originally adjacent. If v_i and v_j are connected by an edge, this edge is simply removed.

ESC is an iterative process. At each step, it samples one edge $(v_i, v_j, w_{ij}) \in E$ and performs the following tasks: *(i)* it substitutes nodes v_i and v_j with only one node v_{ij} representing both of them; *(ii)* it substitutes each edge involving v_i or v_j with an edge involving v_{ij}; *(iii)* it substitutes all the possible edges involving v_{ij} and the same node v_k with a unique edge involving the same nodes, whose weight is suitably determined from the weights of the merged edges, depending on the application context.

3.3.5 Node Sampling with Contraction (NSC)

This is an iterative process. At stage l, it samples one node v^l and contracts v^l and the nodes of $\nu(v^l)$ into one node. In carrying out this task, it suitably removes or modifies the corresponding edges. It is possible to show that NSC is a more constrained version of ESC.

3.3.6 Breadth First Sampling (BFS), Depth First Sampling (DFS), Random FirstSampling (RFS)

The Breadth First Sampling approach uses a support list L of nodes. As pointed out in Section 3.1, L is a queue for BFS and a stack for DFS and RFS. Initially, it selects a starting node v^0 and sets L to $\{v^0\}$, V_s to $\{v^0\}$ and E_s to $\emptyset$. Then, it repeats the following tasks until to the available budget B is exhausted: *(i)* it takes *the first element* v^l from L; *(ii)* for each $v_j \in \nu(v^l)$ such that $v_j \notin V_s$ and $v_j \notin L$, it adds v_j to L; v^l is called the "father" of v_j and is indicated as $f(v_j)$; *(iii)* it adds v^l to V_s; *(iv)* it adds the edge (v^l, v_j) to E_s; *(v)* it subtracts the cost of the current iteration from B.

DFS and RFS differ from BFS only in step *(i)* above. In fact, in DFS, *the last element* is selected from L, whereas, in RFS, *a random element* is chosen.

3.3.7 Snowball Sampling (SBS)

Snowball Sampling, or Network Sampling, or Chain Referral Sampling, is often used in sociology when it is necessary to perform an investigation on a hidden population (e.g., alcoholics).

It starts from an initial set V^0 of nodes, which can be obtained randomly or based on the side knowledge of the hidden population.

At stage l, it first sets the set $\overline{V^l}$ of visited nodes and the set E^l of visited edges to $\emptyset$. Then, for each node, $v^l \in V^{l-1}$, it selects k nodes belonging to the neighborhood, $\nu(v^l)$, of v^l uniformly at random, or according to some policy, adds them to $\overline{V^l}$, and adds the edges from v^l to each of these nodes to E^l. The methodology to perform the selection of the k nodes may depend on the application context. At the end of stage l, $V^l = \overline{V^l} - \bigcup_{j=0..l-1} V^j$.

The process is repeated for t stages until the budget B is exhausted.

The final sampled network $\mathcal{N} = \langle V_s, E_s \rangle$ is constructed by setting $V_s = \bigcup_{j=0..t} V^j$ and $E_s = \bigcup_{j=1..t} E^j$.

Note that SBS is very similar to BFS. Indeed, the difference is that BFS considers the whole neighborhood of the current node, whereas SBS considers only k nodes of this neighborhood.

3.3.8 Random Walk (RW)

Random Walk starts from an initial node v^0. Initially, it sets the set $\overline{E_s}$ of visited edges to $\emptyset$. At step l, it chooses one node, v_j, of the neighborhood, $\nu(v^{l-1})$, of v^{l-1}. This choice can be performed uniformly at random, or according to some policy. Then, it sets $v^l = v_j$ and adds to $\overline{E_s}$ the edge from v^{l-1} to v^l.

This process continues for t stages until to the budget B is exhausted. The final sampled network $\mathcal{N}_s = \langle V_s, E_s \rangle$ can be constructed in two different ways, namely:

- By setting $V_s = \{v^0, v^1, \ldots, v^t\}$ and $E_s = \overline{E_s}$.
- By setting $\overline{V_s} = \{v^0, v^1, \ldots, v^t\}$, $E_s = \bigcup_{v^l \in \overline{V_s}} \iota(v^l)$ and $V_s = \{v^l, v_j | (v^l, v_j) \in E_s\}$. In this case, RW reduces to Node Sampling with Neighborhood.

RW is also related to SBS. In fact, it can be considered as a specific case of SBS where $k = 1$. However, there is an important difference between them because RW is memoryless. In fact, in SBS, the participants from previous stages are excluded, whereas, in RW, the same node can be visited more than once.

It is possible to show that, when RW is applied on an undirected network, it returns a uniform distribution of edges. In this sense, it can be considered equivalent to ES.

Finally, it is worth pointing out that, if the choice of the next node to visit is performed uniformly at random, a node has a degree-proportional probability to be in V_s.

3.3.9 Metropolis-Hastings Random Walk (MHRW)

Metropolis-Hastings Random Walk is capable of returning a desired node distribution from an arbitrary undirected network. It uses two parameters, namely the probability P_{v^l,v_j} to pass from v^l to v_j and the desired distribution δ_v of a node v.

MHRW behaves analogously to RW. However, if v^l is the current node at stage l, the next node v_j to visit is determined according to the parameter P_{v^l,v_j}. The value of this parameter can be determined taking three possible cases into account. Specifically:

- If $v^l \neq v_j$ and $v_j \in \iota(v^l)$, then $P_{v^l,v_j} = M_{v^l,v_j} \cdot min\left\{1, \frac{\delta_{v_j}}{\delta_{v^l}}\right\}$.
- If $v^l \neq v_j$ and $v_j \notin \iota(v^l)$, then $P_{v^l,v_j} = 0$.
- If $v^l = v_j$ then $P_{v^l,v_j} = 1 - \sum_{v_k \in V_s, v_k \neq v^l} P_{v^l,v_k}$.

Here, $M_{v^l,v_j} = M_{v_j,v^l}$ is a normalization factor for the pair $\langle v^l, v_j \rangle$. It allows the condition $\sum_{v_k \in V_s, v_k \neq v^l} P_{v^l,v_j} \leq 1$ to be satisfied. Since adding more higher-weight self-loops makes the mixing time longer, M_{v^l,v_j} should be selected to be as large as possible. A possible choice for it is $M_{v^l,v_j} = min\left\{\frac{1}{|\iota(v^l)|}, \frac{1}{|\iota(v_j)|}\right\}$.

The application scenario of MHRW is more limited than the one of RW. In fact, to calculate P_{v^l,v_j}, the degree of the neighboring nodes should be known. This information is often unavailable even if, in some cases, it is fixed (e.g., in P2P) or it can be obtained through a suitable API (e.g., in Online Social Networks).

3.3.10 Random Walk with Escaping (RWE)

Random Walk with Escaping, or Random Jump, is analogous to RW. However, if v^l is the current node, to determine the next node to visit, besides walking to a node of $\iota(v^l)$, RWE can jump to an arbitrary random node $v_j \in V$. RWE is not very useful as a sampling technique. Indeed, it is classified as a TBS technique. However, TBS generally operates when the whole network cannot be reached, or at least direct Node Sampling or Edge Sampling is hard. By contrast, RWE needs an efficient Node Sampling as a support. As a consequence, it cannot be adopted in several scenarios. Furthermore, it is possible to show that, even when RWE can be adopted, it is hard to construct unbiased estimators for the properties of $\mathcal{N}$ starting from the ones of $\mathcal{N}_s$.

3.3.11 Multiple Independent Random Walkers (MIRW)

One problem of RW is that it tends to be trapped in locally dense regions. Therefore, it could have high bias, depending on the choice of initial nodes. Multiple Independent Random Walkers was proposed to address this problem. First, it applies NS to choose l initial nodes. Then, it splits the budget B among l Random Walks and lets them execute independently from each other. Finally, it merges the results produced by the l Random Walkers. As a matter of fact, it has been shown that the estimation errors of MIRW are higher than those of MDRW (see Section 3.3.12). As a consequence, we have mentioned MIRW only for completeness.

3.3.12 Multi-Dimensional Random Walk (MDRW)

Multi-Dimensional Random Walk, or Frontier Sampling, starts by determining the number of dimensions, k. Then, it initializes a list, L, of nodes by assigning k nodes, determined randomly via NS, to it. After this, it performs several iterations until the Budget, B, is exhausted.

During one of these iterations, it first chooses one node, v^l, from L with a probability $p(v^l)$ proportional to $|\iota(v^l)|$. Then, it selects a node $v_j \in \iota(v^l)$. Finally, it adds the edge (v^l, v_j, w_{lj}) to E_s and substitutes v^l with v_j in L.

It has been shown that: *(i)* MDRW provides very good estimations of some graph properties; *(ii)* when $l \to \infty$, MDRW obtains a uniform distribution of both nodes and edges.

3.3.13 Forest Fire Sampling (FFS)

Forest Fire Sampling can be considered as a probabilistic version of Snowball Sampling (see Section 3.3.7). Specifically, in SSB, k neighbors are selected at each round, whereas, in FFS, a geometrically distributed number of neighbors is selected at each round. If the parameter p of the geometric distribution is set to $\frac{1}{k}$, then the corresponding expectation is equal to k and FFS behaves very similarly to SBS.

An important common point between FFS and SBS, which differentiates both of them from RW and its variants, is that, in FFS and SBS, once a node is visited, it will not be visited again. By contrast, in RW and its variants, repeated nodes are included in the sample for estimation purposes.

3.3.14 Respondent Driven Sampling (RDS)

The original idea of Respondent Driven Sampling is to run SBS and to correct the bias according to the sampling probability of each node of V_s. Currently, RW is often substituted for SBS because the bias of RW can be more easily corrected. In this case, RDS is also called Re-Weighted Random Walk (RWRW). We point out that, actually, RDS itself is not a standalone network sampling technique. Indeed, it uses SBS or RW for sampling and, then, corrects the corresponding bias.

The principle underlying this approach is the following: it does not matter what sampling technique is adopted (NS, ES or TBS); as long as the sample probability is known, a suitable bias correction technique can be invoked.

If we consider sampling and estimating tasks as a whole activity, RWRW and MHRW seem to have the same objective and similar results. RWRW is a practical approach to estimate several properties without knowing the full graph.

4 Analysis and Assessment

In this section, we propose a comparison of network sampling approaches as far as network property preservation and network property estimation are concerned. Although these two goals are different, their results are strictly related and can be transformed into each other.

In the literature, it has been shown that the Node Sampling or the Edge Sampling approaches and their variants (i.e., NS, ES, NSN, ESC and NSC) are completely dominated by Traversal Based Sampling approaches across all network features [12]. Among the TBS approaches, there is no clear single winner. Each approach is the best one for at least some network feature of a particular network configuration.

More specifically, it has been shown that, in presence of a Poisson degree distribution, approaches such as SBS and FFS, configured with the mean of their geometric distribution set to 80% of the number of the remaining unselected neighbors (we call this configuration FFS80%), can reconstruct a good representation of some local parts of the network nodes relatively well. Furthermore, in the presence of a power law degree distribution, approaches as RW and FFS, configured with the mean of the geometric distribution set to 20% of the number of the remaining unselected neighbors (we call this configuration FFS20%), which explores nodes farther away from the focal ones, performs better.

A closer examination of the approaches provides an, at least partial, explanation of these results. Indeed, SSB tends to return sampled networks whose degree distributions contain inflated proportions of nodes with the highest and the lowest degrees. Clearly, this causes these approaches to perform poorly when applied to networks with a power-law degree distribution, which are characterized by a small proportion of high-degree nodes. On the contrary, RW tends to return sampled networks, whose nodes never have the highest degrees. Now, since the proportion of nodes with the highest degree is lower in the power-law degree distribution than in the Poisson distribution, RW performs better when applied to networks with the former distribution than to networks with the latter one. Furthermore, networks with Poisson distributions tend to be homogeneous throughout their regions; as a consequence, a locally-oriented approach, like SSB, can provide good results. On the contrary, networks with power-law degree distributions require a more global exploration; as a consequence, for

this kind of network, FFS and RW appear more adequate.

Summarizing and, at the same time, deepening this topic, we can say that SBS is well suited for sampling social networks with Poisson degree distribution, RW is adequate for sparse social networks with power-law degree distribution and FF is well suited for dense social networks with power-law degree distribution. To implement this recommendation, the degree distribution of network nodes must be known. However, this information may be unavailable in many cases. In the literature, it has been shown that FFS presents the best overall performance in determining degree distributions across different kinds of network and sample size. Therefore, it could be useful to exploit an adaptive sampling procedure using different sampling approaches at different stages. For instance, this procedure could start with FFS when no knowledge about the distribution of network nodes is available. Then, after a certain number of nodes have been included in the sample, it would be possible to determine the degree distribution of the current sample and, based on it, to continue with FFS or to switch to SBS or RW.

5 Closing Remarks

In this chapter, we have provided a general presentation of algorithms for traversing, searching and sampling graphs. We have seen that these algorithms have been intensively investigated in many research fields. On the other hand, they have been little employed in bioinformatics and biomedicine, where the most important adoption cases regard knowledge classification and motif search. In this chapter, we have introduced a formalism to represent a complex network, we have provided three taxonomies of sampling approaches, we have presented a brief description of each of them and, finally, we have compared them. We think that network traversing/searching/sampling approaches could have many more use cases in the future. As a matter of fact, the amount of available data is increasing enormously. This fact could give rise to increasingly sophisticated networks. In several cases, it could be impossible to perform the analysis of the whole network; when this happens, the possibility of generating reliable samples could be extremely beneficial.

Acknowledgement

This work was partially supported by Aubay Italia S.p.A.

References

[1] Y.Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. of the International Conference on World Wide Web (WWW'07)*, pages 835–844, Banff, Alberta, Canada, 2007. ACM.

[2] N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. Sahinalp. Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24(13):i241–i249, 2008. Oxford Univ Press.

[3] W. Boomsma, K. Mardia, C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932–8937, 2008. National Acad Sciences.

[4] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Experiences using BDS, a crawler for Social Internetworking Scenarios. *Social Networks: Analysis and Case Studies*, pages 149–177, 2014. Lecture Notes in Social Networks. Springer.

[5] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Moving from social networks to social internetworking scenarios: The crawling perspective. *Information Sciences*, 256:126–137, 2014. Elsevier.

[6] P. Carrington, J. Scott, and S. Wasserman. *Models and Methods in Social Network Analysis*. 2005. Cambridge University Press.

[7] D.H Chau, S. Pandit, S. Wang, and C. Faloutsos. Parallel crawling for online social networks. In *Proc. of the International Conference on World Wide Web (WWW'07)*, pages 1283–1284, Banff, Alberta, Canada, 2007. ACM.

[8] A. Coulet, N. Shah, Y. Garten, M. Musen, and R. Altman. Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics*, 43(6):1009–1019, 2010. Elsevier.

[9] V. Freschi. Protein function prediction from interaction networks using a random walk ranking algorithm. In *Proc. of the International Conference on Bioinformatics and Bioengineering (BIBE 2007)*, pages 42–48, Harvard, MA, USA, 2007. IEEE.

[10] A.C. Gilbert and K. Levchenko. Compressing network graphs. In *Proc. of the International Workshop on Link Analysis and Group Detection (LinkKDD'04)*, Seattle, WA, USA, 2004. ACM.

[11] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *Proc. of the International Conference on Computer Communications (INFOCOM'10)*, pages 1–9, San Diego, CA, USA, 2010. IEEE.

[12] Pili Hu and Wing Cheong Lau. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865*, 2013.

[13] B. Jin, B. Muller, C. Zhai, and X. Lu. Multi-label literature classification based on the Gene Ontology graph. *BMC Bioinformatics*, 9(1):525, 2008. BioMed Central.

[14] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004. Oxford Univ Press.

[15] D. Knoke and S. Yang. *Social Network Analysis*, volume 154. Sage, 2008.

[16] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J.H. Cui, and A. Percus. Reducing Large Internet Topologies for Faster Simulations. In *Proc. of the International Conference on Networking (Networking 2005)*, pages 165–172, Waterloo, Ontario, Canada, 2005. Springer.

[17] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of BFS (Breadth First Search). In *Proc. of the International Teletraffic Congress (ITC 22)*, pages 1–8, Amsterdam, The Netherlands, 2010. IEEE.

[18] S.H. Lee, P.J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102, 2006. APS.

[19] P. Leong and S. Morgenthaler. Random walk and gap plots of DNA sequences. *Computer Applications in the Biosciences: CABIOS*, 11(5):503–507, 1995. Oxford Univ Press.

[20] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, pages 631–636, Philadelphia, PA, USA, 2006. ACM.

[21] W. Liu, C. Li, Y. Xu, H. Yang, Q. Yao, J. Han, D. Shang, C. Zhang, F. Su, and X. Li. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics*, 29(17):2169–2177, 2013. Oxford Univ Press.

[22] Y. Liu, X. Zeng, Z. He, and Q. Zou. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016. IEEE.

[23] K. Macropol, T. Can, and A. Singh. RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC bioinformatics*, 10(1):283, 2009. BioMed Central.

[24] S. Navlakha and C. Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063, 2010. Oxford Univ Press.

[25] L. Plaza, A. Diaz, and P. Gervas. A semantic graph-based approach to biomedical summarisation. *Artificial Intelligence in Medicine*, 53(1):1–14, 2011. Elsevier.

[26] D. Rafiei and S. Curial. Effectively visualizing large networks through sampling. In *Proc. of the IEEE Visualization Conference 2005 (VIS'05)*, page 48, Minneapolis, MN, USA, 2005. IEEE.

[27] J. Scott. *Social Network Analysis*. Sage, 2012.

[28] S. Wasserman and J. Galaskiewicz. *Advances in social network analysis: Research in the social and behavioral sciences*, volume 171. Sage Publications, 1994.

[29] E. Wong, B. Baur, S. Quader, and C. Huang. Biological network motif detection: principles and practice. *Briefings in Bioinformatics*, page bbr033, 2011. Oxford Univ Press.

[30] S. Ye, J. Lang, and F. Wu. Crawling online social graphs. In *Proc. of the International Asia-Pacific Web Conference (APWeb'10)*, pages 236–242, Busan, Korea, 2010. IEEE.