Building Real-World Complex Networks by Wandering on Random Graphs

Bruno Gaume, Fabien Mathieu, Emmanuel Navarro

Building Real-World Complex Networks by Wandering on Random Graphs

Bruno Gaume[⋆‡], Fabien Mathieu[†], Emmanuel navarro[‡]

[⋆] CLLE-ERSS – CNRS, UTM, Toulouse, France
gaume@univ-tlse2.fr
[†] France Telecom – R & D, Issy-les-Moulineaux, France
fabien.mathieu@orange-ftgroup.com
[‡] IRIT – CNRS, UPS, INPT, Toulouse, France
navarro@irit.fr

Abstract

La plupart des graphes de terrain représentant des phénomènes du monde réel partagent des propriétés similaires de connectivité et de distribution des degrés, cependant, la génération artificielle de graphes possédant ces propriétés reste encore une question difficile.
Dans cet article, nous proposons d'utiliser des marches aléatoires sur des graphes aléatoires pour créer des graphes dont la connectivité et la distribution des degrés sont semblables aux graphes de terrain.

Mots-clés *: graphe, graphe de terrain, réseau petit monde, chaîne de Markov, graphe aléatoire*

Abstract

While most real-world graphs are known to share similar properties with respect to connectivity or degree distribution, generating artificial graphs with those properties is still a challenging issue.
In this paper, we propose to use random walks on random graphs to create graphs similar to real-world ones.

Key-words*: graph, real-world complex network, small world, Markov chain, random graph*

1 INTRODUCTION

In 1998, Watts and Strogatz showed that many real networks are characterized by a high clustering coefficient (their number of edges is sparse, yet thay contain a lot of triangles) and a short average path length, similar to the one of random graphs [38]. Another frequently encoutered characteristic is the

heavy-tail degree distribution: while most nodes have a degree close to the average degree, there are a few nodes of very high degree.

The fact that graphs built on real data from different domains share common properties has been confirmed by many studies [3, 27, 10, 1, 15, 20, 6, 32, 5, 31, 16, 4]. The concerned areas include, but are not limited to: epidemiology (contact graphs), economy (exchange graphs), sociology (knowledge graphs), linguistic (lexical networks), psychology (semantic association graphs), biology (neural networks, proteinic interactions graphs), IT (Internet, Web). We call such graphs *Real-World Complex Networks* (RWCNs).

So RWCNs are (1) globally sparse but (2) locally dense, with (3) a short average path length (APL), and (4) a heavy-tailed degree distribution. The combination of these four properties is very unlikely in random graphs, explaining the interest that those networks have raised in various scientific communities.

In this article, we propose a method to artificialy generate RWCN-like graphs. This method, which is based on random walks on random graphs, may help to better understand how RWCNs from various origins can share similar structures.

In Section 2, we formally introduce the characteristics of RWCNs. Section 3 surveys the different existing methods to generate complex networks. We analyze the dynamics of random walks in a graph in Section 4. In Section 5, we introduce a first version of our algorithm. In Section 6 we propose some improvements to reduce the length of the used random walks and therefore to speed up our method. In Section 7 we focus on the importance of the size of the initial random graphs on the properties of the final graphs. Lastly, section 8 concludes.

2 PROPERTIES OF REAL-WORLD COMPLEX NETWORKS

Let $G = (V, E)$ be a reflexive, symmetric graph:
V is the set of vertices, and $E \subset V \times V$ is the set of edges;
$n = |V|$ is the order of G (its number of nodes);
$m = |E|$ its size (its number of edges, with multiplicity);
$deg(u) = |\{v \in V/(u, v) \in E\}|$ is the degree of the node u;
$d = \frac{m}{n}$ is the average degree;

The four main properties of RWCNs are the following:

Edge sparsity Most of known RWCNs are sparse in edges, and the average degree stays low, it does not grow more than logarithmically with n: $m = O(n \log(n))$.

Short paths The APL, denoted L, is close to the APL L_{rand} in the main connected component of a random reflexive symmetric Erdős-Rényi

graph $\mathcal{G}(n,p)$ with same order and expected size (each symmetric edge between two distinct nodes exist with probability p; for the random graph to have the same expected size, we need to choose $p = \frac{m-n}{n(n-1)}$). According to [14], for $p \geq (1+\epsilon)\frac{\log(n)}{n}$, $\mathcal{G}(n,p)$ is almost surely connected, and $L_{\text{rand}} \approx \frac{\log(n)}{\log(m)-\log(n)}$ ($L_{\text{rand}} = O(\log(n))$).

High clustering The clustering coefficient, denoted C, that expresses the probability that two distinct nodes adjacent to a given third node are adjacent, is an order of magnitude higher than for Erdős-Rényi graphs: $C >> C_{\text{rand}} = p = \frac{m-n}{n(n-1)}$. This indicates that the graph is locally dense (there are a lot of triangles), although it is globally sparse (in terms of edges).

Heavy-tailed degree distribution Most RWCNs are heavy-tailed, having a few nodes with a very high degree. One frequently proposed model for such distribution is the power law, the probability for a given node to have degree k being proportional to $k^{-\lambda}$ for some constant λ.

Example:

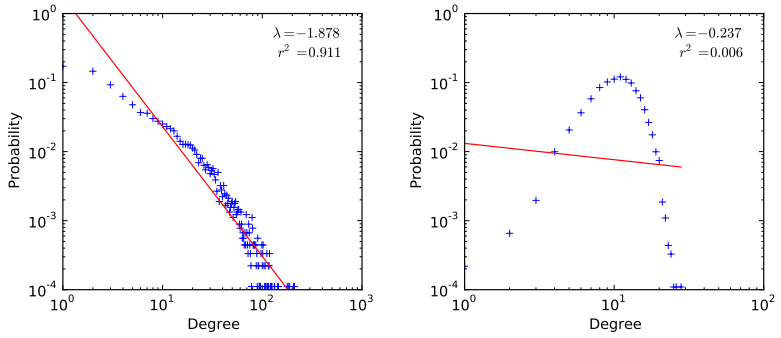
To illustrate how those properties appear in a RWCN, we propose to consider the graph DicoSyn.Verb[4] It is a reflexive symmetric graph with 9147 nodes and 111993 edges. For the sake of convenience, we only consider the largest connected component G_c of DicoSyn.Verb, which admits 8993 nodes and 111659 edges. With an average degree of 12.4, G_c is sparse. Other parameters of G_c are $L \approx 4.19$ (to be compared with $L_{\text{rand}} = 3.71$) and $C \approx 0.14$ (to compare with $C_{\text{rand}} = p = 0.0013$). The degree distribution is plot on log-log scale in Figure 1(a), it is clearly an heavy-tailed distribution.

Note, that the degree distribution for random Erdős-Rényi graphs is far from being heavy-tailed. It is in fact a kind of Poisson distribution : the probability that a node of a $\mathcal{G}(n,p)$ graph has degree k is $p(k) = p^k(1-p)^{n-1-k}\binom{n-1}{k}$. Figure 1(b) give the degree distribution of an Erdős-Rényi graph with same number of nodes and average degree than G_c.

To sum up, compared to Erdős-Rényi graph, RWCNs have the same sparsity (by construction), a similar short characteristic path lengths, but a higher clustering, and a heavy-tailed degree distribution (instead of Poisson distribution).

---

[4]DicoSyn is a french synonyms dictionnary built from seven canonical french dictionnaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse et Robert). The ATILF (`http://www.atilf.fr/`) extracted the synonyms, and the CRISCO (`http://elsap1.unicaen.fr/`) consolidated the results. DicoSyn.Verb is the subgraph induced by the verbs of Dicosyn: an edge exists between two verbs a and b iff DicoSyn tells a and b are synonyms. Therefore DicoSyn.Verb is a symmetric graph, made reflexive for convenience. A visual representation based on random walks [18] can be consulted on `http://Prox.irit.fr`.



(a) DicoSyn.Verb: Verbs synonymy graph (b) Erdős-Rényi graph of the same size

Figure 1: Degree distribution of a typical real world complex network (a) and of a random Erdős-Rényi graph having same number of veritces and edges (b). Plots are on a log-log scale. The red curves indicate the power-law model (ie. linear model on log-log scale) found by least-square fitting. λ give the slope of this curve, r^2 is the correlation coefficient of this model.

3 Modeling Real-world complex networks: State of Art

Since the paper of Watts and Strogatz [38], RWCNs have been studied intensely. In particular, a lot of work has been done in order to be able to generate artificial networks having RWCN's characteristics.

3.1 Small-world networks

By analogy with the *small-world phenomenon*[5] [29], Watts and Strogatz called "*small-worlds*" networks, networks having both a high clustering coefficient and a short characteristic path lengths [38]. For modeling such small-worlds networks, Watts and Strogatz alter a regular ring lattice by randomly rewiring some links. Another model was proposed by Kleinberg [25]: a d-dimensional grid is extended by adding extra-links which range follows a d-harmonic distribution.

Note, that both models fail to capture the heavy-tail property met in RWCNs (they are almost regular).

3.2 Heavy-tail property

There is a lot of research devoted to the production of random graphs that follow a given degree distribution [8, 28, 30, 33]. Such generic models easily

[5]This phenomenon is popularly known as *six degrees of separation* [21]

produce heavy-tailed random graphs if we give them a power-law distribution.

As for the models specifically designed to produce heavy-tailed distributions, Barabasi and Albert proposed the preferential attachment model [6], where new nodes are added one by one, and where the probability that an existing node receives a new link from the new node is proportional to the degree of the existing nodes. A more flexible version of the preferential attachment's model is the fitness model [1, 7], where a pre-determined fitness value is used in the process of link creation. Lastly, Aiello *et al.* proposed a model called α, β *graphs* [2], that encompasses the class of power law graphs.

Note, that these models fail to capture the high clustering property met in RWCNs.

3.3 Others models

Many variants of the model from Barabasi and Albert (BA model) have been proposed in order to provide a high clustering as well. In [23, 12, 24, 17], explicit phases of triangles construction are suggested. In [37], at each step of the graph construction, an edge is selected at random and a new vertex is added and connected to both sides of that edge. In [36], a small clique is created at each iteration instead of a single vertex. In [26, 11], vertices are divided into different potential clusters, and the edge creation processes inside or between clusters are handled separately.

Guillaume and Latapy proposed a different approach based on bipartite graphs [22]. The basic idea is that the unipartite projection of a random bipartite graph is a natural candidate for all RWCNs properties but the degree distribution. The power law distribution can be enforced for one class of vertices in the random bipartite graph, for instance by adapting the BA model.

Compared to the approaches previously proposed, the specificity of our solution is that we start from a fully grown random graph, which we turn into a graph having desired properties. It differs from BA variants, which construct a new graph from zero, and it can be viewed as the dual of Watts and Strogatz's small-worlds model: instead of adding random links to a regular structure, we propose to add "regular" links (i.e. local, with some kind of preferential attachment) to a random structure.

4 CONFLUENCE & RANDOM WALK IN NETWORKS

4.1 Random Walk in Networks

We assume that a particle wanders randomly on the graph G if:

- At any time $t \in \mathbb{N}$ the particle is on a node $u(t) \in V$;

- At time $t+1$, the particle reaches a uniformly randomly selected neighbor of $u(t)$.

This process is a *simple random walk* (SRW) on G [9] which can be defined by a *Markov chain* on V with the $n \times n$ *transition probability matrix* $[G]$ defined as follow:

$$[G] = (g_{u,v})_{u,v \in V}, \text{ with } g_{u,v} = \begin{cases} \dfrac{1}{\deg(u)} \text{ if } (u,v) \in E, \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

As G is reflexive no node has null degree, so $[G]$ is well defined. Moreover, it is a stochastic matrix by construction: $\forall u \in V, \sum_{v \in V} g_{u,v} = 1$.

For any initial probability distribution P_0 on V and any given integer t, $P_0[G]^t$ is the result of the random walk of length t starting from P_0 whose transitions are defined by $[G]$. As a special case, for any u, v in V, the probability P_t of being in v after a random walk of length t starting from u is equal to $(\delta_u[G]^t)_v = ([G]^t)_{u,v}$, where δ_u is the certitude of being in u. Using the Perron-Frobenius theorem [35], it can be shown that if $G = (V, E)$ is a connected, reflexive and symmetric graph, then:

$$\forall u, v \in V, \lim_{t \to \infty} (\delta_u[G]^t)_v = \lim_{t \to \infty} ([G]^t)_{u,v} = \frac{\deg(v)}{\sum_{x \in V} \deg(x)} \quad (2)$$

In other words, as t goes to infinity, the probability of being on node v at time t no longer depends on the departure node u, and is simply proportional to the degree of v.

4.2 Confluence in Networks

Equation (2) tells that the only information retained after an infinite random walk is the degree of the nodes. However, some information can be extracted from transitional states. Indeed, the dynamics of the particle's trajectory on its random walk is completely determined by the graph's topological structure: after t steps, every node v at a distance of t edges or less[6] from the initial vertex u can be reached. Furthemore when t remains small, the probability of reaching a vertex at the t^{th} step depends on the number of paths between the initial vertex u and the vertex v, on their length and on the degree of nodes along these paths: the more paths there are, the shorter the paths, and the weaker the degree of the intermediary nodes, then the probability of reaching v from the initial vertex u at the t^{th} step is higher[7]. For instance, assume the existence of three nodes u, v_1 and v_2 such that :

[6] Thanks to the reflexivity of the graph.

[7] Note that it is not only the length of the shortest path between u and v (ie. classical distance between graph vertices) whith is taken into account. This is an important point since this shortest path length is always short (cf. Section 2).

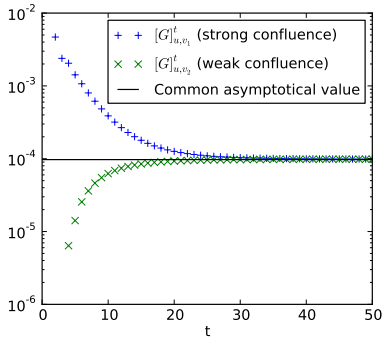
- u, v_1 and v_2 belong to the same connected component,
- v_1 and v_2 have the same degree,
- v_1 is *close* from u, in the sense that *many short paths* exist between u and v_1,
- v_2 is *distant* from u, in the sense that *few short paths exist* between u and v_2.

From Eq. (2), we know that the sequences $(([G]^t)_{u,v_1})_{1\leq t}$ and $(([G]^t)_{u,v_2})_{1\leq t}$ share the same limit $\deg(v_1)/\sum_{x\in V}\deg(x) = \deg(v_2)/\sum_{x\in V}\deg(x)$. However these two sequences are not identical: after a limited amount of steps t, one should expect a greater value for $([G]^t)_{u,v_1}$ than for $([G]^t)_{u,v_2}$ because v_1 is *closer* from u than v_2.

This can be illustraded on the synonymy graph of french verbs G_c (Graph introduced in Section 2), with :

- $u =$ *déshabiller* ("to undress");
- $v_1 =$ *effeuiller* ("to thin out");
- $v_2 =$ *rugir* ("to roar");

The nodes *effeuiller* and *rugir* have the same degree: $deg(effeuiller) = deg(rugir) = 11$, and intuitively, *effeuiller* should be closer (in G_c) to *déshabiller* than *rugir*, because this is the case semantically.



(a) French verbs graph G_c

Figure 2: $([G]^t)_{u,v_1}$ and $([G]^t)_{u,v_2}$ for G_c and a random graph

The values of $([G]^t)_{u,v_1}$ and $([G]^t)_{u,v_2}$ with respect to t are shown in Figure 2(a), along with the common asymptotic value $\frac{11}{\sum_{x\in V}\deg(x)}$. One can

observe that, after a few steps, $([G]^t)_{u,v_1}$ is above the asymptotic value. We claim that this is typical of nodes that are close to each other, and call this phenomenum *strong confluence*. On the other hand, $([G]^t)_{u,v_2}$ is always below the asymptotic value (*weak confluence*).

This phenomenum of strong and weak confluences is particulary clear in real graph thanks to their structure. Indeed it is quite easy to find some vertices connected by more and shorter path than others (typicaly vertices in a same "cluster" vs. vertices in different ones). However, strong and weak confluences also occur in Erdős-Rényi random graphs. Indeed such graphs are not completly uniform, they present an "embryo" of structure (at least, as graphs are sparces, some vertices are neighbors some are not). This can be illustraded by the Figure **??**, it shows $([\mathcal{G}]^t)_{u,v_1}$ and $([\mathcal{G}]^t)_{u,v_2}$ for three nodes u, v_1 and v_2 carefully selected in $\mathcal{G}$ an Erdős-Rényi graph with same number of nodes and average degree than G_c, there is clearly a strong confluence between u and v_1 and a weak confluence between u and v_2.

In the following Section, we will use this to detect and amplify this "embryo" of structure present in random graph in order to turn it into graphs having properties of RWCNs.

5 FROM RANDOM GRAPHS TO *shaped-like* REAL-WORLD COMPLEX NETWORKS

To generate graphs of small-world type, Watts and Strogatz [38] add random links in a locally linked graph. We propose here a dual approach by adding local links in a random graph. In order to provide a way for measuring locality of a possibly added link, we introduce the mutual confluence *conf* between two nodes of a graph G at a time t:

$$\mathrm{conf}_G(u, v, t) = \max([G]^t_{u,v}, [G]^t_{v,u}) \tag{3}$$

For *not too large* values of t, a strong mutual confluence between two nodes may indicate a local link for adding. We claim that a good way to obtain a *shaped-like* RWCNs from a random graph is to set links between the pairs of nodes with the highest confluence.

5.1 Extracting the confluence graph

Given an input graph $G_{in} = (V, E_{in})$, symmetric and reflexive, with n nodes and m_{in} edges, a time parameter t and a target number of edges m, one can extract a strong confluence graph $G = \mathrm{scg}(G_{in}, t, m)$ such that:

- G a symmetric, reflexive graph with the same nodes than G_{in} and with m edges,
- $\forall r \neq s, u \neq v \in V$, if $(r, s) \in E$ and $(u, v) \notin E$, then $\mathrm{conf}_{G_{in}}(r, s, t) \geq \mathrm{conf}_{G_{in}}(u, v, t)$.

Algorithm 1: scg (strong confluence graph), extract highest confluences

Input: An undirected graph $G_{in} = (V, E_{in})$, with n nodes and m_{in} edges
A walk length $t \in \mathbb{N}^*$
A target number of edges $m \in [n, n^2]$
Output: A graph $G = (V, E)$, with n nodes and m edges
begin
  $E \longleftarrow \emptyset$
  for $i \leftarrow 1$ **to** n **do**
    $E \longleftarrow E \cup \{(i, i)\}$ /* Make G reflexive */
  end
  while $|E| < m$ **do** /* Is there unset edges? */
(a) $(r, s) \longleftarrow \arg\max_{(u,v) \notin E}([G_{in}]^t_{u,v})$
(b) $E \longleftarrow E \cup \{(r, s)\}$
(c) $E \longleftarrow E \cup \{(s, r)\}$ /* Stay symmetric */
  end
end

Algorithm 1 proposes a way to construct $\mathrm{scg}(G, t, m)$. Note, that because of possible confluences with same values, line (a) is not deterministic. Furthermore, there is no guarantee that the strong confluence graph is unique, but the possible graphs can only differ by their (few) edges of lowest confluence. In practice, confluences are distinct most of the time [8]

5.2 Making *shaped-like* real-world complex networks

We propose to construct graphs with the properties of RWCNs by extracting the confluences of Erdős-Rényi graphs, as described in Algorithm 2. Note, that the confluence extraction may produce disconnected graphs. Therefore we have to select the main connected component if we want to study properties like the average path length. However, our experiments show that the size of the main connected component is always more than 80%, which appears to be a fair proportion.

5.3 Focus on the parameter t

In order to obtain a good graph, with the properties of RWCNs, the values of m_{in} and t must be carefully selected (for a given n and m). In the following, we set $n = 1000$, $m_{in} = 4000$, and $m = 10000$, and we focus on the importance of the parameter t.

[8] If uniqueness really matters, it suffices to use a total order on the pairs of V in order to break ties in line (a).

Algorithm 2: makesl, Making a shaped-like real-world complex networks

Input: A target number of nodes for the output graph $n \in \mathbb{N}$
A target number of edges for the random graph $m_{in} \in \mathbb{N}$
A walk length $t \in \mathbb{N}^*$
A target number of edges $m \in \mathbb{N}$
Output: A graph $G = (V, E)$, with n nodes and m edges
begin
  $G_{in} \longleftarrow$ a symmetric, reflexive, Erdős-Rényi Random Graph with n nodes and m_{in} edges
  $G \longleftarrow \mathrm{scg}(G_{in}, t, m)$
  $G \longleftarrow$ largest connected component of G
end

Like stated in Section 2, there is no strict definition of RWCNs properties, but typical values of average path length, clustering coefficient and degree distribution. We arbitrary propose to say that $G = \mathrm{makesl}(n, m_{in}, t, m)$ is shaped like RWCNs if it satisfies:

- $m \leq 10n\log(n)$ (satisfied for $n = 1000$, $m = 10000$),
- Clustering coefficient C_G is greater than $\frac{10m}{n^2}$,
- Average path length is shorter than $3\log(n)$,
- A least square fitting on the degree log-log distribution gives a negative slope of absolute value λ greater than 1, with a correlation coefficient r^2 grater than 0.8.

These constraints are certainly too strong (for instance one could be more flexible with the correlation coefficient r^2), but they guarantee that a graph within these constraints has RWCN-like properties.

Remark

The power law estimation we give is not very accurate (see for instance [34]) However, giving a correct estimation of the odds that a given discrete distribution is heavy-tailed is a difficult issue ([19, 13]), and refining the power-law estimation is beyond the scope of this paper.

It is easy to verify that with those requirements, a random Erdős-Rényi graph with 1000 nodes and 10000 edges is not shaped like RWCNs with high probability (for instance because of the clustering coefficient). On the other hand, $G = \mathrm{makesl}(n, m_{in}, t, m)$ satisfies the four properties of RWCNs for some values of t, as shown in Figure 3:

- The upper curve shows the number of nodes of the giant component of the result graph,

- The next curve shows the average path length L (remember that we only consider the main connected component, therefore the average path length is always well defined). The average path length L is always low and consistent with a RWCNs structure.

- The next curves indicates the clustering coefficient C. For $2 \leq t \leq 40$, C is very high. It drops after 40, as the confluences converge to the nodes' degrees, meaning that most of the edges come from the highest degree nodes of the input graph. This leads to star-like structures, that explain the poor clustering coefficient.

- The two next curves indicates that the degree distribution may be a power-law, with a relatively high confidence, for $34 \leq t \leq 45$.

- Lastly, the lower curve summarizes the values of t that satisfy the shaped-like RWCNs requirements (mainly $34 \leq t \leq 45$).

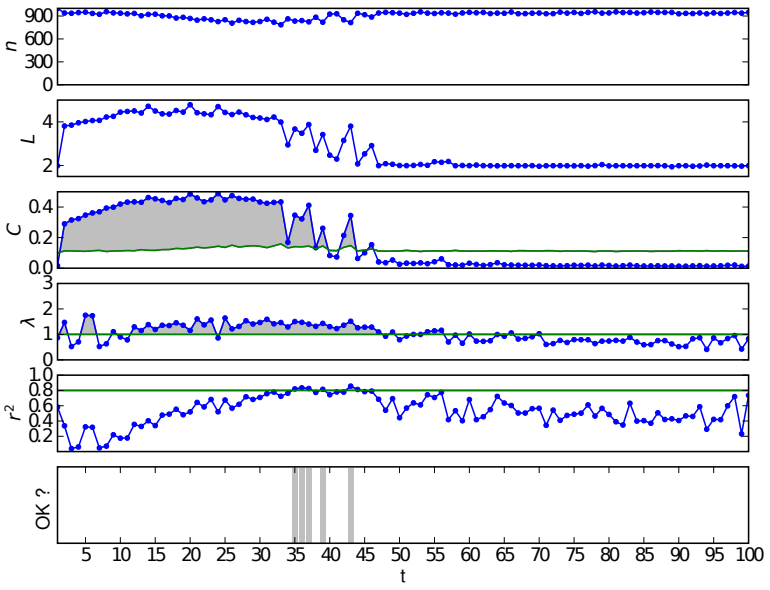


Figure 3: Properties of $G = \text{makesl}(n, m_{in}, t, m)$ with respect to t.

As shown Figure 3, all RWCNs properties are achieved quickly except the heavy-tailed degree distribution. We note that this distribution tends to fit a power-law just when confluences converge to the nodes' degrees ($t \approx 40$ in this example). After this point for all nodes u, v, $([G]^t)_{u,v}$ is proportional to $deg(v)$. Therefore, as already said, it produces a star-like structure where

each vertex is link to vertices of higher degree (in the initial graph)[9]. Nevertheless before this critical point, confluence is already strongly influenced by vertex degrees. That is certainly why heavy-tailed distribution appears here, by a phenomenon close to preferential attachment [6].

6 HEAVY-TAILED DISTRIBUTION WITH SMALLER WALKS

We present in this section two variants of the previous algorithm which significantly reduce the required walk length. Besides the computational cost[10], having walks of length greater than the average path length L is not completely satisfactory, as one would like the RWCN properties to emerge from "local" interactions, i.e. walks of very short length.

To reduce walk length, we propose to enforce some preferential attachment in edge selection phase (strong confluence extraction algorithm), and then to apply the strong confluence extraction algorithm iteratively. We show that the last method produces RWCN-like graphs after solely two iterations of two steps long random walks.

6.1 preferential attachment

In order to speed up the apparition of a heavy-tailed distribution, we propose to balance the edge selection with the degree of vertices in the strong confluence extraction algorithm. Therefore, edges are created preferentially between vertices having already a strong degree, like for the BA model [6]. As our algorithm consist in selecting edges instead of vertices, two potential vertices may be used in our preferential attachment: the source or the target.

If we consider the target only, the line (a) of Algorithm 1 should be replaced by:

$$(r, s) \longleftarrow \arg \max_{(u,v) \notin E} ([G_{in}]^t_{u,v} * \deg(v)) \tag{4}$$

The replacement for source weighting [11] should be:

$$(r, s) \longleftarrow \arg \max_{(u,v) \notin E} ([G_{in}]^t_{u,v} * \deg(u)) \tag{5}$$

Lastly, for taking both sides into account, we propose to replace line (a) by:

$$(r, s) \longleftarrow \arg \max_{(u,v) \notin E} ([G_{in}]^t_{u,v} * \deg(v) * \deg(u)) \tag{6}$$

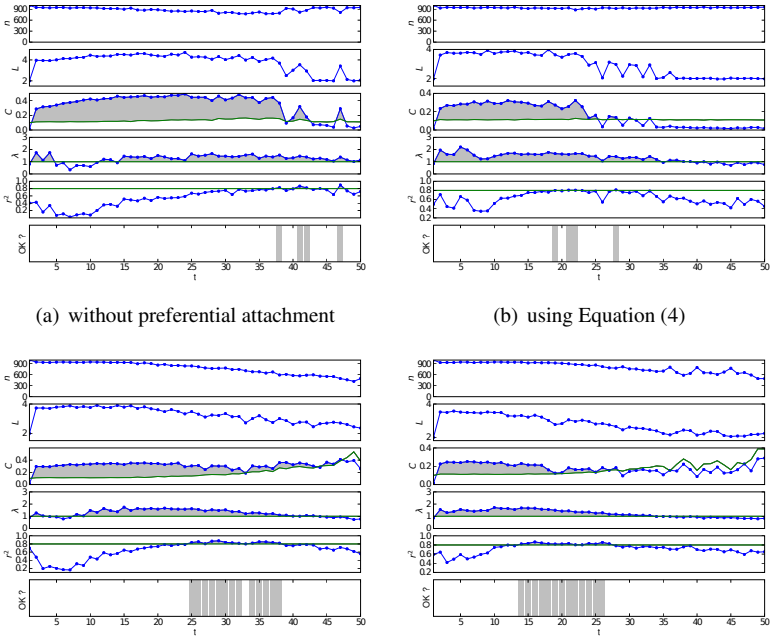
[9] In experiments a different random graph is use for each t, that explains fluctuations in the curves after this point.

[10] $[G]^t$ is no more sparse when t grows. For t greater than the diameter, all entries of $[G]^t$ are non null.

[11] Note that the matrix $[G_{in}]^t$ is not symmetric, hence these two replacements are not equivalent.

6.1.1 Results

We employ the same validation process than in Section 5.3, with $n = 1000$, $m_{in} = 4000$, and $m = 10000$. Figure 4 shows that preferential attachment is effective for speeding up the heavy-tailed distribution.



(a) without preferential attachment

(b) using Equation (4)

(c) using Equation (5)

(d) using Equation (6)

Figure 4: Properties of $G = \mathrm{makesl}(n, m_{in}, t, m)$ with respect to t.

The drawback is that the clustering coefficient is lowered. With the double-weight method, it decreases down to 0.2, that is about half the clustering observed without preferential attachment. However, it is still much higher than for an equivalent random graph, so the resulting graph can be considered as RWCN-like nevertheless. Intuitively, preferential attachment creates highly connected nodes, which can be seen as bridges between clusters. The price for these bridges is that they may have a low clustering themselves as a consequence. That can be an explanation of this trade-off between heavy-tailed distribution and strong clustering.

6.2 Iterative algorithm

Until now, the confluence was only computed once, in the input random graph. We propose to iterate the process as shown in Algorithm 3, extracting

iteratively several stronger confluence graphs.

Algorithm 3: makesl_{iter}, Making a shaped-like real-world complex network iteratively

Input: A target number of nodes for the output graph $n \in \mathbb{N}$
A target number of edges for the random graph $m_{in} \in \mathbb{N}$
A walk length $t \in \mathbb{N}^*$
A number of iteration $k \in \mathbb{N}^*$
A target number of edges $m \in \mathbb{N}$
Output: A graph $G = (V, E)$, with n nodes and m edges
begin
  $G \longleftarrow$ Erdős-Rényi Random Graph (n nodes, m_{in} edges, symmetric, reflexive)
  for $i \leftarrow 1$ **to** k **do**
    $G \longleftarrow \mathrm{scg}(G, t, m)$
  end
  $G \longleftarrow$ largest connected component of G
end

The underlying idea is that even if a short length walk does not produce a truly RWCN-like graph, the output is somehow "closer" to a RWCN than the original input, and should be a more promising input itself.

6.2.1 Results

Figure 5 presents the results for preferential attachment as defined by Equation (6)[12] for several values of k and t. The other parameters are the same as for the other experiments (Erdős-Rényi initial input graph, $n = 1000$, $m_{in} = 4000$, $m = 10000$).

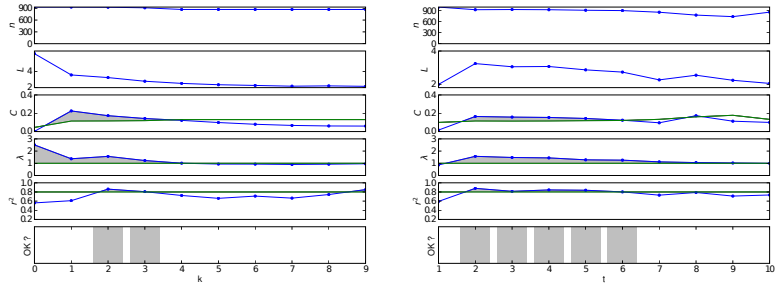
Under the proposed scenario, the iterative algorithm efficiently builds RWCN-like graphs while using short walks only. In fact, two iterations of two-steps walks seem to be enough, which is a great improvement.

7 FOCUS ON THE PARAMETER m_{in}

In this section we set $n = 9147$, $m = 111993$, (wich are respectively the number of nodes and edges of DicoSyn.Verb) with $t = 2$, $k = 2$, and we focus on the importance of the parameter m_{in}.

Figure 6 presents the results for $m_{in} \in [25000, 100000]$ by steps of 5000. When $m_{in} \in [40000, 70000]$, the output graphs is shaped-like RWCNs.

[12]The use of Equations (4) or (5) gives less significant performance.



(a) $G = \text{makesl}_{iter}(n, m_{in}, t, k, m)$ with respect to k, with $t = 2$.

(b) $G = \text{makesl}_{iter}(n, m_{in}, t, k, m)$ with respect to t, with $k = 2$.

Figure 5: Properties of graphs given by Algorithm 3

Table 1 gives the properties of graphs generated by the algorithm $makesl_{iter}$ with $n = 9147, m = 111993, m_{in} = 40000, t = 2, k = 2$, and the properties of DicoSyn.Verb. One can note that they are very similar.

	makesl_{iter}	DicoSyn.Verb
n	8615 (26.3)	8993
m	111407 (26.3)	111659
L	3.82 (0.02)	4.19
C	0.13 (0.00)	0.14
λ	-1.97 (0.03)	-1.88
r^2	0.88 (0.01)	0.91

Table 1: Properties of graphs generated by the algorithm $makesl_{iter}$ with $n = 9147, m = 111993, m_{in} = 40000, t = 2, k = 2$, compared to DicoSyn.Verb properties. $makesl_{iter}$ algorithm has been run 20 times, so for each property the given number corespond to the mean over this 20 graphs, the standard deviation is given in parenthesis.

8 Conclusion

We proposed in this paper to use algorithms based on random walks to turn random graphs into RWCN-like graphs. Our approach allows to get a graph with a given number of nodes and edges, having all properties expected from a RWCN: short average path length, low edge density, high clustering and heavy-tailed degree distribution.

However, being able to generate artificial RWCN-like graphs is not sufficient to answer one of the most interesting questions about RWCNs, which

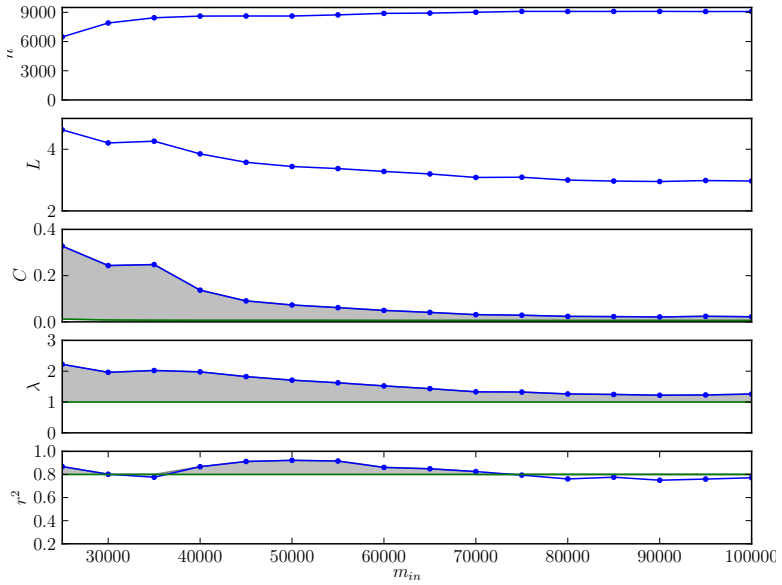


Figure 6: Properties of $G = \mathrm{makesl}_{iter}(n, m_{in}, t, k, m)$ with respect to m_{in}.

is *Why most of real-world complex network have a similar structure, despite the fact that this structure is very unlikely among possible graphs?* In order to bring a contribution to the answer to that question, a RWCN-like graph generator should emulate real-world interactions in its algorithms. As real-world interactions are based on local knowledge, the algorithm should be able to be decentralized, which is not the case for Algorithm 3.

However, there are simple variants of Algorithm 3 that can be decentralized: for instance, if we introduce a confluence bound s, an algorithm where each node u decide to connect with any node it can find with a mutual confluence greater than s would have the same behavior that Algorithm 3, except that the size m would not be directly tunable anymore. Understanding the relationship between m and s is part of our future work, which would more generally aim at providing a better analytical understanding of the reasons that explain why our solution succeeds in providing graphs shaped like real-world complex networks.

In this framework, we will have to analyticaly study the relationship between the properties of the output graphs and the inputs (t, k, n, m_{in} and s) of a decentralized algorithm.

REFERENCES

[1] Lada A. Adamic, Bernardo A. Huberman;, A. Barab'asi, R. Albert, H. Jeong et G. Bianconi;. Power-law distribution of the world wide web. *Science*, 287(5461):2115, March 2000.

[2] William Aiello, Fan Chung et Linyuan Lu. A random graph model for massive graphs. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, pages 171–180, Portland, Oregon, United States, 2000. ACM.

[3] R. Albert, H. Jeong et A. L. Barabasi. The diameter of the world wide web. *Nature*, 401:130–131, 1999.

[4] Reka Albert et Albert-Laszlo Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[5] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert et T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.

[6] Albert-László Barabási et Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[7] G. Bianconi et A. L. Barabási. Bose-einstein condensation in complex networks. *Phys Rev Lett*, 86(24):5632–5635, June 2001.

[8] B. Bollobás. *Random Graphs*. Cambridge University Press, 2001.

[9] Bela Bollobas. *Modern Graph Theory*. Springer-Verlag New York Inc., Octobre 2002.

[10] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins et Janet Wiener. Graph structure in the web. *Comput. Networks*, 33(1-6):309–320, 2000.

[11] Shouliang Bu, Bing-Hong Wang et Tao Zhou. Gaining scale-free and high clustering complex networks. *Physica A: Statistical Mechanics and its Applications*, 374(2):864–868, 2007.

[12] Michele Catanzaro, Guido Caldarelli et Luciano Pietronero. Assortative model for social networks. *Phys. Rev. E*, 70(3):037101, Sep 2004.

[13] Aaron Clauset, Cosma R. Shalizi et M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, Novembre 2009.

[14] P. Erdos et A. Rényi. On random graphs. *Publicationes Mathemticae*, 6(26):290–297, 1959.

[15] Michalis Faloutsos, Petros Faloutsos et Christos Faloutsos. On power-law relationships of the internet topology. In *Proceedings of SIGCOMM*, pages 251–262, Cambridge, Massachusetts, United States, 1999. ACM.

[16] Ramon Ferrer-i-Cancho et Ricard V. Sole. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, November 2001.

[17] Peihua Fu et Kun Liao. An evolving scale-free network with large clustering coefficient. In *Proceedings of ICARCV '06*, pages 1–4, Singapore, Décembre 2006.

[18] Bruno Gaume. Balades aléatoire dans les petits mondes lexicaux. *I3 Information Interaction Intelligence*, 4(2), 2004.

[19] Michel L. Goldstein, Steven A. Morris et Gary G. Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 41(2):255–258, Septembre 2004.

[20] Ramesh Govindan et Hongsuda Tangmunarunkit. Heuristics for internet map discovery. In *Procedings of IEEE INFOCOM 2000*, pages 1371–1380, Tel Aviv, Israel, Mars 2000. IEEE.

[21] John Guare. *Six Degrees of Separation: A Play*. Vintage, 1st vintage books ed edition, Novembre 1990.

[22] Jean-Loup Guillaume et Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical and Theoretical Physics*, 371(2):795–813, 2006.

[23] Petter Holme et Beom Jun Kim. Growing scale-free networks with tunable clustering. *Phys. Rev. E*, 65(2):026107, Janvier 2002.

[24] Liu Jian-Guo, Dang Yan-Zhong et Wang Zhong-Tuo. Multistage random growing small-world networks with power-law degree distribution. *Chinese Physics Letters*, 23(3):746, 2006.

[25] Jon Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, pages 163–170, Portland, Oregon, United States, 2000. ACM.

[26] Konstantin Klemm et Víctor M. Eguíluz. Growing scale-free networks with small-world behavior. *Phys. Rev. E*, 65(5):057102, May 2002.

[27] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins et Eli Upfal. The web as a graph. In *Proceedings of 19th ACM SIGACT-SIGMODAIGART Symp. Principles of Database Systems*, pages 1–10, Dallas, Texas, United States, 2000. ACM.

[28] T. Luczak. Sparse random graphs with a given degree sequence. *Random Graphs*, 2:165–182, 1992.

[29] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[30] M. Molloy et B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6(2-3):161–179, 1995.

[31] J. M. Montoya et R. V. Solé. Small world patterns in food webs. *Small World Patterns in Food Webs*, 214(3):405–412, Février 2002.

[32] M. E. J. Newman. Scientific collaboration networks: I. network construction and fundamental results. *Physical Review E*, 64:016131, 2001.

[33] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89:208701, 2002.

[34] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351, Septembre 2005.

[35] G. W. Stewart. Perron-frobenius theory: a new proof of the basics. Technical report, College Park, MD, USA, 1994.

[36] Jianwei Wang et Lili Rong. Evolving small-world networks based on the modified ba model. In *Proceedings of ICCSIT '08*, pages 143–146, Los Alamitos, CA, USA, 2008. IEEE.

[37] Lei Wang, Hua ping Dai et You xian Sun. Random pseudofractal networks with competition. *Physica A: Statistical Mechanics and its Applications*, 383(2):763–772, 2007.

[38] D.J. Watts et S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.