Little Ball of Fur: A Python Library for Graph Sampling

Citation for published version:

Digital Object Identifier (DOI):
10.1145/3340531.3412758

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 29th ACM International Conference on Information & Knowledge Management

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Little Ball of Fur: A Python Library for Graph Sampling

Benedek Rozemberczki  
The University of Edinburgh  
Edinburgh, United Kingdom  
benedek.rozemberczki@ed.ac.uk

Oliver Kiss  
Central European University  
Budapest, Hungary  
kiss_oliver@phd.ceu.edu

Rik Sarkar  
The University of Edinburgh  
Edinburgh, United Kingdom  /rsarkar@inf.ed.ac.uk

ABSTRACT
Sampling graphs is an important task in data mining. In this paper, we describe Little Ball of Fur, a Python library that includes more than twenty graph sampling algorithms. Our goal is to make node, edge, and exploration-based network sampling techniques accessible to a large number of professionals, researchers, and students in a single streamlined framework. We created this framework with a focus on a coherent application public interface which has a convenient design, generic input data requirements, and reasonable baseline settings of algorithms. Here we overview these design foundations of the framework in detail with illustrative code snippets. We show the practical usability of the library by estimating various global statistics of social networks and web graphs. Experiments demonstrate that Little Ball of Fur can speed up node and whole graph embedding techniques considerably with mildly deteriorating the predictive value of distilled features.

KEYWORDS
network sampling, graph sampling, subsampling, network analytics, graph mining

ACM Reference Format:

1 INTRODUCTION
Modern graph datasets such as social networks and web graphs are large and can be mined to extract detailed insights. However, the large size of the datasets pose fundamental computational challenges on graphs [9, 16]. Exploratory data analysis and computation of basic descriptive statistics can be time consuming on real world graphs. More advanced graph mining techniques such as node and edge classification or clustering can be completely intractable on full size datasets such as web graphs.

One of the fundamental techniques to deal with large datasets is sampling. On simple datasets such as point clouds, sampling preserves most of the distributional features of the data and forms the basis of machine learning [40]. However, graphs represent complex interrelations, so that naive sampling can destroy the salient features that constitute the value of the graph data. Graph sampling algorithms therefore need to be sensitive to the various features that are relevant to the downstream tasks. Such features include statistics such as diameter, clustering coefficient [6], transitivity or degree distribution. In more complex situations, graphs are used for community detection, classification, edge prediction etc [13]. A sampling algorithm should be representative with respect to such downstream learning tasks.

Various graph sampling procedures have been proposed with different objectives [14]. The implementation of the graph sampling technique and choice of its parameters used for the subgraph extraction can affect its utility for the task in question. A toolbox of well understood graph sampling techniques can make it easier for researchers and practitioners to easily perform graph sampling, and have consistent reproducible sampling across projects. Our goal is to make a large number of graph sampling techniques available to a large audience.

Present work. We release Little Ball of Fur – an open-source Python library for graph sampling. This is the first publicly available and documented Python graph sampling library. The general design of our framework is centered around an end-user friendly application public interface which allows for fast paced development and interfacing with other graph mining frameworks.

We achieve this by applying a few core software design principles consistently. Sampling techniques are implemented as individual classes and have pre-parametrized constructors with sensible default settings, which include the number of sampled vertices or edges, a random seed value and hyperparameters of the sampling method itself. Algorithmic details of the sampling procedures are implemented as private methods in order to shield the end-user from the inner mechanics of the algorithm. These concealed workings of samplers rely on the standard Python library and Numpy [47]. Practically, sampling techniques only provide a single shared public method (sample) which returns the sampled graph. Sampling procedures use NetworkX [12] and NetworkKit [41] graphs as the input and the output adheres to the same widely used generic formats.

We demonstrate the practical applicability of our framework on various social networks and web graphs (e.g. Facebook, LastFM, Deezer, Wikipedia). We show that our package allows the precise estimation of macro level statistics such as average degree, transitivity, and degree correlation. We provide evidence that the use of sampling routines from Little Ball of Fur can reduce the runtime of node and whole graph embedding algorithms. Using these embeddings as input features for node and graph classification tasks we establish that the embeddings learned on the subsampled graphs extract high quality features.
Our contributions. Specifically, the main contributions of our work can be summarized as:

1. We present Little Ball of Fur a Python graph sampling library which includes various node, edge and exploration based subgraph sampling techniques.
2. We use code snippets to discuss the design principles which we applied when we developed our package. We examine the presence of standard hyperparameter settings, inner sampling mechanics which are implemented as private methods, the unified application public interface and the use of open-source scientific data structures.
3. We demonstrate on various real world social network and webgraph datasets how sampling a subgraph with Little Ball of Fur affects the estimation accuracy of graph-level macro statistics. We present real world graph mining case studies where using our sampling framework speeds up graph embedding.
4. We provide a detailed documentation of our sampling package with a tutorial and case studies with code snippets.

The rest of this paper has the following structure. In Section 2 we overview the relevant literature about graph sampling. This discussion covers node, edge, and exploration sampling techniques, and the possible applications of sampling from graphs. The design principles which we followed when Little Ball of Fur was developed are discussed in Section 3 with samples of illustrative Python code. The subsampling techniques provided by our framework are evaluated in Section 4. We present results about network statistic estimation performance, machine learning case studies about node and graph classification. The paper concludes with Section 5 where we discuss our main findings and point out directions for future work.

We open sourced the software package and it can be downloaded from https://github.com/benedekrozemberczki/littleballoffur; the Python package can be installed via the Python Package Index. A comprehensive documentation can be accessed at https://little-ball-of-fur.readthedocs.io/en/latest/ with a step-by-step tutorial.

2 RELATED WORK

In this section we briefly overview the types of graph subsampling techniques included in Little Ball of Fur and the node and graph level representation learning algorithms used for the experimental evaluation of the framework.

2.1 Graph sampling techniques

Graph subsampling procedures have three main groups – node, edge, and exploration-based techniques. We give a brief overview of these techniques in this section.

2.1.1 Node sampling. Techniques which sample nodes select a set of representative vertices and extract the induced subgraph among the chosen vertices. Nodes can be sampled uniformly without replacement (RN) [42], proportional to the degree centrality of nodes (RDN) [1] or according to the pre-calculated PageRank score of the vertices (PRN) [19]. All of these methods assume that the set of vertices and edges in the graph is known ex-ante.

2.1.2 Edge sampling. The simplest link sampling algorithm retains a randomly selected subset of edges by sampling those uniformly without replacement (RE) while another approach is to randomly select nodes and an edge that belongs to the chosen node (RNE) [17]. These techniques can be hybridized by alternating between node-edge sampling and random edge selection with a parametrized probability (HRNE) [17].

By randomly selecting a set of retained edges one implicitly samples nodes. Because of this the random edge selection can be followed up by an induction step (TIES) [2] in which the additional edges among chosen nodes are all added. This step can be a partial induction (PIES) [2] if the edges were sampled in a streaming fashion and only edges with already sampled nodes are selected in the induction step.

2.1.3 Exploration based sampling. Node and edge sampling techniques do not extract representative subsamples of a graph by exploring the neighbourhoods of seed nodes. In comparison exploration based sampling techniques probe the neighborhood of a seed node or a set of seed vertices.

A group of exploration based sampling techniques uses search strategies on the graph to extract a subsample. The simplest search based strategies include classical traversal methods such as breadth first search (BFS) and depth first search (DFS) [5]. Snow ball sampling (SB) [10] is a restricted version of BFS where a maximum fixed k number of neighbors is traversed. Forest fire (FF) sampling [20] is a parametrized stochastic version of SB sampling where the constraint on the maximum number of traversed neighbours only holds in expectation. A local greedy search based technique is community structure expansion sampling [26] (CSE) which starting with a seeding node adds new nodes to the sampled set which can reach the largest number of unknown nodes. Another simpler search based sampling technique is the random node-neighbor (RNN) [19] algorithm which randomly selects a set of seed nodes, takes the neighbors in a single hop and induces the edges of the resulting vertex set. Searching for shortest paths (SP) [33] between randomly sampled nodes can be used for selecting sequences of nodes and edges to induce a subsampled graph.

A large family of exploration based graph sampling strategies is based on random walks (RW) [8]. These techniques initiate a random walker on a seed node which traverses the graph and induces a subgraph which is used as the sample. Random walk based sampling has numerous shortcomings and a large number of sampling methods tries to correct for specific limitations.

One of the major limitations is that random walks are inherently biased towards visiting high degree nodes in the graph [14], Metropolis-Hastings random walk (MHRW) [15, 43] and its rejection constrained variant (RC-MHRW) [24] address this bias by making the walker prone to visit lower degree nodes.

Another major shortcoming of random walk based sampling is that the walker might get stuck in the closely knit community of the seed node. There are multiple ways to overcome this. The first one is the use of non-backtracking random walks (NBTRW) [18] which removes the tottering behaviour of random walks. The second one is circulating the neighbors of every node with a vertex specific queue (CNRW) [50]. A third strategy involves teleports - the random walker jumps (RWJ) [34] with a fixed probability to a random node from the current vertex. A fourth approach is to make the walker biased towards weak links by creating a common
neighbor aware random walk sampler (CNARW) [25] which is biased towards neighbors with low neighborhood overlap. A fifth strategy is using multiple random walkers simultaneously which form a so called frontier of random walkers (FRW) [34]. These techniques can be combined with each other in a modular way to overcome the limitations of random walk based sampling.

There are other possible modifications to traditional random walks which we implemented in Little Ball of Fur. One example is random walk with restart (RWR) [19], which is similar to RWJ sampling, but the teleport always ends with the seed node or loop erased random walks (LERW) [48] which can sample spanning trees from a source graph uniformly.

2.2 Node and whole graph embedding

Our experimental evaluation includes node and graph classification for which we use features extracted with neighbourhood preserving node embeddings and whole graph embedding techniques.

2.2.1 Neighbourhood preserving node embedding. Given a graph \( G = (V, E) \) neighbourhood preserving node embedding techniques [3, 11, 28, 31, 32, 38, 44] learn a function \( f : V \to \mathbb{R}^d \) which maps the nodes \( v \in V \) into a \( d \) dimensional Euclidean space. In this embedding space a pre-determined notion of node-node proximity is approximately preserved by learning the mapping. The vector representations created by the embedding procedure can be used as input features for node classifiers.

2.2.2 Whole graph embedding and statistical fingerprinting. Starting with a set of graphs \( \mathcal{G} = (G_1, \ldots, G_n) \) whole graph embedding and statistical fingerprinting procedures [4, 27, 39, 45, 46] learn a function \( h : \mathcal{G} \to \mathbb{R}^d \) which maps the graphs \( G \in \mathcal{G} \) to a \( d \) dimensional Euclidean space. In this space those graphs which have similar structural patterns are close to each other. The vector representations distilled by these whole graph embedding techniques are useful inputs for graph classification algorithms.

3 DESIGN PRINCIPLES

We overview the core design principles that we applied when we designed Little Ball of Fur. Each design principle is discussed with illustrative examples of Python code which we explain in detail.

3.1 Encapsulated sampler hyperparameters, random seeding, and parameter inspection

Graph sampling methods in Little Ball of Fur are implemented as individual classes which all inherit from the Sampler class. A Sampler object is created by using the constructor which has default out-of-the-box hyperparameter settings. These default settings are available in the documentation and can be customized by re-parametrizing the Sampler constructor. The hyperparameters of the sampling techniques are stored as public attributes of the Sampler instance which allows for inspection of the hyperparameter settings by the user. Each graph sampling procedure has a seed parameter – this value is used to set a random seed for the standard Python and NumPy random number generators. This way the subsample extracted from a specific graph with a fixed seed is always going to be the same.

The code snippet in Figure 1 illustrates the encapsulated hyperparameter and inspection features of the framework. We start the script by importing a simple random walk sampler from the package (line 1). We initialize the first random walk sampler instance without changing the default hyperparameter settings (line 3). As the seed and hyperparameters are exposed we can print the seed parameter which is a public attribute of the sampler (line 4) and we can see the default value of the seed. We create a new instance with a parametrized constructor which sets a new seed (line 6) which modifies the value of the publicly available random seed (line 7).

```python
from littleballoffur import RandomWalkSampler
sampler = RandomWalkSampler()
sampler.seed
sampler = RandomWalkSampler(seed=41)
sampler.seed
```

Figure 1: Re-parametrizing and initializing the constructor of a random walk sampler by changing the random seed.

3.2 Achieving API consistency and non-proliferation of classes

The graph sampling procedures included in Little Ball of Fur are implemented with a consistent application public interface. As we discussed the parametrized constructor is used to create the sampler instance and the samplers all have a single available public method. The subsample of the graph is extract by the use of the sample method which takes the source graph and calls the private methods of the sampling algorithm.

We limited the number of classes and methods in Little Ball of Fur with a straightforward design strategy. First, the graph sampling procedures do not rely on custom data structures to represent the input and output graphs. Second, inheritance from the Sampler ensures that private methods which check the input format requirements do not have to be re-implemented on a case-by-case basis for each sampling procedure.

```python
from networkx as nx
import littleballoffur
import RandomWalkSampler

graph = nx.watts_strogatz_graph(1000, 10, 0)
sampler = RandomWalkSampler()
sampled_graph = sampler.sample(graph)
print(nx.transitivity(sampled_graph))
```

Figure 2: Using a random walk sampler on a Watts-Strogatz graph without changing the default sampler settings.

In Figure 2, first we import NetworkX and the random walk sampler from Little Ball of Fur (lines 1-2). Using these libraries we create a Watts-Strogatz graph (line 4) and a random walk sampler
with the default hyperparameter settings of the sampling procedure (line 6). We sample a subgraph with the public sample method of the random walk sampler (line 7) and print the transitivity calculated for the sampled subgraph (line 8).

```python
import networkx as nx
from littleballoffur import ForestFireSampler

graph = nx.watts_strogatz_graph(1000, 10, 0)
sampler = ForestFireSampler()
sampled_graph = sampler.sample(graph)
print(nx.transitivity(sampled_graph))
```

Figure 3: Using a forest fire sampler on a Watts-Strogatz graph without changing the default sampler settings.

The piece of code presented in Figure 2 can be altered seamlessly to perform Forest Fire sampling by modifying the sampler import (line 2) and changing the constructor (line 7) – these modifications result in the example in Figure 3.

These illustrative sampling pipelines presented in Figures 2 and 3 demonstrate the advantage of maintaining API consistency for the samplers. Changing the graph sampling technique that we used only required minimal modifications to the code. First, we replaced the import of the sampling technique from the Little Ball of Fur framework. Second, we used the constructor of the newly chosen sampling technique to create a sampler instance. Finally, we were able to use the shared sample method and the same pipeline to calculate the transitivity of the sampled graph.

### 3.3 Backend deployment, standardized dataset ingestion and limitations

Little Ball of Fur was implemented with a backend wrapper. Sampling procedures can be executed by the NetworKit [41] or NetworkX [12] backend libraries depending on the format of the input graph. Basic graph operations such as random neighbor or shortest path retrieval of the backend libraries have standardized naming conventions, data generation and ingestion methods. The generic backed wrapper based design allows for the future inclusion of other general graph manipulation backend libraries such as GraphTool [30] or SNAP [22].

The shared public sample method of the node, edge and exploration based sampling algorithms takes a NetworkX/NetworKit graph as input and the returned subsample is also a NetworkX/NetworKit graph. The subsampling does not change the indexing of the vertices.

The rich ecosystem of graph subsampling methods and the consistent API required that Little Ball of Fur was designed with a limited scope and we made restrictive assumptions about the input data used for sampling. Specifically, we assume that vertices are indexed numerically, the first index is 0 and indexing is consecutive. We assume that the graph that is passed to the sample method is undirected and unweighted (edges have unit weights). In addition, we assume that the graph forms a single strongly connected component and orphaned nodes are not present. Heterogeneous, multiplex, multipartite and attributed graphs are not handled by the 1.0 release of the sampling framework.

The sampler classes all inherit private methods that check whether the input graph violates the listed assumptions. These are called within the sample method before the sampling process itself starts. When any of the assumptions is violated an error message is displayed for the end-user about the wrong input and the sampling is halted.

### 4 EXPERIMENTAL EVALUATION

To evaluate the sampling algorithms implemented in Little Ball of Fur we perform a number of experiments on real world social networks and webgraphs. Details about these datasets are discussed in Subsection 4.1. We show how randomized spanning tree sampling can be used to speed up node embedding in Subsection 4.2 without reducing predictive performance. Our ablation study about graph classification in Subsection 4.3 demonstrates how connected graph subsampling can accelerate the application of whole graph embedding techniques. We present results about estimating graph level descriptive statistics in Subsection 4.4.

<table>
<thead>
<tr>
<th></th>
<th>Facebook Page-Page</th>
<th>Wikipedia Crocodiles</th>
<th>LastFM Asia</th>
<th>Deezer Hungary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>22,470</td>
<td>11,631</td>
<td>7,624</td>
<td>47,538</td>
</tr>
<tr>
<td>Density</td>
<td>0.0007</td>
<td>0.0025</td>
<td>0.0010</td>
<td>0.0002</td>
</tr>
<tr>
<td>Transitivity</td>
<td>0.2323</td>
<td>0.0261</td>
<td>0.1786</td>
<td>0.0929</td>
</tr>
<tr>
<td>Diameter</td>
<td>15</td>
<td>11</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Labels</td>
<td>4</td>
<td>2</td>
<td>18</td>
<td>84</td>
</tr>
</tbody>
</table>

### 4.1 Datasets

We use real world social network and webgraph datasets to compare the performance of sampling procedures and test their utility for speeding up classification tasks.

#### 4.1.1 Node level datasets

The datasets used for graph statistic estimation and node classification are all available on SNAP [21], and descriptive statistics can be found in Table 1.

- **Facebook Page-Page** [35] is a webgraph of verified official Facebook pages. Nodes are pages representing politicians, governmental organizations, television shows and companies while the edges are links between the pages. The related task is multinomial node classification for the 4 page categories.
- **Wikipedia Crocodiles** [35] is a webgraph of Wikipedia pages about crocodiles. Nodes are the pages and edges are mutual links between the pages. The potential task is binary node classification.
- **LastFM Asia** [39] is a social network of LastFM (English music streaming service) users who are located in Asian countries. Nodes are users and links are mutual follower relationships. The task on this dataset is multinomial node classification – one has to predict the location of the users.
• Deezer Hungary [36] is a social network of Hungarian Deezer (French music streaming service) users. Nodes are users located in Hungary and edges are friendships. The relevant task is multi-label multinomial node classification - one has to list the music genres liked by the users.

4.1.2 Graph level datasets. Our classification study on subsampled sets of graphs utilized forum threads and small sized social networks of developers. The respective descriptive statistics of these datasets are in Table 2.

• Reddit Threads 10K [37] is a random subsample of 10 thousand graphs from the original Reddit threads datasets. Threads can be discussion and non-discussion based and the task is the binary classification of them according to these two categories.

• GitHub StarGazers [37] is a set of small sized social networks. Each social network is a community of developers who starred a specific machine learning or web development package on Github. The task is to predict the type of the repository based on the community graph.

Table 2: Descriptive statistics and size of the graph datasets for graph subsampling and whole graph classification.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Graphs</th>
<th>Nodes</th>
<th>Density</th>
<th>Diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reddit Threads 10K</td>
<td>10,000</td>
<td>11</td>
<td>97</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.291</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>GitHub StarGazers</td>
<td>12,725</td>
<td>10</td>
<td>957</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.561</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18</td>
</tr>
</tbody>
</table>

4.2 Node classification with randomly sampled spanning tree embeddings of networks

Node embedding vectors [31, 32] are useful compact descriptors of vertex neighbourhoods when it comes to solving classification problems. In traditional classification scenarios the whole graph is used to learn the node embedding vectors. In this experiment we study a situation where the embedding vectors are learned from a randomly sampled spanning tree of the original graph. We compare the predictive value of node embeddings learned on the whole graph with ones learned from spanning trees extracted with randomized BFS [17], DFS [17] and LERW [48]. The main advantage of randomized spanning trees is that storing the whole graph requires \(O(|E|)\) memory when we learn the node embedding. In contrast storing a sampled spanning tree only requires \(O(|V|)\) space.

4.2.1 Experimental settings. The experimental pipeline which we used for node classification has four stages.

1) Graph sampling. The BFS, DFS and LERW sample based embeddings start with the extraction of a random spanning tree using Little Ball of Fur. This sample is fed to the embedding procedure.

2) Upstream model. The sampled graph is fed to the unsupervised upstream models DeepWalk [31] and Walklets [32] which learn the neighbourhood preserving node embedding vectors. We used the Karate Club [37] implementation of these models with the default hyperparameter settings.

3) Downstream model. We inputted the node embedding vectors as input features for a logistic regression classifier – we used the scikit-learn implementation [29] with the default hyperparameter settings. The downstream models were trained with a varying ratio of nodes.

4) Evaluation. We report average AUC values on the test set calculated from a 100 seeded sampling, embedding and downstream model training runs.

![Graph classification performance on the Facebook Page-Pe page graph [35] evaluated by average AUC scores on the test set calculated from a 100 seeded experimental runs.](image1)

![Graph classification performance on the LastFM Asia graph [39] evaluated by average AUC scores on the test set calculated from a 100 seeded experimental runs.](image2)

Figure 4: Node classification performance on the Facebook Page-Pe page graph [35] evaluated by average AUC scores on the test set calculated from a 100 seeded experimental runs.

4.2.2 Experimental results. We report the predictive performance for the Facebook Page-Pe page and LastFM Asia graphs respectively on Figures 4 and 5. First, we see that the features extracted from the BFS, DFS and LERW sampled spanning trees are less valuable for node classification based on the predictive performance on these two social networks. In plain words node embeddings of randomly sampled spanning trees produce inferior features. Second, the marginal gains of additional training data are smaller when the embedding is learned from a subsampled graph. Third, DFS sampled node embedding features have a considerably lower quality compared to the BFS and LERW sampled node embedding features. Finally, the Walklets based higher order proximity preserving embeddings have a superior predictive performance compared to the DeepWalk based ones even when the graph being embedded is a randomly sampled spanning tree of the source graph.

Figure 5: Node classification performance on the LastFM Asia graph [39] evaluated by average AUC scores on the test set calculated from a 100 seeded experimental runs.

4.3 An ablation study of graph classification

Graph classification procedures use the whole graph embedding vectors as input to discriminate between graphs based on structural patterns. Using subsamples of the graphs and extracting structural patterns from those can speed up this classification process. We will investigate how exploration based sampling techniques perform when they are used to obtain the samples used for the embedding.
4.3.1 Experimental settings. The data processing which we used for evaluating the performance of graph classification has four stages.

(1) Graph sampling. We sample both datasets using the RW [8] and RWR [19] methods implemented in Little Ball of Fur 100 times for each retainment rate with different random seeds. Using these algorithms ensures that the graphs’ connectivity patterns are unchanged.

(2) Upstream model. Following the sampling, all of the samples are embedded using the Graph2Vec [27] algorithm. This procedure uses the presence of subtrees as structural patterns.

(3) Downstream model. With the embedding vectors as covariates, we estimate a logistic regression for each dataset and retainment rate. We rely on the scikit-learn implementation [29] of the classifier with the default hyperparameter settings. We use 80% of the graphs to train the classifier.

(4) Evaluation. The classification performance is evaluated using the AUC based on the remaining 20% of graphs which form the test set.

4.3.2 Experimental results. We report mean AUC values along with a standard deviation band in Figure 6 for the Reddit Threads and GitHub Stargazers datasets with the Random Walk and Random Walk with Restart sampling methods. Lower retainment rate is associated with a lower classification performance, as it can be expected. The more ragged, step function-like pattern observed in case of the Reddit threads dataset is likely to be due to the interplay of structural pattern downsampling and the generally smaller graphs in the dataset.

We report the mean runtime of the graph embedding process with a band of standard deviations in Figure 7. As we decrease the retainment rate, a significant decrease in runtime is prevalent. There is a clear trade-off between runtime and predictive performance. It is, however, worth noting that while the runtime associated with the 50% retainment rate is, in most cases close to half of the runtime using the whole graphs, the loss in classification power in case of the Reddit Threads dataset is less significant.

4.4 Estimating descriptive statistics

A traditional task for the evaluation of graph sampling algorithms is the estimation of graph level descriptive statistics. The graph level descriptive statistic is calculated for the sampled graph and it is compared to the ground truth value which is calculated based on the whole set of nodes and edges. A well performing sampling procedure is ought to give a precise estimate for the graph level quantity of interest with a reasonably small subsample of the graph.

4.4.1 Experimental settings. The pipeline used for estimating the graph level descriptive statistics had two stages.

(1) Graph sampling. Node and exploration based sampling procedures sample 50% of vertices, while the edge sampling techniques select 50% of the edges to extract a subgraph.

(2) Descriptive statistic estimation. We calculated the average of the clustering coefficient (transitivity), average node degree and the degree correlation for the sampled graphs. We did 10 seeded experimental runs to get an estimate of the statistics and calculated the standard error around these averages.

4.4.2 Experimental results. The ground truth and estimated descriptive statistics are enclosed in Table 3 for all of the node level datasets. Blocks of rows correspond to node, edge, random walk based and non random walk based exploration sampling algorithms. In each block of methods bold numbers denote the best performing sampling technique (closest to the ground truth) in a given category for a specific estimated descriptive statistic and dataset.

We can make a few generic observations about the quality of descriptive statistic estimates. First, there is not a clearly superior sampling technique. This holds generally and within all of the main categories of considered algorithms. Specifically, there is not a superior node, edge or expansion based sampling procedure. Second, the induction based edge sampling techniques (TIES and PIES) give a good estimate of the statistics, but the induction step includes more than 50% of the edges. Because of this, the obtained good estimation performance is somewhat misleading as the majority of edges is still retained after the induction step. Third, edge sampling algorithms sometimes fail to estimate the direction of the degree correlation properly. Finally, the random walk based techniques generally tend to overestimate the average degree. This is not surprising considering that these are biased towards high degree nodes.
Table 3: Ground truth and estimated descriptive statistics of the web graphs and social networks. We calculated average statistics from 10 seeded experimental runs and included the standard errors below the mean. We included the ground truth values based on the whole graph (first block) with estimates obtained with node (second block), edge (third block) and exploration (fourth and fifth blocks) sampling algorithms. Bold numbers denote for each category the best estimate for a given dataset.

<table>
<thead>
<tr>
<th>Social Network</th>
<th>Clustering Coefficient</th>
<th>Average Degree</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>0.232</td>
<td>15.205</td>
<td>0.085</td>
</tr>
<tr>
<td>RN [42]</td>
<td>0.229</td>
<td>7.531</td>
<td>0.070</td>
</tr>
<tr>
<td>DRN [1]</td>
<td>0.261</td>
<td>21.514</td>
<td>0.080</td>
</tr>
<tr>
<td>PRN [19]</td>
<td>0.270</td>
<td>16.228</td>
<td>0.136</td>
</tr>
<tr>
<td>RE [17]</td>
<td>0.116</td>
<td>8.470</td>
<td>0.084</td>
</tr>
<tr>
<td>RNE [17]</td>
<td>0.092</td>
<td>7.602</td>
<td>0.075</td>
</tr>
<tr>
<td>HRNE [17]</td>
<td>0.081</td>
<td>7.194</td>
<td>0.005</td>
</tr>
<tr>
<td>TIES [2]</td>
<td>0.215</td>
<td>16.564</td>
<td>0.083</td>
</tr>
<tr>
<td>PIES [2]</td>
<td>0.231</td>
<td>15.357</td>
<td>0.087</td>
</tr>
<tr>
<td>RW [8]</td>
<td>0.255</td>
<td>22.355</td>
<td>0.073</td>
</tr>
<tr>
<td>RWR [19]</td>
<td>0.235</td>
<td>20.282</td>
<td>0.092</td>
</tr>
<tr>
<td>RWJ [34]</td>
<td>0.271</td>
<td>18.615</td>
<td>0.123</td>
</tr>
<tr>
<td>MHRW [15, 43]</td>
<td>0.280</td>
<td>17.903</td>
<td>0.134</td>
</tr>
<tr>
<td>RC-MHRW [24]</td>
<td>0.266</td>
<td>21.070</td>
<td>0.106</td>
</tr>
<tr>
<td>FRW [34]</td>
<td>0.663</td>
<td>5.745</td>
<td>0.069</td>
</tr>
<tr>
<td>CNBW [50]</td>
<td>0.255</td>
<td>22.355</td>
<td>0.072</td>
</tr>
<tr>
<td>CNARW [25]</td>
<td>0.259</td>
<td>21.157</td>
<td>0.076</td>
</tr>
<tr>
<td>NBT-RW [18]</td>
<td>0.257</td>
<td>22.533</td>
<td>0.076</td>
</tr>
<tr>
<td>SB [10]</td>
<td>0.238</td>
<td>20.671</td>
<td>0.069</td>
</tr>
<tr>
<td>FF [20]</td>
<td>0.238</td>
<td>19.219</td>
<td>0.079</td>
</tr>
<tr>
<td>CSE [26]</td>
<td>0.229</td>
<td>13.116</td>
<td>0.070</td>
</tr>
<tr>
<td>SP [33]</td>
<td>0.221</td>
<td>12.842</td>
<td>0.106</td>
</tr>
</tbody>
</table>

5 CONCLUSION AND FUTURE DIRECTIONS

In this paper we described Little Ball of Fur – an open-source Python graph sampling framework built on the widely used scientific computing libraries NetworkX [12], NetworkX [41], and NumPy [47]. In detail it provides techniques for node, edge, and exploration based graph sampling.

We reviewed the general conventions which we used for implementing graph sampling algorithms in Little Ball of Fur. The framework offers a limited number of public methods, ingests and outputs data in widely used graph formats, and embodies preset default hyperparameters. We presented practical implications of these design features with illustrative examples of Python code snippets. Using various social networks and web graphs we had shown that using sampled graphs extracted with Little Ball of Fur one can approximate ground truth graph level statistics such as transitivity and the degree correlation coefficient. We also found evidence that sampling subgraphs with our framework can accelerate node and graph classification without extremely reducing the predictive performance.

As we have emphasized Little Ball of Fur assumes that the inputted graph is undirected and unweighted. In the future we envision to relax these assumptions about the input. We plan to include additional high performance backend libraries such as SNAP [22] and GraphTool [30]. Furthermore, we aim to extend our framework by including multiplex [7], attributed, and heterogeneous [23, 49] graph sampling algorithms with new releases of the framework.

ACKNOWLEDGEMENTS

Benedek Rozemberczki was supported by the Centre for Doctoral Training in Data Science, funded by EPSRC (grant EP/L016427/1).

REFERENCES
