# A generative model for sparse, evolving digraphs

Georgios Papoudakis, Philippe Preux, Martin Monperrus

# A generative model for sparse, evolving digraphs

Georgios Papoudakis and Philippe Preux and Martin Monperrus

**Abstract** Generating graphs that are similar to real ones is an open problem, while the similarity notion is quite elusive and hard to formalize. In this paper, we focus on sparse digraphs and propose SDG, an algorithm that aims at generating graphs similar to real ones. Since real graphs are evolving and this evolution is important to study in order to understand the underlying dynamical system, we tackle the problem of generating series of graphs. We propose SEDGE, an algorithm meant to generate series of graphs similar to a real series. SEDGE is an extension of SDG. We consider graphs that are representations of software programs and show experimentally that our approach outperforms other existing approaches. Experiments show the performance of both algorithms.

## 1 Introduction

We wish to generate artificial graphs that are similar to real ones: by "real", we mean a graph that is observed in the real world; as we know, there is ample evidence that graphs coming from the real world are not Gilbert, or Erdös-Rényi graphs, but exhibit more structure. The motivations range from pure intellectual curiosity to, for instance, being able to test ideas on a set of graphs when only one is available (the WWW, a social network), or understanding which are the key properties of a graph. This paper is considering directed, un-looped, un-weighted, sparse graphs of moderate sizes (number of nodes ranging from 100 to a couple of thousands of nodes); by sparse, we mean that the number of edges is of the order of the number of vertices, and typically scales like $aN$, with $a$ very small with regards to $N$ (say

Georgios Papoudakis e-mail: giwrpapoud@gmail.com
Philippe Preux e-mail: firstname.lastname@inria.fr
Université de Lille, CRIStAL & Inria, Villeneuve d'Ascq, France.

Martin Monperrus
KTH Royal Institute of Technology, Sweden e-mail: firstname.lastname@csc.kth.se

$a \leq 10$ to give an idea of its value). We assume weak connectivity of the graph. As a case study, we experiment with graphs extracted from software programs; beyond a better understanding of software programs, such graphs may be used *e.g.* to improve software development and track the sources of bugs [8, 7].

To generate a graph in this context, one may use an algorithm that builds a graph given the degree distribution of the real graph (see [5] and followers), or its adjacency matrix [3], or some other structure (see [11] and references therein). As we wish to understand and model the creation of the graph, and as real graphs are often dynamic, we are more interested in a second type of algorithms that build a graph incrementally. Another motivation is that we do not want to generate graphs that have the exact same number of vertices, or the exact same degree distribution, or anything identical to the real one. A reason for this is that if we consider the degree distribution, two graphs having the same degree distribution may be very different regarding their other properties; in the other way around, two graphs that have more or less slightly different degree distribution, may have very similar properties. The properties we are interested in are of various natures: connectivity, diameter, average path length, transitivity, modularity, assortativity, spectral properties, degree distribution. Furthermore, when considering degree distribution or spectral properties, it is not clear how to meaningfully measure the difference between two degree distributions: mean squared distance, Kolmogorov-Smirnov statistics, Kullback-Leibler divergence, Jensen-Shannon distance. Finally, an important property of the generator is its stability. We identify two types of stability: the first is that for a given set of parameters, the graphs that are generated should have approximately the same characteristics; the other is that the graphs generated by a set of parameters should not change too much when the value of the parameters change a bit (sort of continuity of the properties of the generated graphs in the space of parameters of the generator).

Modeling and generating static graphs is important, but we are really interested in modeling the evolution of a graph. Though some works exist [4], the issues mentioned above take yet another aspect when considering the evolution of a graph. We see the evolution of a graph as time series of graphs, that is a set of couples $\{(t,g)\}$. We wish to generate the whole series of graphs with a single algorithm. Succeeding in this endeavor, we would access to general properties of the graph and the evolution process, as well as being able to predict the next graphs.

The content of this paper is as follows: in section 2, we propose the Sparse Digraph Generator (SDG) which is an algorithm that generates graphs that fit our requirements; we then show that the degree distribution follows a power law distribution; we also show that the in-degree and the out-degree distributions are not identical, something often observed in real digraphs. Then, we put SDG to the test: we introduce the real graphs we work with and show how our generator performs. As we are interested in the modeling of the evolution of a dynamic graph, we introduce Sparse Evolving Digraph GEnerator (SEDGE) which is an incremental version of SDG in section 4 and put it to the test in section 5. Then, we conclude and draw some final remarks.

For the sake of reproducible research, all the experiments may be reproduced with the material freely available at `https://github.com/papoudakis/sparse-digraph-generator`.

## 2 The Sparse Digraph Generator: SDG

We present a novel algorithm that aims at generating sparse digraphs. It is outlined in algorithm 1. SDG starts by creating a digraph made of $N$ isolated nodes and then, at each iteration, it adds a link between two nodes. To add a link, SDG selects two nodes, one as output, and the other as an input node. The selection of either node is performed either at random or following a preferential attachment rule.

**Algorithm 1** Outline of SDG

1: ***Input:*** *Number of nodes: N*
2: ***Input:*** *Number of edges: E (assumed to be* $\ll N^2$*)*
3: ***Input:*** *Parameters* $e_1$ *and* $e_2$*, both in the range* $[0,1]$
4: ***Output:*** *Generated graph G*
5: $G \leftarrow$ *DiGraph (with N nodes and no edge)*
6: **for** t $\in \{1,...,E\}$ **do**
7: ▷ Selection of the node that the edge will start from
8: ***With*** *probability* $e_1$*: out* $\leftarrow$ *select_a_node_uniformly_at_random()*
9: ***Otherwise****: out* $\leftarrow$ *select_a_node_by_preferential_attachment*
10: ▷ Selection of the node that the edge will end to
11: ***With*** *probability* $e_2$*: in* $\leftarrow$ *select_a_node_of_in-degree_0()*
12: ***Otherwise****: in* $\leftarrow$ *select_a_node_by_preferential_attachment*
13: *G.add_edge(out, in)*
**return** *G*

We consider sparse digraphs in which the number of edges $E$ is $aN$, where $a \in (1,10)$. Such digraphs are quite common in applications and they are quite specific with regards to their properties: for instance, there is usually a very small number of paths to navigate from one node to another. It is often the case that the in-degree and the out-degree distributions do not have the same shape. SDG achieves this: is $e_1 \neq e_2$, the parameters of the power law of in-degree and out-degree distributions are different.

The selection of a node to connect to or from is either uniformly at random (among all nodes at line 8, among nodes of in-degree 0 at line 11), or with a probability proportional to the degree of the node, that is we use a linear preferential attachment rule.

In the rest of this section, we derive the form of the in-degree and out-degree distributions resulting from SDG. We show that both distributions follow a power law, though of different parameters.

## 2.1 The in-degree distribution

After the completion of the $t^{\text{th}}$ iteration of SDG, the graph is made of $t$ edges. So, the probability for a node of degree $k$ to be selected by linear preferential attachment is $\frac{k}{t}$. Additionally, we assume that $e_2 < \frac{N}{E}$, so the expected number of nodes that have in-degree 0 is bigger than 0, $\mathbb{E}[N - e_2 E] > 0$

Let $D_k(t)$ be the number of nodes with in-degree $k$ at timestep $t$. For $k > 1$, $D_k(t)$ decreases at timestep $t$ only if a node with in-degree $k$ is selected due to preferential attachment (line 12). So the probability that $D_k$ decreases at iteration $t$ is:

$$\underbrace{(1-e_2)}_{\text{probability of selecting a node by preferential attachment}} \underbrace{\frac{k}{t}}_{\text{probability of choosing a degree k node}} D_k(t) \quad (1)$$

Similarly, $D_k(t)$ increases only if a node with in-degree $k-1$ is selected due to preferential attachment. So the probability that $D_k$ increases at iteration $t$ is:

$$(1-e_2)\frac{(k-1)}{t} D_{k-1}(t) \quad (2)$$

Let $d_k(t) = \mathbb{E}[D_k(t)]$. It follows that the expected change in the number of nodes of degree $k$ at iteration $t$ is:

$$d_k(t+1) - d_k(t) = (1-e_2)\frac{(k-1)d_{k-1}(t) - k d_k(t)}{t} \quad (3)$$

We set $c_2 = 1 - e_2$ and we assume that $d_k(t) = p_k t$ so we get:

$$p_k = c_2((k-1)p_{k-1} - k p_k) \quad (4)$$

$$p_k = \left(1 - \frac{(1+c_2)/c_2}{1/c_2 + k}\right) p_{k-1} \quad (5)$$

Assuming that $k \gg \frac{1}{c_2}$ and using the binomial approximation we come up with:

$$p_k \approx \left(1 - \frac{(1+c_2)/c_2}{k}\right) p_{k-1} \approx \left(\frac{k-1}{k}\right)^{\frac{1+c_2}{c_2}} p_{k-1} \quad (6)$$

Finally, we calculate the values of $p_0$ and $p_1$ and we iterate the equation until $k = 2$.

$$p_k \approx \left(\frac{k-1}{k}\right)^{\frac{1+c_2}{c_2}} \left(\frac{k-2}{k-1}\right)^{\frac{1+c_2}{c_2}} \ldots \left(\frac{1}{2}\right)^{\frac{1+c_2}{c_2}} p_1 \quad (7)$$

$$p_k \approx p_1 k^{-\frac{1+c_2}{c_2}} \quad (8)$$

## 2.2 The out-degree distribution

In this section, $D_k(t)$ is the number of nodes with out-degree $k$ at iteration $t$. Starting with the same assumptions as before, we can write that the number of nodes with out-degree distribution $k$ decreases if a node with out-degree $k$ is selected due to preferential attachment with probability $1-e_1$ or if such a node is selected from a uniform distribution with probability $e_1$. This second possibility is different from the analysis we did for the in-degree distribution. So, the probability that $D_k$ decreases at iteration $t$ is:

$$e_1\frac{D_k(t)}{n}+(1-e_1)k\frac{D_k(t)}{t} \tag{9}$$

Similarly, $D_k(t)$ increases with probability:

$$e_1\frac{D_{k-1}(t)}{n}+(1-e_1)(k-1)\frac{D_{k-1}(t)}{t} \tag{10}$$

After following the same steps as before we end up with:

$$d_k(t+1)-d_k(t)=(1-e_1)\frac{(k-1)d_{k-1}(t)-kd_k(t)}{t}+e_1\frac{d_{k-1}(t)-d_k(t)}{N} \tag{11}$$

Assuming that the solution is like $d_k(t)=p_kt$ and by setting $c_1=1-e_1$, we can prove that at the final timestep $t=E$:

$$p_k\approx p_1(k+\frac{(1-c_1)}{c_1}\frac{E}{N})^{-\frac{1+c_1}{c_1}} \tag{12}$$

## 2.3 Discussion & Related Work

We have shown that the in-degree and the out-degree distributions of the graphs generated by SDG exhibit a power law. This may come as a surprise to the reader, well aware of earlier works, such as [1]. Indeed, our graph is not growing, keeping a set of $N$ nodes, connecting them along the iterations of the algorithm. However, the departure from a power law is expected when the number of iterations is approximately $N^2$, that is when the graph gets dense. However, as we emphasized it earlier, we only consider sparse graphs, and the number of iterations, hence the number of edges, remains $\mathcal{O}(N)$, hence much less than $N^2$.

It is worth noting that the power law coefficients of graphs generated by SDG are the same as those of graphs produced by Bollobas *et al.*, though the algorithms are slightly different. Actually Bollobas *et al.* results come as special cases of our analysis.

SDG departs from the usual Barabasi-Albert type of algorithms because it generates directed graphs. Strictly speaking, our algorithm generates a variant of a Price

graph [9] and setting $e_1$ to 0, $e_2$ to 1, a kind of Price's algorithm which adds one edge at a time is recovered. SDG comes very close to the one studied by Bollobas *et al.* [2] though only SDG is able to add two vertices at once, in a single iteration.

## 3 Experimental study of SDG

In this experimental section, we mainly study two questions:

- which algorithm performs the best to produce graphs that are similar to some real graphs?
- the stability of SDG with regards to its parameters.

We compare our algorithm with GDGNC [6] where it is shown to be the best graph generator available in the context of software graphs. We also compare our model with Bollobas *et al.*'s since they are quite similar: it is interesting to check how the small difference in these 2 algorithms convert into difference of performance. We have compared SDG with other algorithms (Kronecker graphs, ...) but since they perform poorly and due to space limitations, we do not report them. The experiments are performed with 10 major software programs taken from the maven dataset [10]. Table 1 summarizes the basic features of our dataset.

| Software (version) | Nodes | Edges | Edges/Nodes | Diameter |
|---|---|---|---|---|
| ant (1.5.1) | 266 | 1427 | 5.36 | 6 |
| findbugs (0.6.4) | 56 | 183 | 3.27 | 5 |
| freemarker (1.5.3) | 76 | 358 | 4.71 | 7 |
| hibernate (1.2) | 365 | 1916 | 5.25 | 7 |
| htmlunit (1.10) | 219 | 934 | 4.26 | 7 |
| jasperreports (3.1.2) | 1139 | 7460 | 6.54 | 7 |
| jparsec (0.2.2) | 75 | 203 | 2.71 | 5 |
| ojb (0.5.200) | 179 | 766 | 4.28 | 6 |
| pmd_jdk14 (4.1.1) | 521 | 3049 | 5.85 | 8 |
| spring_core (1.0.1) | 112 | 337 | 3.01 | 7 |

**Table 1** Statistics of the dataset used in the experiments reported in section 3.

In the literature, the measure of similarity between two graphs is not very well defined. In this paper, we measure the similarity between the generated graph ($gg$) and the original graph ($go$) using the following set of metrics:

- The Kolmogorov-Smirnov statistic (KS) of the in-degree and out-degree distributions. Let $CDF_g$ denote the cumulative degree distribution function of a graph $g$, so that $CDF_g(k) = \sum_{i \leq k} D_k$ where $D_k$ is the degree distribution of graph $g$. Then, $KS = \max_k |CDF_{gg}(k) - CDF_{go}(k)|$. We denote $KS_{in}$ (resp. $KS_{out}$) the KS statistics regarding in-degree (resp. out-degree) distribution.
- The mean squared distance (MSD) of the sorted in-degree and out-degree distributions. For each generated graphs $g$ we consider the in-degree and out-degree

of each node, sort these two lists to obtain $d_{in,g}$ and $d_{out,g}$. Then: $MSD_{in} = \frac{\sum_i (d_{in,gg}(i) - d_{in,go}(i))^2}{N}$ and $MSD_{out} = \frac{\sum_i (d_{out,gg}(i) - d_{out,go}(i))^2}{N}$.
The MSD can only be used for SDG and GDGNC because they generate the same number of nodes as the original graph. On the contrary, Bollobas *et al.*' model does not necessarily produce graphs with the same number of nodes.

We perform a grid search in order to determine the parameters of each model that best fit for each graph. SDG and GDGNC are optimized to minimize the maximum value between $MSD_{in}$ and the $MSD_{out}$: $minimize\{max(MSD_{in}, MSD_{out})\}$. As $MSD_{in}$ and $MSD_{out}$ are irrelevant for it, Bollobas *et al.* model is optimized to minimize the KS statistic. This may be seen as a caveat in our experiments, but we provide ample observations below to convince the reader that if we were tuning the parameters of the 3 models with the same metrics, the conclusions of the experiments would not change much. The experiments presented in table 2 below are performed with the optimal parameters for each software, averaged over 100 generated graphs. Table 2 provides the average value of KS and MSD for each model and each software.

| | $KS_{in}$ | | | $KS_{out}$ | | | $MSD_{in}$ | | $MSD_{out}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| Software | SDG | GDGNC | Bollobas | SDG | GDGNC | Bollobas | SDG | GDGNC | SDG | GDGNC |
| ant | 0.26 | **0.24** | 0.39 | **0.16** | 0.17 | 0.34 | **17.4** | 30.45 | **1.89** | 2.58 |
| findbugs | **0.29** | 0.41 | 0.37 | **0.33** | 0.35 | 0.37 | **2.32** | 3.74 | **1.24** | 2.65 |
| freemarker | **0.23** | **0.23** | 0.4 | 0.48 | 0.49 | **0.38** | **3.11** | 6.14 | **4.89** | 6.76 |
| hibernate | 0.38 | 0.41 | **0.33** | **0.22** | 0.32 | 0.32 | **14.38** | 21.87 | **3.14** | 9.23 |
| htmlunit | **0.37** | **0.37** | 0.42 | **0.31** | 0.36 | 0.44 | **12.67** | 20.68 | **3.92** | 8.45 |
| jasperreports | **0.24** | **0.24** | 0.28 | 0.35 | 0.43 | **0.29** | **32.37** | 97.72 | **16.1** | 37.43 |
| jparsec | **0.22** | **0.22** | 0.41 | **0.36** | 0.47 | 0.42 | **0.69** | 2.72 | **4.6** | 9.98 |
| ojb | 0.26 | **0.25** | 0.44 | **0.21** | 0.27 | 0.4 | **3.6** | 6.36 | **0.77** | 3.41 |
| pmd_jdk14 | **0.27** | 0.28 | 0.28 | 0.5 | 0.56 | **0.41** | **14.92** | 114.9 | **32.08** | 54.54 |
| spring_core | **0.36** | 0.4 | 0.4 | **0.23** | 0.34 | 0.3 | **2.54** | 4.67 | **1.2** | 3.72 |

**Table 2** Comparison of SDG with GDGNC and Bollobas *et al.* in terms of MSD and KS for 10 Java software graphs. Bold faces indicate best results.

We can clearly see that SDG performs better than both GDGNC and Bollobas *et al.* model. Additionally, SDG is much more stable than the other models. That means that given the parameters of the generator the graphs that are produced are similar. In table 3, we give the average of the standard deviation for the experiments that appear in table 2.

| Model | $KS_{in}$ | $KS_{out}$ | $MSD_{in}$ | $MSD_{out}$ |
|---|---|---|---|---|
| SDG | $0.093 \pm 0.012$ | $0.084 \pm 0.023$ | $3.94 \pm 2.55$ | $1.33 \pm 1.07$ |
| GDGNC | $0.091 \pm 0.01$ | $0.081 \pm 0.013$ | $19.78 \pm 26.6$ | $4.01 \pm 3.89$ |
| Bollobas | $0.102 \pm 0.029$ | $0.099 \pm 0.025$ | | |

**Table 3** Mean and standard deviation of standard deviation values of MSD and KS on 10 Java software graphs.

From table 3 we can see the standard deviation values of SDG are on the same level or smaller than both GDGNC and Bollobas *et al.* But the most important property of SDG is that it can create graphs similar to the original one without the parameter optimization process, that both other models require in order to perform decently. For each software, we generate 100 graphs and we compute the average $KS_{in}$, $KS_{out}$, $MSD_{in}$, and $MSD_{out}$. All the experiments are performed with the same values $e_1 = 0.45$ and $e_2 = \frac{N}{E} - 0.05$ for all software graphs; these values result from our experiments. Table 4 provides the results; in ()'s, we report the ratio between the SDG without and with tuning: *e.g.*, $0.14(0.9)$ is the first row of column $KS_{out}$ means that $KS_{out}$ is $0.14$ without tuning, and $0.14/0.9$ with tuning. The value of $KS$ without tuning may be smaller than with tuning because the parameter tuning is performed to minimize $MSD$.

| Software | $KS_{in}$ | $KS_{out}$ | $MSD_{in}$ | $MSD_{out}$ |
|---|---|---|---|---|
| ant | 0.25 (1.0) | 0.14 (0.9) | 20.54 (1.2) | 0.89 (0.5) |
| findbugs | 0.3 (1.0) | 0.34 (1.0) | 2.66 (1.1) | 1.34 (1.1) |
| freemarker | 0.24 (1.0) | 0.46 (1.0) | 3.43 (1.1) | 5.31 (1.1) |
| hibernate | 0.29 (0.8) | 0.3 (1.4) | 27.16 (1.9) | 13.27 (4.2) |
| htmlunit | 0.33 (0.9) | 0.29 (0.9) | 12.84 (1.0) | 5.24 (1.3) |
| jasperreports | 0.21 (0.9) | 0.43 (1.2) | 119.42 (3.6) | 49.16 (3) |
| jparsec | 0.25 (1.1) | 0.42 (1.2) | 1.52 (2.2) | 8.41 (1.8) |
| ojb | 0.33 (1.3) | 0.27 (1.3) | 13.47 (3.7) | 2.36 (3.1) |
| pmd_jdk14 | 0.33 (1.2) | 0.54 (1.1) | 61.67 (4.1) | 45.43 (1.4) |
| spring_core | 0.3 (0.8) | 0.25 (1.1) | 2.63 (1.0) | 2.43 (2.0) |

**Table 4** MSD and KS without tuning parameters: numbers in ()'s gives the ratio between the measurement without tuning and the measurement with tuning.

From table 4 we see that in most cases, SDG, without parameter tuning, performs better than both GDGNC and Bollobas *et al.* model after parameter tuning. Another very nice property is that the performance does not change very much as the value of a parameter is changing: there is some sort of continuity of the performance of SDG with regards to the value of parameters. This is a very nice property, as this implies that to tune the parameters of SDG, a coarse grid search is enough and computationally cheaper.

Figure 1 provides a graphical illustration of these measurements: we plot the in-degree distribution, the out-degree distribution, and the spectra of the adjacency matrix for the real graph and for the graphs generated by each algorithm we compare to.

To conclude this part, let us stress that SDG uses two pieces of information: the number of nodes $N$ and the number of edges $E$. We have shown that SDG produces graphs which degree distributions follow power laws. When we want to generate graphs similar to a real one, both $N$ and $E$ are available, and we have shown that $e_1$ and $e_2$, the parameters of SDG, are not that important to obtain satisfying graphs.

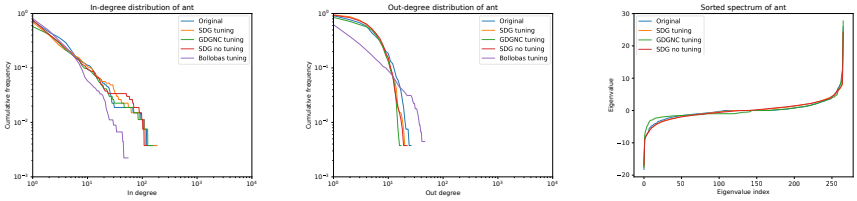


**Fig. 1** In-degree distribution, out-degree distribution, and spectrum for the real graph and the generated graphs.

Another point is that the occasional addition of 2 nodes instead of 1 seems beneficial since this is the only difference between SDG and Bollobas *et al.* approach.

Finally, it is important to refer to the metrics we use to compare graphs and the metrics we use to optimize the parameters of the algorithm. As said earlier, it is not known how to assess the similarity of two graphs using a single metric; instead, we use a series of metrics (and more may be used) to formalize the idea of similarity between two graphs. The metrics we use are recognized as very important to characterize a graph: the degree distribution, and the spectrum. We have found that optimizing using the degree distributions leads to better results. We see that as a primary observation, other spectral information might be used, and other properties may be used too. Furthermore, a combination of metrics may be optimized or used to judge the similarity: this is left as future work.

## 4 SEDGE: modeling the evolution of a graph

We consider a model of evolution of the real graph that is version-oriented. As the real graphs we consider are software, considering a sequence of versions of a software, the graphs along this sequence evolve by part: by that, we mean that the set of nodes and the set of edges evolve by chunks: from one graph to the next one (one version of a software to the next one), a set of nodes are added, some nodes are removed, and it is the same for the edges. So, we consider an algorithm that takes a graph as input, and then adds a set of nodes and a set of edges, possibly removing some existing nodes and edges.

We propose the "Sparse Evolving Digraph GEnerator" SEDGE (see algorithm 2), a model to capture the evolution of software graphs based on the generative model that we proposed in section 2. SEDGE is an extension of SDG. It distinguishes existing nodes from new nodes. At each timestep, SEDGE chooses two nodes to connect, sampling them from either set of nodes, based on 2 parameters that act as probabilities $\alpha$ and $\beta$.

**Algorithm 2** SEDGE: a generative model for sparse digraph evolution. The SAMPLE_A_NODE samples nodes in exactly the same way algorithm 1 does. new_nodes refers to the N_new nodes that are added to the current graph. all_nodes refers to all nodes of the new graph.

***Input:*** *Number of nodes to add N_new*
***Input:*** *Number of edges to add E_new*
***Input:*** *Parameters* $\alpha, \beta, e_1, e_2$*, all in the range* $[0,1]$
***Input:*** *Current graph G_cur*
***Output:*** *Generated graph G_new*
**function** SAMPLE_A_NODE(so, si, $e_1$, $e_2$)
    ***With*** *probability* $e_1$*: out* $\leftarrow$ *select_a_node_uniformly_at_random(so)*
    ***Otherwise****: out* $\leftarrow$ *select_a_node_by_preferential_attachment (so)*
    ***With*** *probability* $e_2$*: in* $\leftarrow$ *select_a_node_of_in-degree_0(si)*
    ***Otherwise****: in* $\leftarrow$ *select_a_node_by_preferential_attachment (si)*
    **return** (in, out)
**End function**
*G_new* $\leftarrow$ *G_cur.add_nodes(N_new)*
**for** t $\in \{1, ..., E\_new\}$ **do**
    ***With*** *probability* $\alpha$*:* (in, out) $\leftarrow$ SAMPLE_A_NODE(all_nodes, new_nodes, $e_1$, $e_2$)
    ***With*** *probability* $\beta$*:* (in, out) $\leftarrow$ SAMPLE_A_NODE(new_nodes, all_nodes, $e_1$, $e_2$)
    ***Otherwise****:* (in, out) $\leftarrow$ SAMPLE_A_NODE(all_nodes, all_nodes, $e_1$, $e_2$)
    *G_new.add_edge(out, in)*
**return** *G_new*

## 5 Experimental study of SEDGE

In this section, we evaluate the ability of SEDGE to capture the software evolution. For the experiments, we use 10 pairs of consecutive versions of software graphs[1] from the maven dataset. With the term "first graph", we refer to the first version of the software and with the term "second graph" to the second version. In each pair of these graphs, the second graph has at least 20% more nodes than the first graph.

The degree distributions and the spectrum of the graphs of two successive versions are close. For this reason, in order to perform a better evaluation of SEDGE we compute KS and MSD only for the new nodes: doing so, we amplify the difference between the two versions. In table 5, we report on the values of KS and MSD averaged over 100 experiments, for each real graph, given the optimal parameters of the model.

[1] (ant.1.4.1→ant.1.5), (commons_collections.20030418.083655→commons_collections.20031027.000000), (hibernate.2.0.3→hibernate.2.1.1), (jasperreports.0.6.7→jasperreports.1.0.0), (jasperreports.1.0.3→jasperreports.1.1.0), (ojb.0.8.375→ojb.0.9), (ojb.0.9.5→ojb.0.9.6), (spring.1.0 →spring.1.1), (wicket.1.0.3→wicket.1.1), (wicket.1.1.1 → wicket.1.2)

| First Software | $N_{new}$ | $E_{new}$ | $KS_{in}$ | $KS_{out}$ | $MSD_{in}$ | $MSD_{out}$ |
|---|---|---|---|---|---|---|
| ant.1.4.1 | 116 | 665 | 0.29 (0.8) | 0.4 (1.1) | 5.57 (2.9) | 2.37 (0.6) |
| commons.20030418 | 118 | 385 | 0.41 (1.1) | 0.39 (0.9) | 0.99 (1.5) | 1.05 (0.9) |
| hibernate.2.0.3 | 92 | 853 | 0.39 (0.6) | 0.29 (1.0) | 3.52 (12.4) | 3.22 (1.0) |
| jasperreports.0.6.7 | 170 | 1100 | 0.23 (1.1) | 0.2 (1.2) | 15.39 (1.2) | 6.08 (1.8) |
| jasperreports.1.0.3 | 117 | 1214 | 0.19 (0.9) | 0.25 (1.0) | 22.1 (2.7) | 9.2 (0.9) |
| ojb.0.8.375 | 100 | 555 | 0.31 (0.9) | 0.39 (1.0) | 5.72 (1.6) | 1.27 (1.0) |
| ojb.0.9.5 | 120 | 586 | 0.47 (1.0) | 0.36 (1.0) | 1.51 (0.6) | 2.4 (3.5) |
| spring.1.0 | 199 | 830 | 0.36 (1.0) | 0.4 (0.8) | 2.66 (3.7) | 1.17 (3.2) |
| wicket.1.0.3 | 96 | 569 | 0.36 (0.9) | 0.32 (1.0) | 1.21 (26.4) | 1.93 (1.4) |
| wicket.1.1.1 | 235 | 1800 | 0.25 (1.0) | 0.2 (1.1) | 4.92 (1.0) | 2.75 (2.8) |

**Table 5** MSD and KS for 10 evolutions of software graphs of SEDGE, averaged over 100 runs for each software. We also run the same experiments without tuning parameters: numbers in ()'s gives the ratio between the measurement without tuning and the measurement with tuning: a value below 1 means that it is better without tuning, above 1 that it is worse.

SEDGE has the same fundamental property SDG has: it can capture the structure of the evolved network without tuning its parameters. As in table 4, the values in ()'s in table 5 gives the ratio between tuning and no tuning. We use $\alpha = 0.5$, $\beta = 0.4$, $e_1 = 0.45$ and $e_2 = \frac{N}{E} - 0.05$ in the non tuned parameters experiment.

## 6 Conclusion and future work

In this paper, we consider the problem of generating graphs that are similar to real, sparse digraphs. We propose SDG which generates such graphs, exhibiting power law in their degree distributions. We show that SDG performs very well experimentally; furthermore, SDG is stable in terms of parameter tuning: we show that it behaves very well even if we do not perform parameter tuning. Then, we propose an extension named SEDGE which aims at generating series of sparse digraphs that is similar to a series of real graphs. The similarity between two graphs is not well defined; we have used different ways to measure it and we have discussed the influence on the final result of the generator. Other metrics can also be used and will be investigated in the future. We have used SDG and SEDGE with a type of graphs in mind; we have not defined these algorithms using any knowledge on the graphs being modeled: we have designed the algorithms, tested them on some real graphs, and observed the results. We think they may be used for many types of real graphs. More importantly, considering series of graphs is a very important aspect of our work. As real graphs are evolving, we think that we have to use dynamic models to deal with them to really capture something about the evolution of the real graph, and the understanding of the process underneath.

## *Acknowledgements*

This work was partially supported by CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, and the French Ministry of Higher Education and Research. We also wish to acknowledge the continual support of Inria, and the stimulating environment provided by the SequeL Inria project-team.

## References

1. Barabasi, A., Albert, R.: Emergence of scaling in random networks. Science **286** (1999)
2. Bollobas, B., Borgs, C., Chayes, J., Riordan, O.: Directed scale-free graphs. In: Proc. SODA, pp. 132–139 (2003)
3. Carstens, C.J., Berger, A., Strona, G.: Curveball: a new generation of sampling algorithms for graphs with fixed degree sequence (2016). Arxiv.org, 1609.05137
4. Holme, P.: Modern temporal network theory: a colloquium. The European Physical Journal B **88**(9) (2015)
5. Kleitman, D., Wang, D.: Algorithms for constructing graphs and digraphs with given valences and factors. Discrete Math. **6**(1), 79–88 (1973)
6. Musco, V., Monperrus, M., Preux, P.: A generative model of software dependency graphs to better understand software evolution (2015). Arxiv, 1410.7921
7. Musco, V., Monperrus, M., Preux, P.: Mutation-based graph inference for fault localization. In: Proc. SCAM, pp. 97–106 (2016)
8. Musco, V., Monperrus, M., Preux, P.: A large-scale study of call graph-based impact prediction using mutation testing. Software Quality Journal **25**(3), 921–950 (2017)
9. Newman, M.: The structure and function of complex networks. SIAM Review **45**(2), 167–256 (2003)
10. Raemaekers, S., Deursen, A.v., Visser, J.: The maven repository dataset of metrics, changes, and dependencies. In: Proc MSR, pp. 221–224. IEEE Press (2013)
11. Staudt, C.L., Hamann, M., Safro, I., Gutfraind, A., Meyerhenke, H.: Generating Scaled Replicas of Real-World Complex Networks, pp. 17–28. Springer International Publishing (2017)