EVALUATING A SUBLINEAR-TIME ALGORITHM FOR THE MINIMUM SPANNING TREE WEIGHT PROBLEM

GABRIELE SANTI* AND LEONARDO DE LAURENTIIS†

Abstract. We present an implementation and an experimental evaluation of an algorithm that, given a connected graph G (represented by adjacency lists), estimates in sublinear time, with a relative error e, the Minimum Spanning Tree Weight of G (see [1] for a theoretical exposure of the algorithm). Since the theoretical performances have already been shown and demonstrated in the above-mentioned paper of Chazelle et al. our goal is, exclusively, to experimental evaluate the algorithm and at last to present the results. Some technical insights are given on the implementation of the algorithm and on the dataset used in the test phase, hence to show how the experiment has been carried out even for reproducibility purposes; the results are then evaluated empirically and widely discussed, comparing these with the performances of the Prim algorithm and the Kruskal algorithm, launching several runs on a heterogeneous set of graphs and different theoretical models for them.

We assume hereafter that the reader has knowledge about the cited paper as we will just recap the theoretical results.

Key words. minimum spanning tree, sublinear time algorithms, randomized algorithm, approximation algorithm, minimum spanning tree weight, experimental evaluation

AMS subject classifications. 68W20, 68W25, 68R10

1. Introduction. We would discuss here some preliminary observations and assumptions. First of all, we observe that we need a set of graphs that satisfies the following points:
- they should be finite and should not be multigraphs;
- they should be undirected;
- they should have weighted edges;
- the weights on the edges should be integers; since the graph is finite, it is enough to show that there exist W such that it is the maximum weight on the edges of the graph;[1]
- they might contain self-loops;
- they have to be connected (the graph has only one connected component);
- they should be represented with adjacency lists;
- they should be represented in the same manner (we need an unvarying file format).

Unfortunately, the graphs and their representations which can easily be found on the Internet doesn't accomplish all of this requirements at the same time, although many standards for the file format are available.

Given this observation, our choice was to use randomly generated graphs, hence to implement our own graphs generator. This gives us the opportunity to generate a wide set of connected graphs, with tunable parameters, carefully chosen looking forward to the tests; these parameters include the number of nodes, the number of edges and the edges weight, nonetheless the distribution law for the edges. The edges between the nodes are step-by-step randomly constructed, respecting the connection

---

*Master's Degree in Computer Science, University of Rome "Tor Vergata" (gabriele.santi@acm.org).
†Master's Degree in Computer Science, University of Rome "Tor Vergata" (ldelaurentiis@acm.org).
[1]more generally, it is enough to have a numerable set of values for the weights which is always true when the graph is finite

requirement. The different types of graphs that we use in our experimental evaluation are presented afterwards.

After studying the paper we made some assumptions. One of the problem we encountered is that the theoretical algorithm assume to have in input the graph G and to have direct access to the family of graphs G_i[2]; with "direct" we intend that no computation is required to extract G_i, which is not true. In fact, we can show easily that a lower bound for the extraction of all the family is, at least, $O(m)$ (i.e. is linear on the number of edges). A naïve approach would cost $O(m)$ for the extraction of G_i for a given i, hence $O(wm)$ for the whole family; a better approach could order the edges in $O(m \log m)$ and build the family in a single pass on the edges, achieving $O(m + m \log m)$. Having in input only G, it would seem that the algorithm is responsible for the extraction of the family, but this is not desirable due to this lower bound that would sabotage the overall performance. We decided finally to consider this cost a part of the construction of the data structure of the graph, to be done prior to the call of the algorithm.

2. Design studies and choices.

2.1. Random Graphs Generator. As mentioned before, our choice is to implement our own graphs generator. The aim is to generate *connected random graphs* with a specific set of parameters, like the number of nodes, the number of edges, the maximum edges weight and the average degree. Moreover, we want to test how our algorithm behaves in different environments, so we want to control the distribution law of the edges. Keep in mind that the graph has to satisfy all the points of the previous section, among which the connectivity requirement. Given a desired number of nodes n and $e \geq n - 1$ edges, the key-concepts under the implementation of our connected graphs generator are the following:

- we begin by generating a random permutation of the vertices $v_0, v_1, \ldots v_n$
- we generate a random spanning tree by iteratively adding edges in this vector in a uniform random manner; suppose we have added τ vertices to the tree T, we randomly select a vertex s in the set $\{T_0, \ldots, T_\tau\}$ and add a new edge $< T_s, v_{\tau+1} >$. At the end we will have an acyclic graph with $n - 1$ edges.
- following a certain probability law, we add the remaining $e - (n - 1)$ edges

note that every time we add an edge, a weight is associated to it, with a value uniformly choosen in $[1, w]$.

We even decided to use a custom file format to save the data, wich is a `.ssv` file, which stands for *space separated values*: every line of the file corresponds to two edges, reporting the values $< v_s, v_t, w >$ that means source and target nodes, edge weight. Being the graph undirected, it is implicit that the edge $< v_t, v_s, w >$ also exists.

2.2. Graph Data Structures. We implemented two versions of the algorithm: the first using the well-known *Boost*'s BGL[3] for the graphs data structures and efficient implementations of Kruskal's and Prim's algorithms; the latter uses our own implementation of both the structures and all the side algorithms. We decided to do so because we wanted more control over the code, for testing purposes, and because at the moment (version `1.61.0`) the BGL subgraphs framework presents some still unfixed bug.

---

[2] we recall that G_i is an induced subgraph of G such that the maximum weight on his edges is i (e.g. G_w is exactly G)

[3] `Boost Graph Libray`, [3]

Unfortunately, our Kruskal algorithm, although using *union-find* structures with path compression[4], is not as fine-tuned as that proposed by the Boost's libraries; we didn't take further time on this because on the other side, our version of Prim's algorithm shows the *same exact performances* of the Boost version and it counts as a theoretical lower bound, being an optimal algorithm and way better than the Kruskal's solution. Our data structures have been called FastGraph for their optimization over the operation required for the test. In any way our data structures *can* nor *do* "help" the CRT[5] algorithm in improving his time complexity.

2.3. Tuning the algorithm. In the implementation of the algorithm, the graph stored in the text file is read into a `FastGraph` structure, which is of course a representation by adjacency lists. We want to compare the performances of CRT algorithm with a standard MST algorithm that in addition computes the total weight.
We want to emphasize now that the CRT algorithm is based on probabilistic assumptions; some parameters, that depend asymptotically on ε should be selected carefully in order to provide a good approximation very fast. These includes:
- r, the number of vertices uniformly choosen in the "approx-number-connected-components". This number is critical for the performances, as it directly determines the running time of the entire algorithm.
- C, the large costant that we use to pick the number of vertices determinants for the application of the original paper's Lemma 4.

Both these values largely affect the overall performances because they undirectly decide how many BFS will be carried out by the CRT algorithm; the BFSes represent the fundamental cost driver of the entire run.

Initially in our opinion, the choice of these parameters must depend on the number of vertices in a way that they are dynamic, keeping their dependency on ε. We tried to untie this bond to n but having static values for this parameters showed poor performances and a relative error exploding too fast as we grow the input instance.

As the paper says, the only requirement for r is that it is $O(\frac{1}{\varepsilon^2})$ and $C \in O(\frac{1}{\varepsilon})$.

The demonstration reported on the original paper bound these values in a way that helped us on choosing the right function to compute them; in fact, we have that[6]

$$r\sqrt{\frac{w}{n}} < \varepsilon < 1/2, \qquad \frac{C}{\sqrt{n}} < \varepsilon < 1/2$$

hence we choose

$$r = \left\lfloor \frac{\lfloor \sqrt{\frac{n}{w}}\varepsilon - 1 \rfloor}{\varepsilon^2} \right\rfloor \in O\left(\frac{1}{\varepsilon^2}\right)$$

$$C = \left\lfloor \frac{\lfloor \sqrt{n}\varepsilon - 1 \rfloor}{\varepsilon} \right\rfloor \in O\left(\frac{1}{\varepsilon}\right)$$

---

[4]it is known that the Kruskal algorithm time complexity is $O(m \log m)$, but can be improved to $O(m \log n)$ using union find structures with some euristic conveniently applied
[5]Short for B. Chazelle, R. Rubinfeld, and L. Trevisan.[1]
[6]see Theorem 7 and Theorem 8 of [1]

2.4. Random Sequences. Another problem we encountered was the random selection of k distinct vertices out of n with $k < n$; after some research, we found that this kind of problem cannot be solved in sublinear time on the number of vertices, which can't be done for the same reasons exposed in the introduction about the subgraph family. The problem here is that we can't extract *with no reinsertion k* values without using an external data structure; this structure has a linear cost for the maintainance that depends on n. A special case is that in which n is a prime that fulfills certain properties; the key idea is to make use of the properties of the quadratic residues of n^7. That permits us to extract a non-repeating random value in constant time in the range $0 \ldots n - 1$; sadly, this solution is not affordable here because we need a dynamic range for each call, so that we cannot fulfill such constraints for n.

The solution we found is to use Fisher-Yates sequences[8] which permits us to prepare in advance the sequences and then get different, distinct values at each call in constant time and with dynamic bounds. The cost of the preparation of those sequences is linear and is not considered in the total cost.

3. Implementation choices. We choose to implement the algorithm in C++ because we wanted a language that offers the advantages of an OOP language, not least a faster development, that had a standard library and sufficiently "near" the machine to have a tighter control over the performances and the memory occupation (e.g. performances high variance due to overhead of a VM or an interpreter). We considered that absolute performances are not important here, whereas relative performances of the CRT algorithm and other MST algorithms are the focus, so finally the choice of the language is something not of the greatest importance. The IDE tool used for the implementation is JetBrains Clion. We also used GitHub, as our code versioning control system.

Also, as already mentioned, we did use the Boost library to extend the STL that C++ already offers; we did implement FastGraph in place of BGL also to address memory issues in relation to the method Boost uses to store subgraphs of a graph; we in fact used an advanced method to store the family of subgraphs $G_i, i = 0, 1, \ldots, w$ that simply store the difference of vertices between them, $\Delta_{G_i} := V(G_i) - V(G_{i-1})$. It is always possibile to reconstruct every G_i because this family only has *induced* subgraphs, and this cost is not taken into account in the computation of the total time.

The main function of the implementation is in the AlgoWEB.cpp file. Launching the program from this file allows us to run either the CRT algorithm or the Prim algorithm,or the Kruskal algorithm, and to view either the running time or the computed weight of the Minimum Spanning Tree. It takes some argument in input, namely the file containing the graph, a suitable value for ε and the path where to save the results of the single run.

Besides, it is possible to run the random graph generator by the `grandom` utility we developed apart. It takes many arguments in input, that you can find just by calling

```
$> grandom -h|--help
```

3.1. Random Graphs Models. In order to give exhaustive results, we designed `grandom` to produce 4 classes of random graphs:

---

[7]this is widely used in cryptography for "format preserving encryption"; see [5] and [6]

[8]originally decribed in [4] and then redesigned for computer use in [7]

- Erdős-Rényi model, that builds random graphs using a uniform distribution of the edges over the vertices; his average degree $d \approx \frac{2m}{n}$;
- Gaussian model, that builds random graphs with a "cluster zone" where the edges are more frequent, hence the gaussian shape of the degree distribution; we have still $d \approx \frac{2m}{n}$;
- Barabási-Albert model, that builds random *scale-free* graphs. The average degree is $d \approx \frac{m}{n}$;
- Watts-Strogatz model, that builds random graphs with *small-world* properties and $d \approx \frac{2m}{n}$.

We want to emphasize here the fact that for the Barabási-Albert model the average degree results to be different respect to the other models; this is due to the algorithm the scientific literature offers to build such graphs. For the other models is always possible to add an arbitrary number of edges keeping the theoretical properties valid; having, on the contrary, for the Barabási-Albert model a certain probability p_k at each step to successfully add an edge, it is not possible to build a graph with an arbitrary number of edges; the user can solely give the number of vertices. But then again, the theory states that if the algorithm starts with a complete graph of m_0 vertices (hence $m_0 - 1$ edges), it will produce a Barabási-Albert graph whose average degree is scaled by this quantity. Our initial complete graph has $m_0 = 2$ vertices, so we will have $d \approx \frac{2m}{m_0 n} = \frac{m}{n}$. A little insight on the proof is the following: the distribution of the degree in the Barabási-Albert model is a power law with a cubic exponent. Fortunately, in that case this distribution has a well defined mean. Applying a little of arithmetic we can easily see the truthfulness of what stated previously.

This difference in the models has to be bore in mind when reading the test results, because when comparing the performances over a Barabási-Albert graph with n vertices and m edges and any other one of a different model with same number of vertices and edges, we will have different average degrees. Since the CRT algorithm is designed to only depend on the latter and not on m nor n, a multiplicative factor of 2 is to be taken into account.

Our implementation provide a *memory-aware*[9] subroutine that calculates the exact value of d at each run; the values obtained so far agree with the statements above.

4. Tests. The following section reports the results we had launching the algorithm over a variegate dataset of graphs, as described in the next paragraph.

4.1. Dataset. For each random graph model listed in the previous section we decided to produce a dataset; each dataset has a "family" of graphs that differs one from each other following a pattern for the parameters. Precisely, we composed the dataset grouping sets of graphs based on the value of the parameters, following the rules:

- build one set of graphs for each model (Uniform, Gaussian, Small-World, Scale-Free);
- every set contains in turn other sets, one for each selected value of n, i.e. the number of nodes (5000, 30000, 55000, 80000, 105000, 130000, 155000, 180000, 205000, 230000);
- then again, for each selected value of d, i.e. the average degree $(20, 100, 200)$; this way we also determine the desired value for m, since we have, for our models and for the fact that $m \propto n$, a proportion between the two. It's easy

---

[9]it calculates the average using the *progressive mean* technique, avoiding arithmetic overflow

to see that $d \xrightarrow{n \rightharpoonup +\infty} \frac{2m}{n}$, so we have the selected values for m $(10, 50, 100$ times $n)$[10];
- finally, a set for each selected value of w, i.e. the weight $(20, 40, 60, 80)$.

In conclusion we have the number of models $\times$ the number of selected values for n $\times$ the number of selected values for d (hence m) $\times$ the number of selected values for w for a total of $4 \cdot 10 \cdot 3 \cdot 4 = 480$ graphs. This dataset, made of plain text files as already described, is quite heavy: 42.1 GiB of data.

4.2. Runs. Using the dataset just described, we used a pattern for the runs, in order to have a complete view of the behavior of the algorithm in the domain of the various parameters; every single result consists of three plots, and we obtained, inspired by the structure of the dataset:
- a set of results for each model;
- this set containing in turn a set for each selected value of ε $(0.2, 0.3, 0.4, 0.49999)$;
- then a set for each selected value of $d/2 \simeq \frac{m}{n}$ $(10, 50, 100)$;
- finally a set for each selected value of w.

The first two plots report the absolute and relative error of the CRT result compared to the correct one, calculated with Prim's algorithm; the third report the trend of the used time. As we mentioned, we indeed used Kruskal too but it's not reported here for it was not faster respect to Prim's times and unuseful for the computation of the MST weight, since already done with Prim's method.

We had this way a total of $3 \cdot 4 \cdot 3 \cdot 4 = 144$ plots, or better 48 different cases of study. A comparison between those cases we consider meaningful finally concludes our work, as reported below in section 5.

As mentioned before, the parameters should be selected carefully in order to provide a good approximation very fast. In this sense ε plays a crucial role here: since the CRT algorithm is a probabilistic algorithm and ε is indeed the driver parameter either for performances and for the accuracy of the estimate MST weight, it does not make sense to choose values too small for this parameter, as the performances could dramatically degrades (as, indeed, is expected). So, although it could be of interest to study the right limit of $1/2$, we empirically noted that values of ε below 0.2 shows undesired behaviour of the CRT algorithm, either for the computed MST weights and for the running times. This is not unusual dealing with theoretical algorithms that shows *asymptotical* properties; the class this algorithm belongs to is known as *property testing algorithms*, whose requirement is to have a query complexity much smaller than the instance size of the problem. Given that and the fact that the algorithm is not required to compute an *exact* value, but to *estimate* a value *probabilistically near* to the correct one[11], we are not surprised that if the instance of the problem is small, the algorithm shows worse performances respect to a "deterministic" method. Because of this, results for $\varepsilon < 0.2$ were not comprehensive nor interesting and are not reported.

Given the intrinsically probabilistic nature of the CRT algorithm, we had to launch several runs on the same graph to have a good estimate of the CRT running time. For this purpose we decided to launch 10 runs for each graph of the dataset, and to take the average running time as the estimate of the running time; the amount of runs has been decided as a compromise between having a low variance between execution time's mode and mean values, and keeping a restrained amount of tests to do over the whole dataset. For the output value of the approximated MST weight we

---

[10]operatively we fixed the desired number of edges m, so d is a random variable with mean $\frac{2m}{n}$
[11]iff the property we are looking for is a *probabilistically checkable proof*, see [2]

instead took one over the ones computed, randomly, to preserve the information on the tolerance of the estimate.

5. Results. Following there are considerations about the meaningful line charts regarding the results.

We know that the CRT time complexity is $O(dw\varepsilon^{-2}\log\frac{dw}{\varepsilon})$, where d is the average degree of a graph; on the other hand, we know that the accuracy depends on ε also, so we expect an inversely proportional relation with the running time. Therefore what we expect is that:

- by increasing one or more of the parameters d, w, there should be a worsening of the average running time;
- keeping all the other parameters unchanged, if we consider an increase in the value of ε there must be an improvement of the running time (as well as a worsening of the result, though);
- viceversa we expect the opposite behaviour if we decrease d and/or w or decrease ε with the other parameters unchanged.

Let us call the above *crucial parameters*.

What we still need to know is what happens to the error; we should expect a direct proportion with the running times, so the above considerations could be also valid for the error. On the contrary, we'll see that this is not exactly respected.

First thing to show is that the CRT algorithm is sublinear in the number of edges, hence is better than any other *exact* algorithm. This can be easily seen in figures from 5.1a to 5.1c. For the rest of the plots listed from now on it is possibile to see, above each graph, the values of the parameters for the presented run.

It is interesting to note that the correct value, computed with Prim's algorithm, it's linear in the number of edges (figure 5.1b); we want to stress the fact that this is not the general case, and that this trend is due to the uniform distribution of the weights over the edges. We will dissect this point more in section 8.

Given that the CRT Algorithm respects the sub-linearity constraint, let's now see the variations that occur when selectively changing other parameters. For the sake of completeness, in figures 5.1 informations about Kruskal's runs are reported, yet we won't report them in the charts that follow.

5.1. Variations of crucial parameters.

5.1.1. Average degree d. Let us now see the behaviour for the variation of d; we will initially concentrate on the running time and solely for the uniform model. The selected values of ε and w are respectivey 0.3 and 40.

As we can see in figures 5.2a to 5.2c, there is a worsening in the performance *for small instances*: an increase of the average degree d is somewhat correlated to a loss of performances to the point that our property testing algorithm needs more time that the deterministic one; still, that seems to be true *under* a certain dimension of the instance of the graph, so that we don't lose the truthfulness of the theoretical time complexity because for a fixed d^* it will always be possibile to find empirically a certain number of edges $m^* \propto d^*$ beyond which the running time function is always below $C \cdot dw\varepsilon^{-2}\log\frac{dw}{\varepsilon}$ for a certain C.[12] We want here to highlight the crucial fact that the algorithm behaves better on *big* instances, where with "*big*" we refer to the parameters the performances depend on, namely d, w. Almost all the trends reported
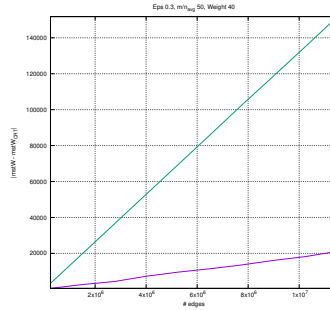
---

[12] that comes directly from the definition of *asymptotic complexity*
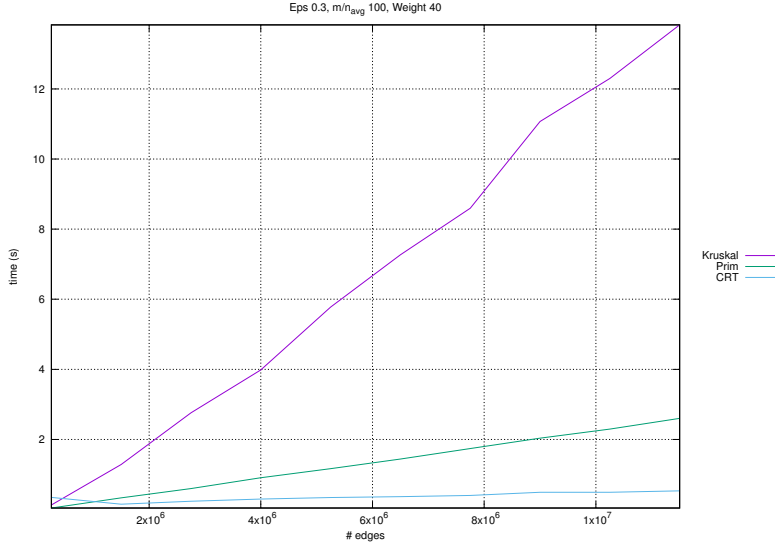
(a) *Running time; for completeness, Kruskal's time is also reported.*
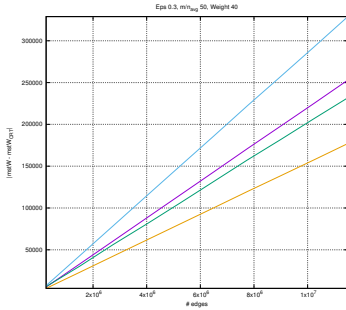
(b) *Relative error.*   (c) *Absolute error.*

Fig. 5.1: Sublinearity of the CRT algorithm and related error trends.

in this paper, in fact, show this initial "bad" curve and, from a certain value onward, a sublinear behaviour. We will discuss further about this in section 6 on page 20.

We would speculate that the error could be related to this. Indeed in figure 5.3 we can see that to an increase of d corresponds a dramatic annihilation of the error; to explain this point we use a simplified example. Let us consider a minimal complete graph[13]; the algorithm launches some BFSes on a *strict subset* of the graph, in more steps that could be interrupted according to certain rules, based in part on a stochastic process. It is easily provable that in this kind of graphs we have a worst case of $n-1$ hops between to vertices u and v; if we add a random edge to this graph, the length of this path decrease *at least* of one hop. By induction we can hence prove that the *diameter* of the graph decreases as $|E(G)|$ grows. In other words having a stronger connection within the graph (i.e. a greater probability to visit a vertex v from u in k hops of a random walk) increases the probability to have a complete view of the

---

[13]a connected graph with the least number of edges, i.e. $n-1$
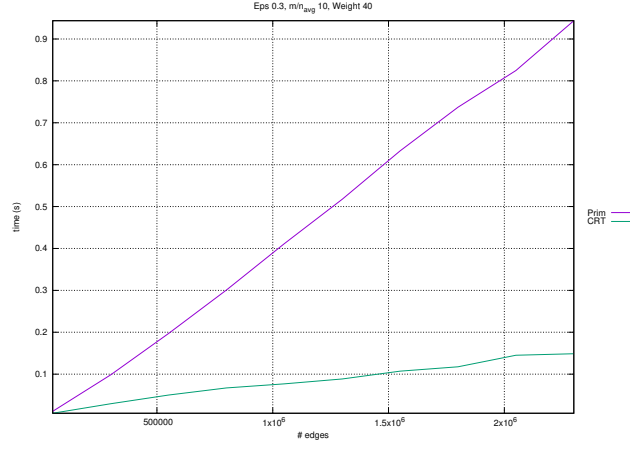
(a) $d \simeq 20$
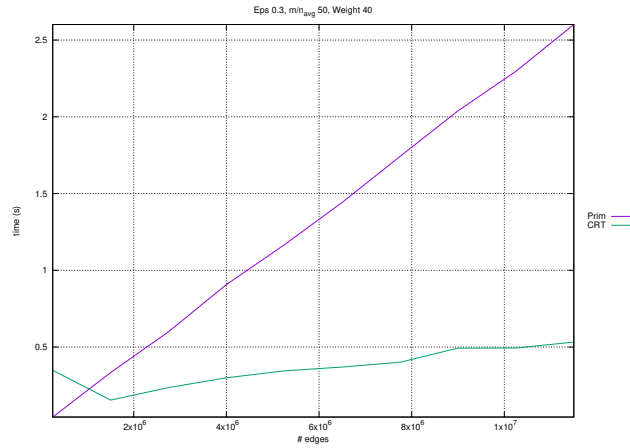
(b) $d \simeq 100$
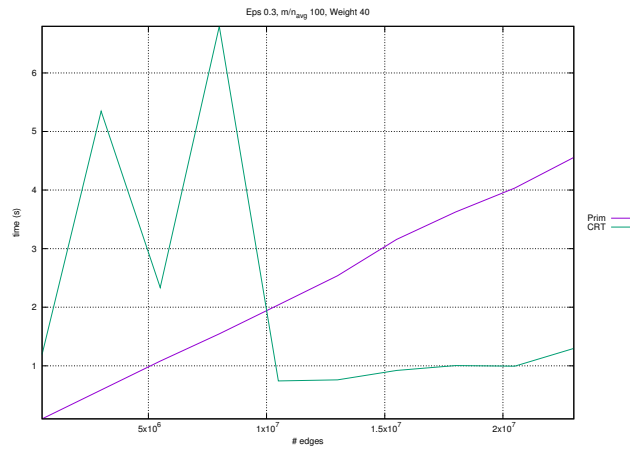
(c) $d \simeq 200$

Fig. 5.2: Behaviour for the increase of d. (Uniform model)
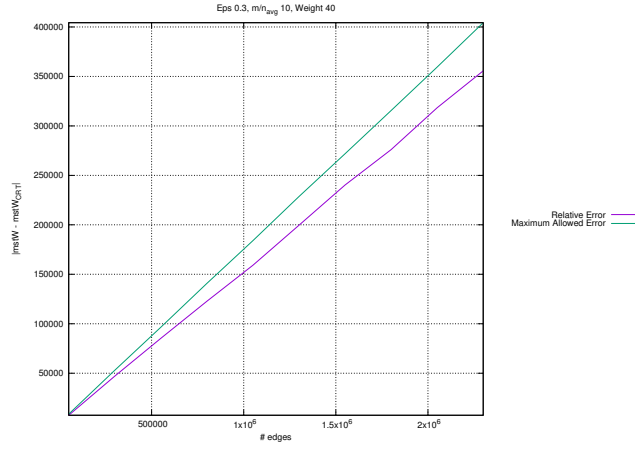
(a) $d \simeq 20$
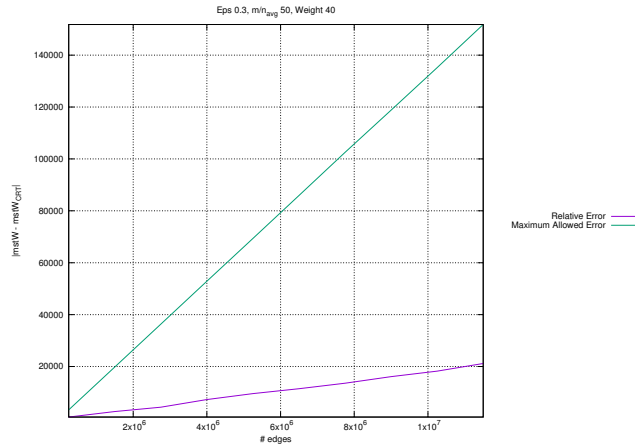
(b) $d \simeq 100$
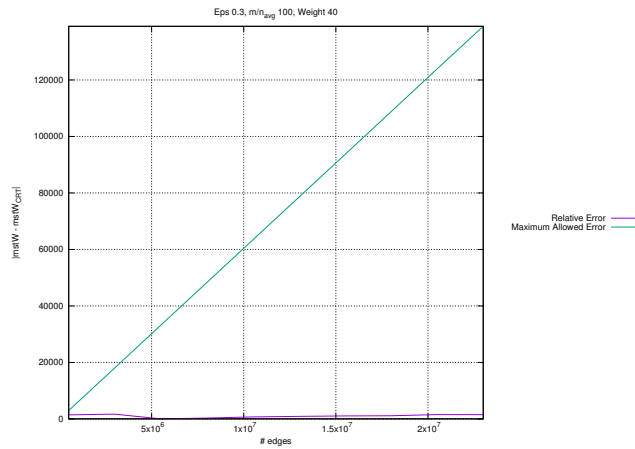
(c) $d \simeq 200$

Fig. 5.3: Error behaviour for the increase of d. (Uniform model)

graph, that is more *information* about the MST weight.

Moreover, we saw in our study of all the results showed in this paper, that the performances of the CRT are completely untied from the number of vertices n and from the number of edges m of the input graph; this suggests us also that the error is in turn driven solely by the parameters responsible of the algorithm's complexity, as the results that follow are in fact going to prove.

In figure 5.4 we summarize the results for the Gaussian and Small-World models, noticing that they equate the one we showed about the uniform model. This suggests that the algorithm complexity does not depend on the dimension and clustering coefficient of the graphs, being those the main differences from one model to another.

In figure 5.5 we summarize instead the trend of the relative error; we see here a slightly different evolution. We cannot conclude, as we did for the time complexity, that the error doesn't suffer from the different graph model. The error in fact depends on the clustering coefficient, because it is going to grow dependently on the number of not accomplished BFSes: each one of them cause in fact a loss of information. The algorithm, as well explained in [1], during the BFSes phase avoid to explore nodes that shows a high degree and even stops when encounters hubs.

In other words, having equals values of d in two different runs of the algorithm, we see that its time complexity trend remain the same; assuming that δ is the sample mean of the vertices degree, this tells us that the time complexity is bound to the *average* of δ and, since we have a growth of the error on graphs that contains hubs, we also conclude that the relative error is bound to the *variance* of δ.

5.2. Maximum weight w**.** Here we will manipulate the value of w, similarly to what we have already done with the average degree. Figures 5.6 and 5.7 are hereafter proposed; this time the other fixed values are $\varepsilon = 0.4$, $d = 50$. Still this graphs has been build using a uniform model.

This time we see the error growing as w increase. So we see here a *direct* proportion of the execution time with the maximum weight, unlike the *inverse* proportion it had with d. This is due to the fact that every iteration of the subroutine `approx-number-connected-components` that the reader can find in the original paper and remembered in pseudocode 1, adds a further approximation to the final result, because approximates the addend $\hat{c}$ that contributes to the total approximation, that is, the final error.

We see also that the dimension of the initial curve described here so far, grows proportionally to the worsening of the excution time's trend. This evidence also is observable in all the trends studied.

5.3. Error tolerance ε**.** We will test our algorithm for the values of $\varepsilon = 0.2, 0.3, 0.4, 0.49999$ over uniform generated graphs. As already explained, no values below 0.2 are investigated. We see in figures 5.8 and 5.9 the trends.

As expected, we do note that the time trends tend to decrease as ε increases, since we tolerate a higher error for the computed value, so the algorithm is less aggressive on computation and takes less time.
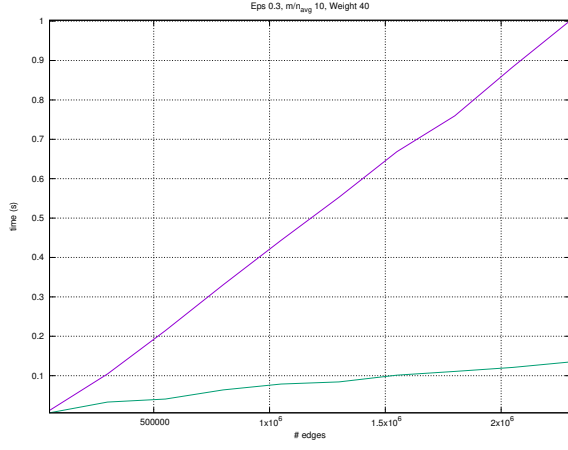
(a) gaussian, $d \simeq 20$
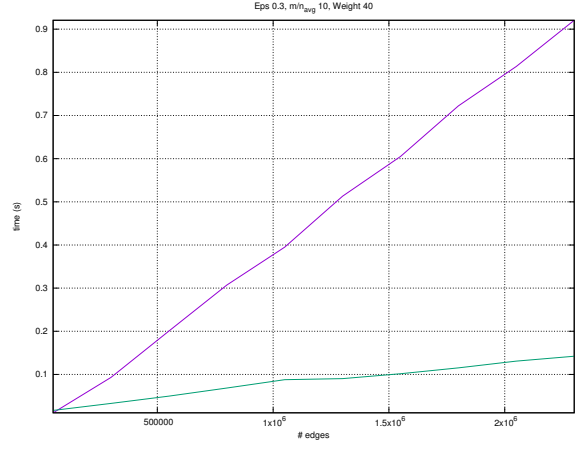
(b) small-world, $d \simeq 20$
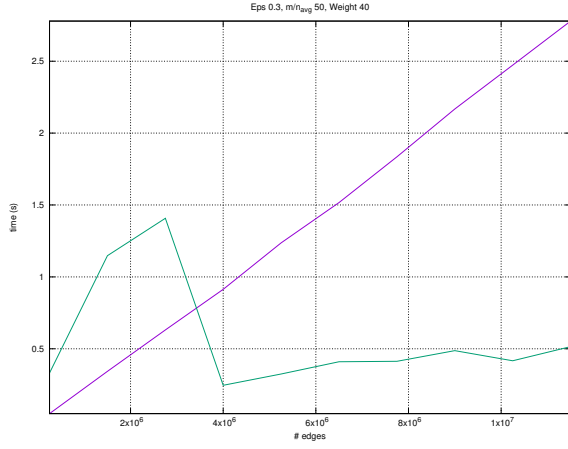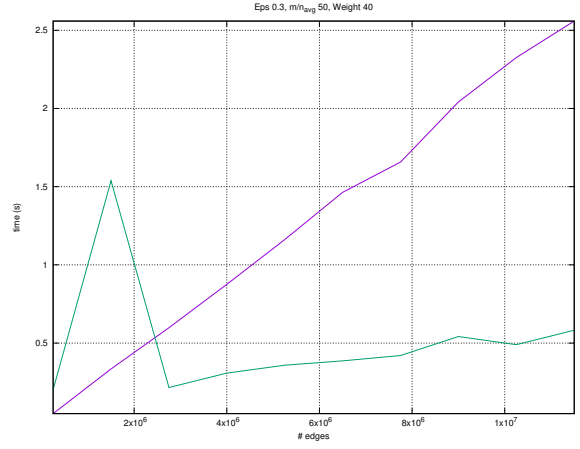
(c) gaussian, $d \simeq 100$
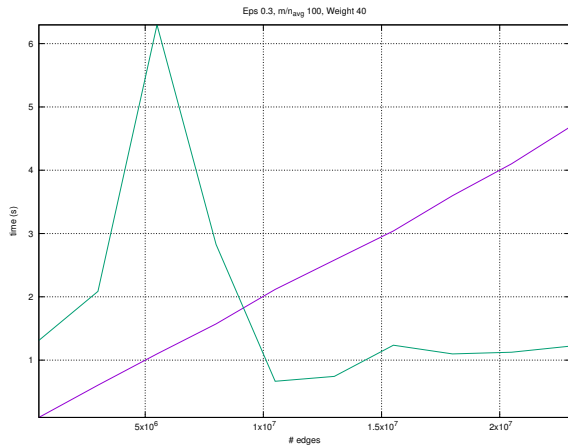
(d) small-world, $d \simeq 100$
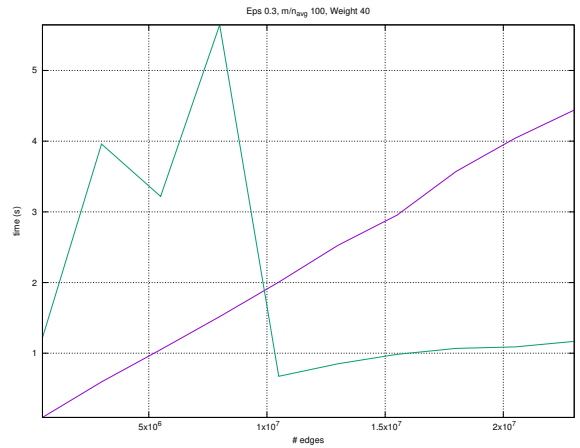
(e) gaussian, $d \simeq 200$

(f) small-world, $d \simeq 200$

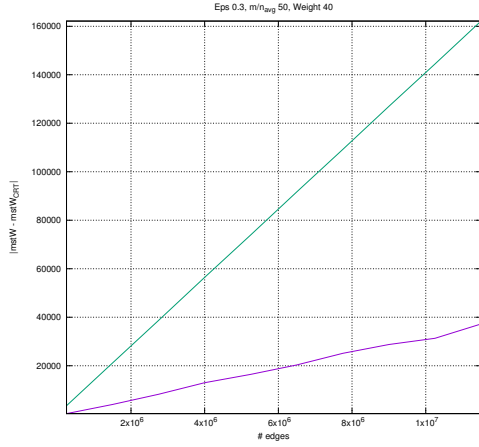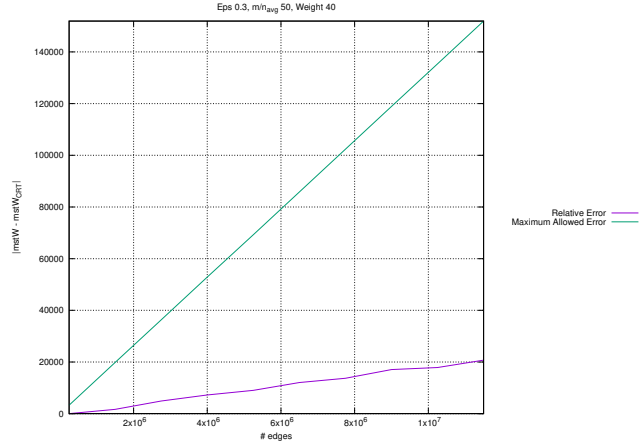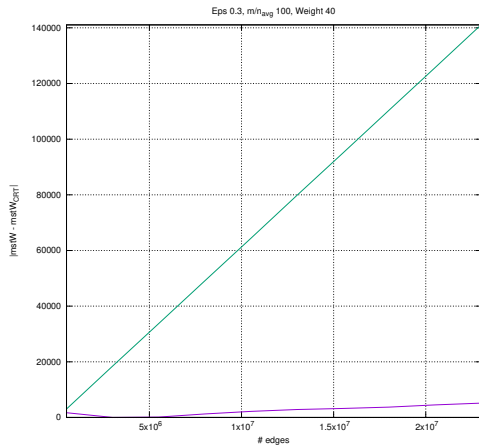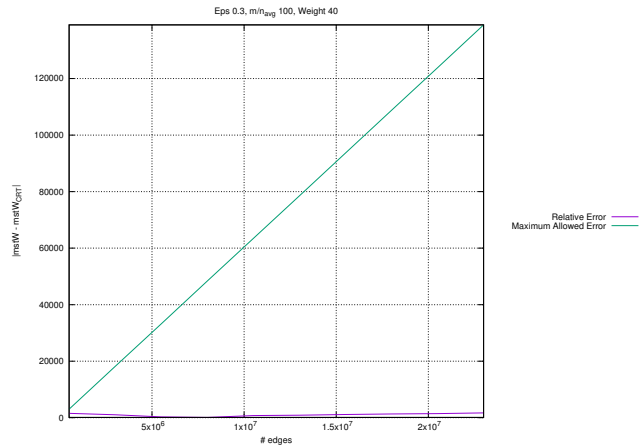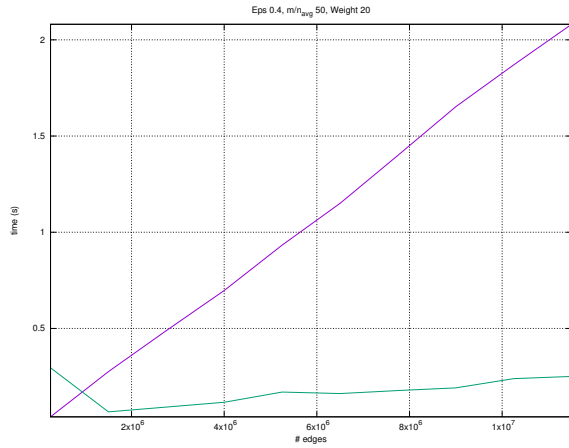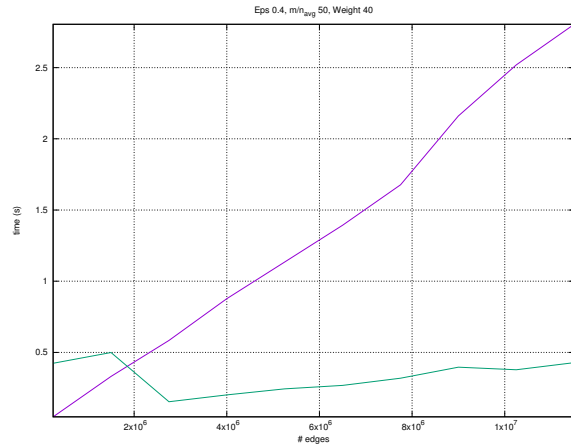Fig. 5.4: Execution times behaviour for the increase of d, different models.
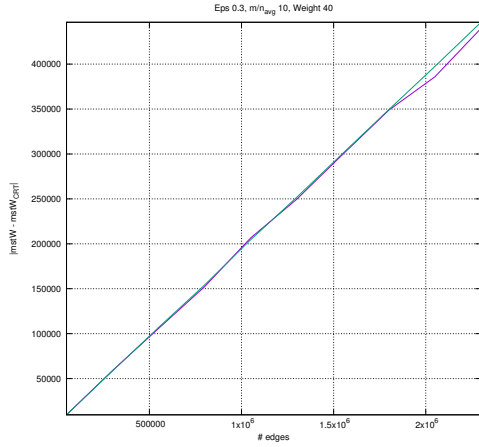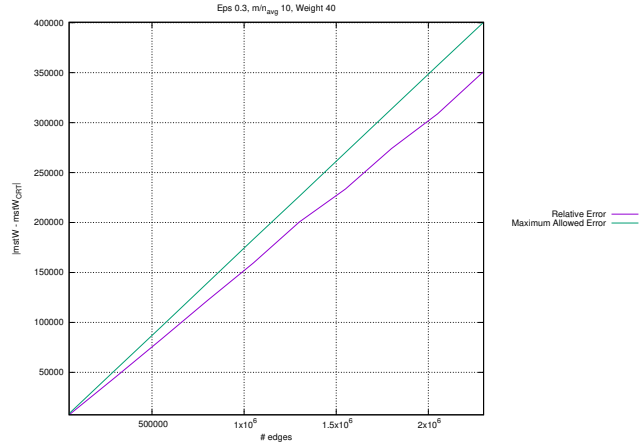
(a) gaussian, $d \simeq 20$

(b) small-world, $d \simeq 20$

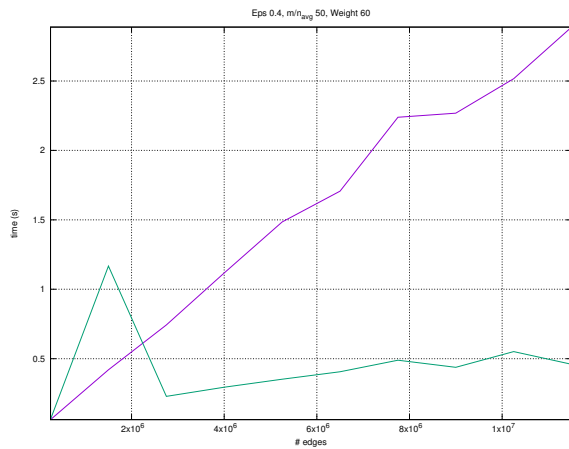(c) gaussian, $d \simeq 100$
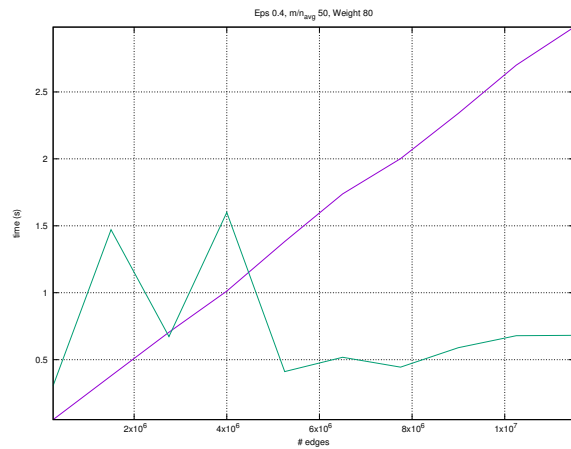
(d) small-world, $d \simeq 100$

(e) gaussian, $d \simeq 200$

(f) small-world, $d \simeq 200$

Fig. 5.5: Error behaviour for the increase of d, different models.

(a) $w = 20$

(b) $w = 40$

(c) $w = 60$

(d) $w = 80$

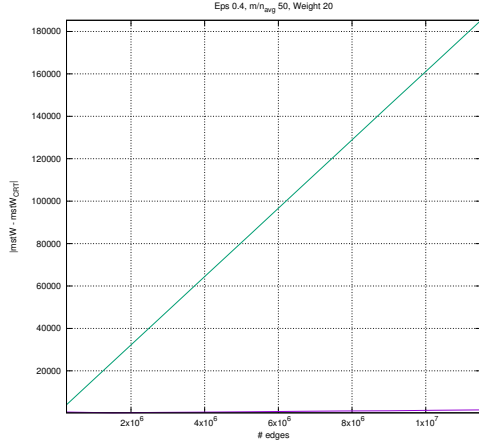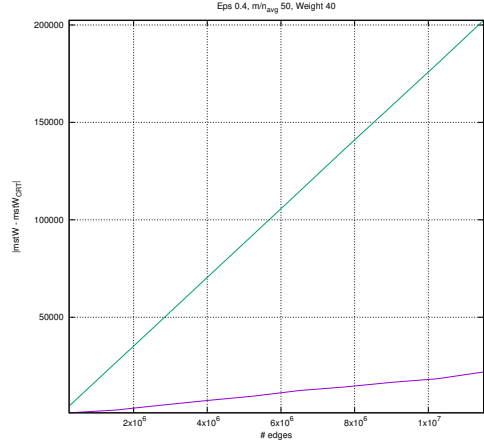Fig. 5.6: Behaviour for the increase of w. (Uniform model)

(a) $w = 20$

(b) $w = 40$

(c) $w = 60$

(d) $w = 80$

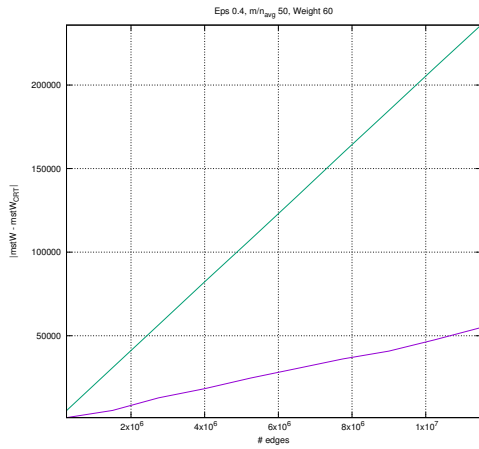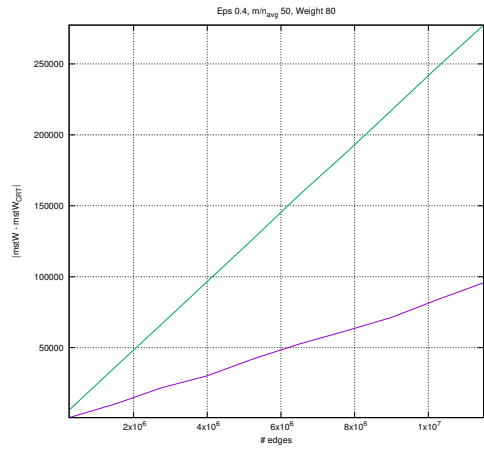Fig. 5.7: Error behaviour for the increase of w. (Uniform model)
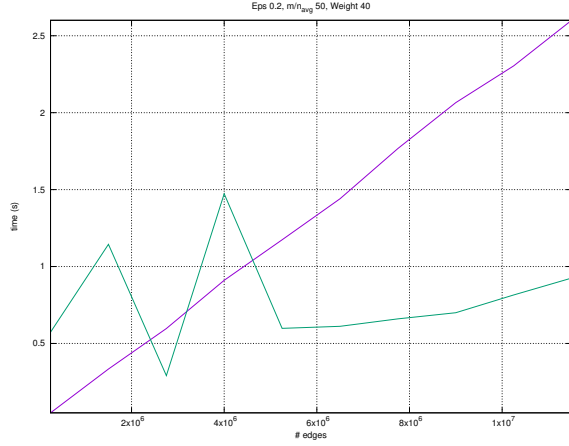
(a) $\varepsilon = 0.2$
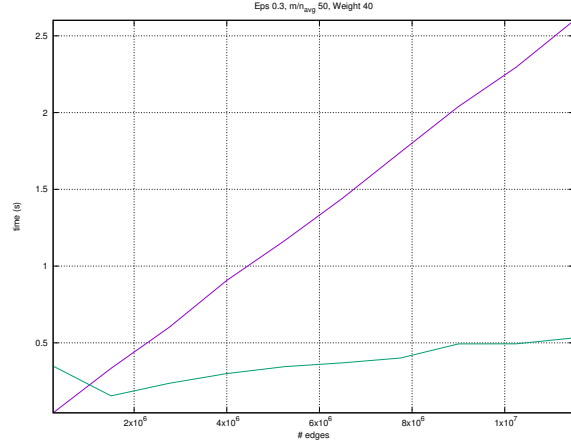
(b) $\varepsilon = 0.3$
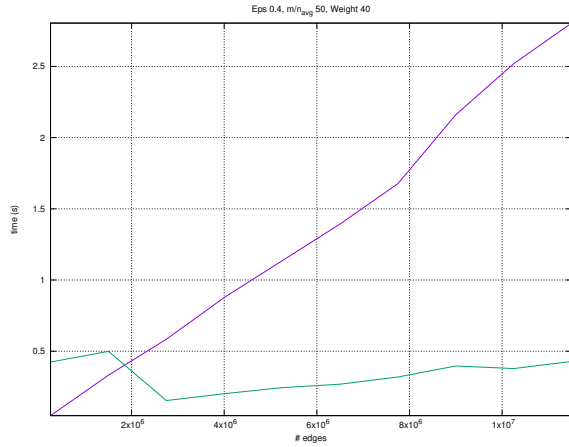
(c) $\varepsilon = 0.4$
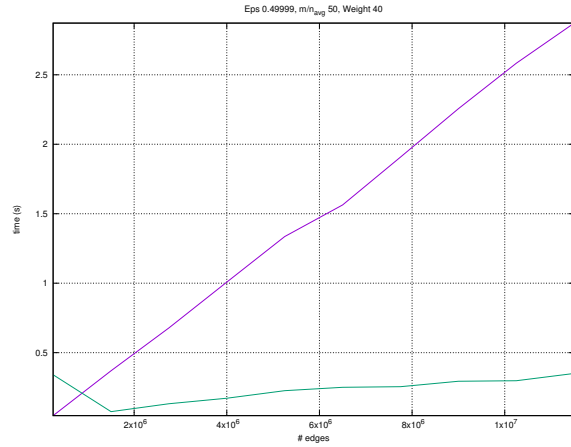
(d) $\varepsilon = 0.49999$

Fig. 5.8: Behaviour for the increase of ε. (Uniform model)

(a) $\varepsilon = 0.2$ (b) $\varepsilon = 0.3$ (c) $\varepsilon = 0.4$ (d) $\varepsilon = 0.49999$

(e) $\varepsilon = 0.2$ (f) $\varepsilon = 0.3$ (g) $\varepsilon = 0.4$ (h) $\varepsilon = 0.49999$

Fig. 5.9: Absolute and relative value of the error in a Uniform dataset, as ε changes; fixed values are $\bar{d} = 50$, $w_{\max} = 40$.

For the error trend instead we note an increase, still as expected; figures from 5.9a to 5.9d has to be read with attention because every function graph has a different scale. The reader must not confuse the apparent lowering of the error since it is not the case. Looking at the other graphs about the absolute error, from 5.9e to 5.9h, we see an expansion of the tolerance cone (light blue and ochre yellow lines) as increasing ε means admitting an higher deviation from the correct MST weight value. Here too the different scaling must not confuse about the increasing trend of the error.
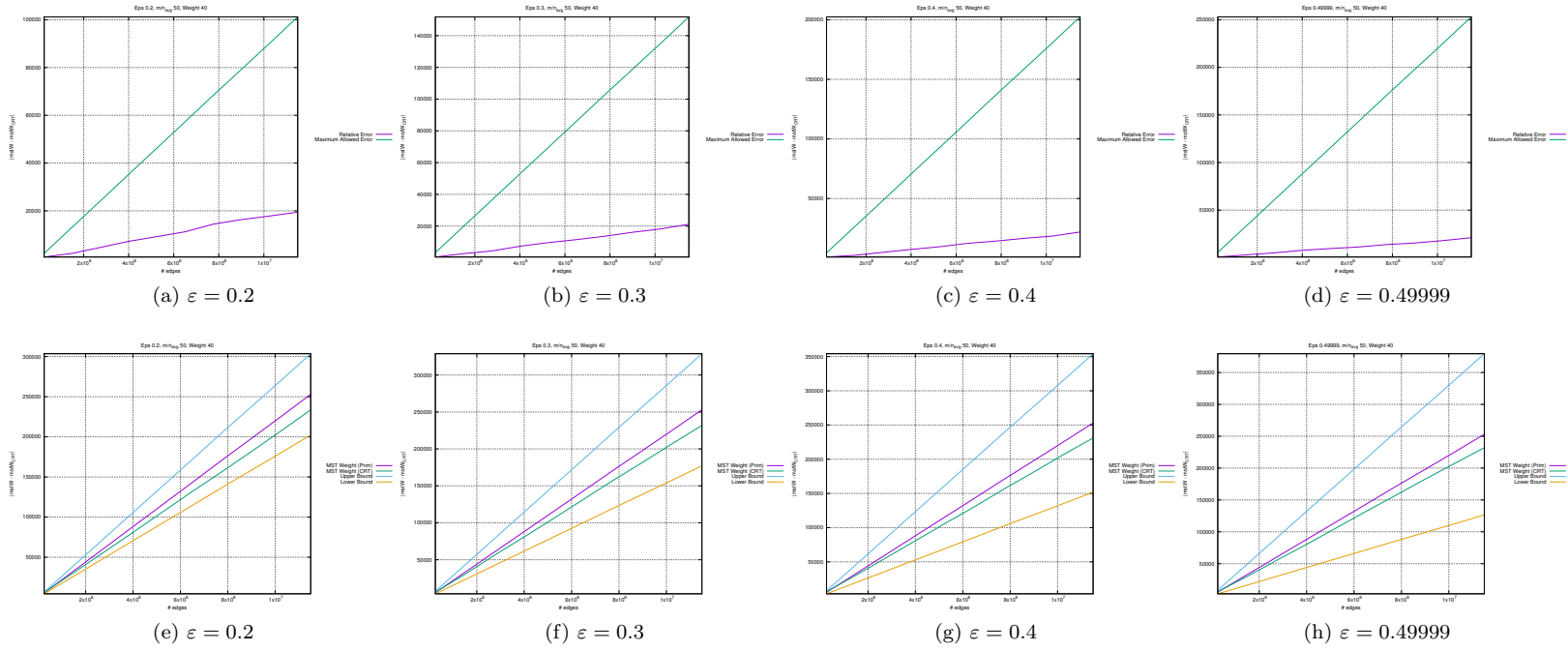
We see that the result is coherent with the theoretical model as the error increases with ε, but his variation is, after all, contained.

5.4. Variations of graph model. As a last comparison, instead of varying one by one the crucial parameters and see the time and error trends, we fix all of them and try to change the model the graph belongs to. Figures 5.10 on the next page show those results. We see here that both uniform and small-world models keep the same trend for the error, but the small-world one behaves slightly better on small instances. On the contrary, the gaussian show the same time trend respect to the uniform case, but his error has a higher growth curve. The scale-free model seems to be the worse case both regarding time and error trend, but we might remember, as observed earlier at 3.1 on page 5, that a real term of comparison requires to considerate a scale factor of 2 as done in figure 5.11 on page 20. We see in fact that, compared to the uniform case, the scale-free model has even a better behaviour, looking carefully at the function graph scale: but after all, both of them show sublinear complexity for instances of more than 10^7 edges, so have both the same transients as the same steady state trend.

Another thing we note is that a bad trend in execution times are always bond to an explosion of the error, as we can see in the charts so far. This means that using more time to compute the value doesn't mean it will be nearer to the correct one.

(a) uniform, time trend (b) gaussian, time trend (c) small world, time trend (d) scale free, time trend

(e) uniform, relative error (f) gaussian, relative error (g) small world, relative error (h) scale free, relative error
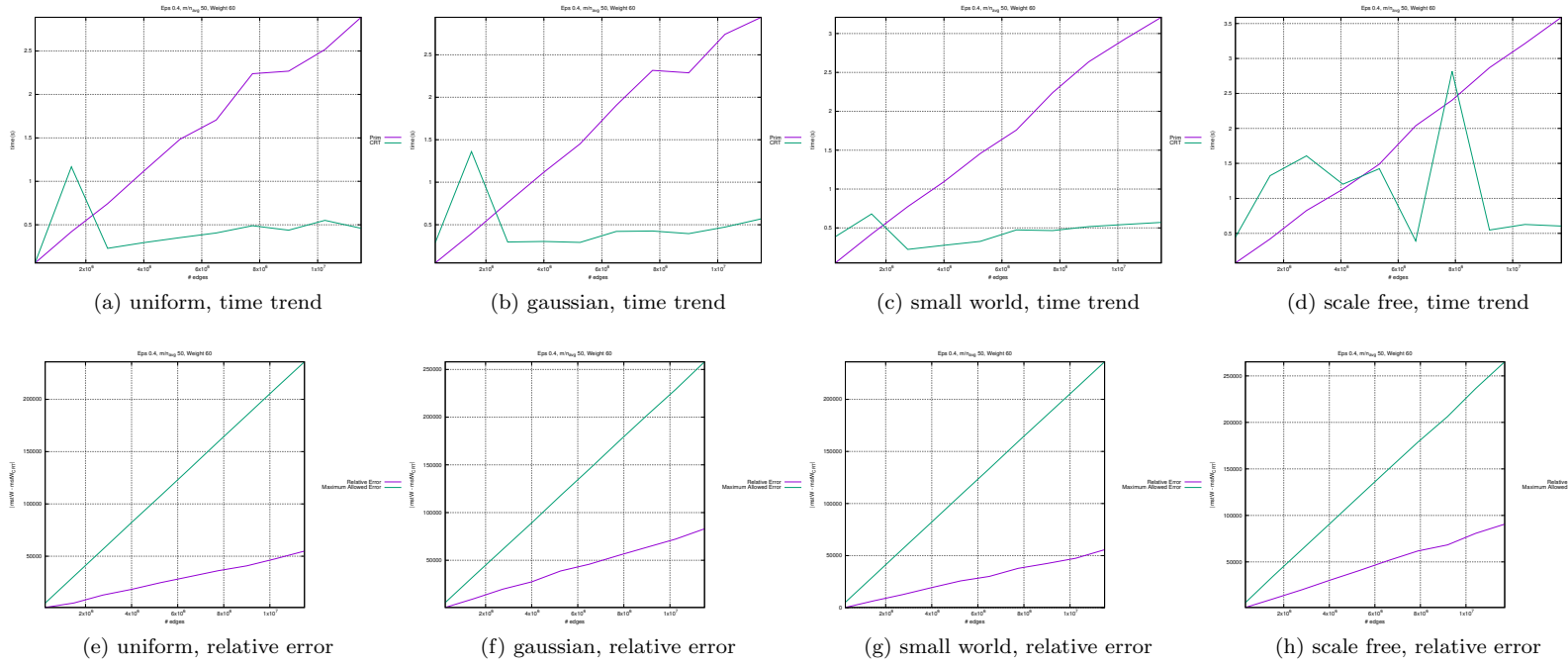
Fig. 5.10: Different behaviours for different graph models; fixed values are $\varepsilon = 0.4$, $\bar{d} = 50$, $w_{\max} = 60$.

(a) uniform, $\bar{d} = 100$ (b) scale free, $\bar{d} = 50$
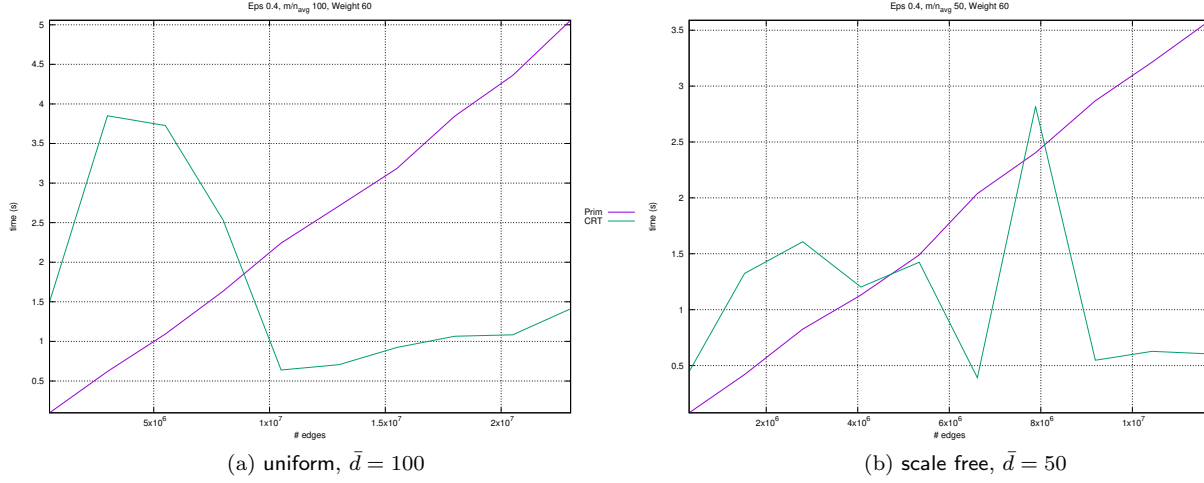
Fig. 5.11: A more correct comparison between the general case and scale-free graphs
($\varepsilon = 0.4$, $w_{\max} = 60$.)

6. A specific case of study. At this point, all of our graphs show a bad ini-
tial curve every time we burden the input instance; in a specific case of study this
anomalous curve was so persistent that all the function graph was more than linear;
we decided to deepen this special case of study, performing a longer run to see the
tail of this particular trend. The results are reported on figure 6.1 on the facing page.

We can see that the original case of study hinted a sublinear trend beyond 20
millions edges instances, but we considered to investigate further, and figure 6.1b
confirms the sublinearity. We have a good behaviour on the error trend (figure 6.1c).

7. Conclusions and final thoughts. As expected, a probabilistic algorithm
like the CRT allows us to compute an approximation of the Minimum Spanning Tree
weight in sublinear time on the number of edges, under certain conditions. Tunable
parameters, that depends on ε, allows us to perform either a better or a worse approx-
imation, implying respectively a very slow and a very fast computation. The choice of
a small value of ε can lead to terrible running times, and for these values it does not
make sense to compare the CRT algorithm with any other deterministic algorithm.
For other ε values, instead, we prove the good performances of the CRT. The reader
can easily view the better performances of CRT algorithm versus Prim algorithm or
Kruskal algorithm watching the line charts in the previous section of this paper.

More in general, we see that execution time and error depend on the number of
BFSes successfully completed during the computation The more BFSes are completed,
the more information the algorithm has to predict a correct value for the MST weight,
but on the other hand, completing the BFSes takes time. If instead we have a lot of
interrupted BFSes, we waste a large amount of time without gathering information,
hence resulting in both high execution time and error.

We considered so far different theoretical graph models, and we conclude that a
high *clustering coefficient* tends to increase the probability to have interrupted BFSes.
This because one of the reasons the algorithm has to interrupt the search is having
encountered a hub, i.e. a vertex with a high degree. We saw in fact that when

(a) Heavy case of uniform, $\varepsilon = 0.3$, $d \simeq 200$, $w = 80$

(b) Long run, up to 50 millions edges
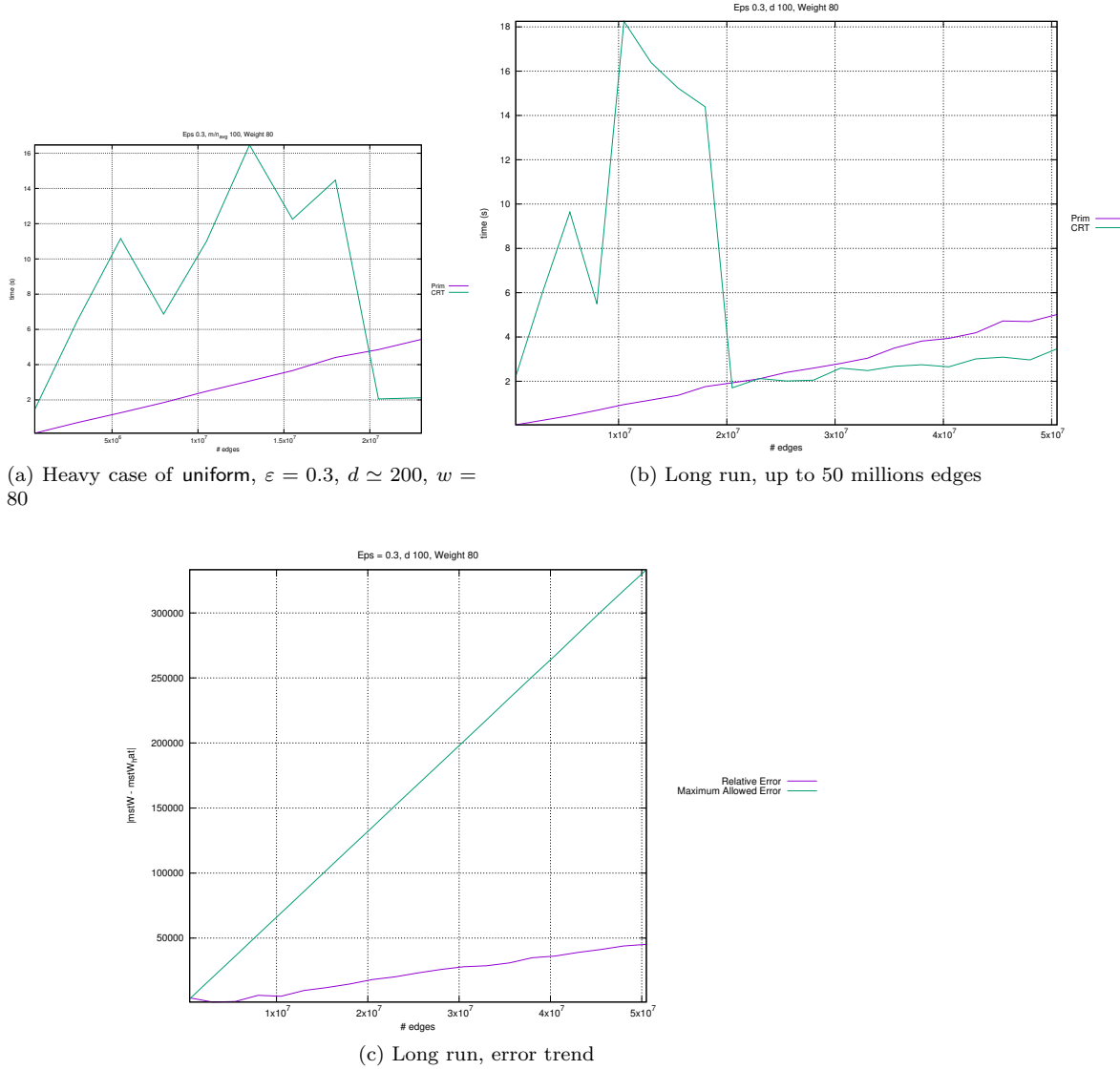

(c) Long run, error trend

Fig. 6.1: Long run results; figure (a) shows the original case of study, while (b) his extension.

changing the model there is a slight perturbation of the trend, although it remains sublinear. The key concept is the "distance" from the hubs of the graph of the root vertex from which our BFS starts.

Generally we also concluded that increasing the average degree let our algorithm gather more information, because there is a growing of the probability to visit the generic node u of our graph. On the other side, increasing the maximum weight correspond to an increase in the number of iterations our algorithm does, that leads to summing more intermediate approximated results that imply a higher final approxi-

mation.

7.1. Parallel implementation. We observe that the CRT algorithm lends itself very well to a parallel implementation. Indeed the majority of the algorithm's code is organized into independent sections, and in most cases they don't need to comunicate to each other. We also observe that three levels of paralellism can be achieved within the code. In the first level we parallelize each of the w independent calls to `approx-number-connected-components`, as depicted in pseudocode 1, below; every of this calls internally performs r independent BFSes from r different roots, that could in turn run in parallel, achieving a second level of parallelism. Moreover, considering that in the academic world there already exist different parallel implementations of the BFS algorithm, we can use one of them to perform an additional third level of paralellism.

Algorithm 1 First level parallelism for the CRT algorithm

---

    function APPROX-MST-WEIGHT(G, ε) $\triangleright$ G - input graph, ε - error tolerance
        $d^* \leftarrow$ APPROX-AVG-DEGREE(ε) $\triangleright$ sequential, runs in $O(d/\varepsilon)$ as shown in [1]
        $\hat{c} \leftarrow 0$
        for $i = 1, \ldots, i = w$ **do** $\triangleright$ this $w-1$ calls can be run in $w-1$ parallel threads
            $\hat{c} +=$ APPROX-NUMBER-CONNECTED-COMPONENTS$(G^{(i)}, \varepsilon, d^*)$
        end for
        return $\hat{v} \leftarrow n - w + \hat{c}$
    end function

To make it even more simpler, the number of different flows is known a priori, so a static pre-instantiation and a smart scheduling of the threads can be performed. At the first and second levels a *master-slave* model can be used in a *fork-join* structure, while in the third level a shared variable is needed between the different BFSes. As a final remark, parallelizing the BFSes could have too much overhead given that the algorithm is optimized to run them very fast and to stop the ones that seem to cost too much.

8. Future prospects. During an e-mail exchange with one of the original authors of [1], Dr. Ronitt Rubinfeld, another topic of discussion and study has emerged, about the distribution of the weight on the edges.

In our code, the generation of random graphs only assume a uniform distribution of the weights on the edges, i.e. a given edge can have an integer weight $k \in [1, w]$ with probability $\frac{1}{w}$. That implies a linear growth of the dimension of $E(G_i), \forall i \in [1, w]$, namely the set of G_i's edges; this is well depicted in figure 8.1a, where, as i grows, the size of $E(G_i)$ increases at each step of a quantity "near" $\frac{|E(G)|}{w}$, and it is more true as $|E(G)|$ is big for the law of large numbers. On the other side, having a generic law of distribution for the edges weight implies having a different behavior as depicted on 8.1b.

This difference could be of interest because it means that the input of different subsequent iterations of `approx-number-connected-components` will have a non regularly increasing size, or even the same size for some calls; it can be easily shown indeed that the function in 8.1 is nondecreasing. Since the cost of those calls finally determines the overall cost, we might argue that this could lead to a minor difference, but at the same time think that only with another set of tests we could conclude something relevant about this observation.
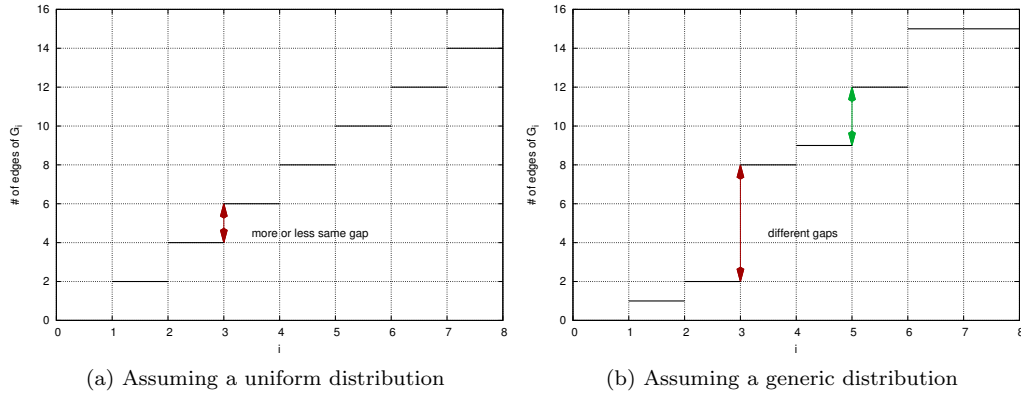
(a) Assuming a uniform distribution

(b) Assuming a generic distribution

Fig. 8.1: Expected behaviour with different laws for the distribution of weights.

REFERENCES

[1] B. CHAZELLE, R. RUBINFELD, AND L. TREVISAN, Approximating the minimum spanning tree weight in sublinear time. SIAM J *Computing*, 34, 2005.

[2] R. RUBINFELD, A. SHAPIRA, Sublinear Time Algorithms. SIAM *Journal on Discrete Mathematics*, 2011, Vol. 25, No. 4 : pp. 1562-1588.

[3] BEMAN DAWES, DAVID ABRAHAMS, RENE RIVERA, Boost C++ Libraries. http://www.boost.org/.

[4] FISHER, RONALD A.; YATES, FRANK (1948) [1938], Statistical tables for biological, agricultural and medical research (3rd ed.). London: Oliver & Boyd. *pp. 2627*.

[5] LENORE BLUM, MANUEL BLUM, MIKE SHUB; Comparison of Two Pseudo-Random Number Generators. Advances in Cryptology: *Proceedings of CRYPTO '82 pp. 61-78*, Plenum 1982.

[6] JOHN BLACK, PHILLIP ROGAWAY; Ciphers with Arbitrary Finite Domains. *Topics in Cryptology - CT-RSA 2002, The Cryptographer's Track at the RSA Conference, 2002, San Jose, CA, USA, February 18-22, 2002, Proceedings, pp. 114-130*

[7] DURSTENFELD, R. (JULY 1964), "Algorithm 235: Random permutation". *Communications of the ACM*.