From Random Graph to Small World by Wandering

Bruno Gaume, Fabien Mathieu

INRIA

From Random Graph to Small World by Wandering

Bruno Gaume — Fabien Mathieu

N° 6489

Avril 2008

Thème COM

R apport de recherche

From Random Graph to Small World by Wandering

Bruno Gaume*, Fabien Mathieu†

Thème COM — Systèmes communicants
Projet GANG

Abstract: Numerous studies show that most known *real-world* complex networks share similar properties in their connectivity and degree distribution. They are called *small worlds*. This article gives a method to turn random graphs into Small World graphs by the dint of random walks.

Key-words: Random graphs, small worlds, random walks

* IRIT–UPS, Toulouse F-31062 Cedex 4, France
† Orange Labs, 38–40 rue du Général Leclerc, 92794 Issy-les-Moulineaux Cedex

Petit-mondisation par marches aléatoires

Résumé : De nombreuses études montrent un fait remarquable qui est que la plupart des réseaux dits de terrain possèdent des propriétés identiques bien particulières et font partie de la classe des graphes *petit-monde*. Un autre fait tout aussi remarquable est que cette classe des petits mondes est très petite au regard de l'ensemble des graphes possibles. Dans cet article, nous proposons une méthode de production de graphes petit-monde au moyen de marches aléatoires.

Mots-clés : Graphes aléatoires, petits mondes, marches aléatoires

1   Introduction

In 1998, Watts and Strogatz showed that many real graphs, coming from different fields, share similar properties [28]. This has been confirmed by many studies since this seminal work [4, 20, 9, 1, 13, 17, 6, 25, 5, 23, 14, 3]. The concerned fields 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, or small-world networks.

The common properties of real-world complex networks are a low diameter, a globally sparse but locally heavy edge density, and a heavy-tailed degree distribution. The combination of these property is very unlikely in random graphs, explaining the interest that those networks have arisen in different scientific communities.

In this article, we propose a method to generate a graph with small-world properties from random graph. This method, which is based on random walks, may be a first step in order to understand why graphs from various origins share a common structure.

In Section 2, we briefly state the properties used to decide wheter a given graph is small world or not. In Section 3, we survey the different existing methods to generate complex networks. In Section 4, we analyse the dynamics or random walks in a graph, and in Section 5 we propose a new method to construct small worlds by *wandering* on random graphs. Section 6 concludes.

2   Small Worlds Structure

let $G = (V, E)$ be a reflexive, symmetric graph with $n = |V|$ nodes and $m = |E|$ edges. G is called *small world* if the following properties are verified:

Edge sparsity Small world graphs are sparse in edges, and the average degree stay low: $m = O(n)$ or $m = O(n \log(n))$

Short paths The average path length (denoted ℓ) is close to the average path length ℓ_{rand} in the main connected component of $G(n, m) = \mathcal{G}(n, \frac{m-n}{n(n-1)})$ Erdös-Rényi graphs. According to [12], for $d := \frac{m-n}{n} \geq (1 + \epsilon) \log(n)$, $\mathcal{G}(n, \frac{m-n}{n(n-1)})$ is almost surely connected, and $\ell_{\mathrm{rand}} \approx \frac{\log(n)}{\log(d)}$. $(l = O(\log(n)))$.

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

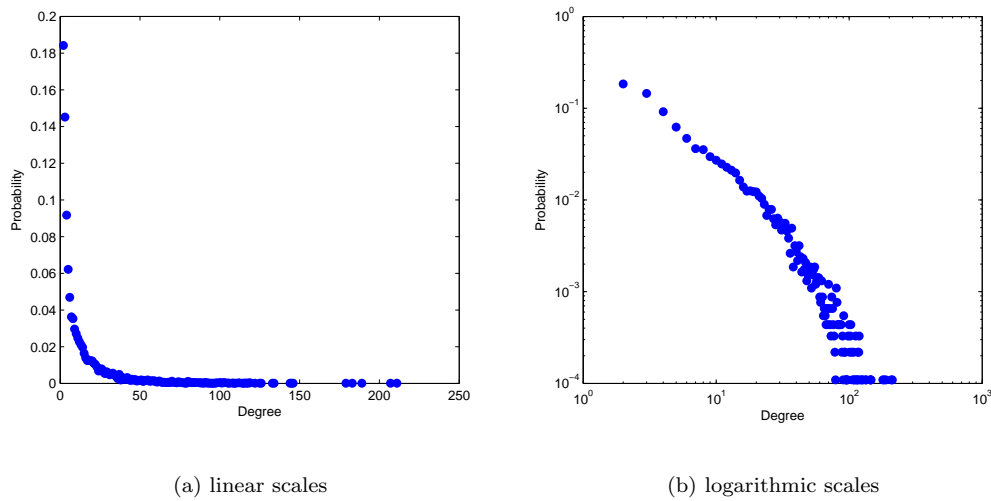**Heavy-tailed degree distribution**

Example: DicoSyn.Verbe[1] is a reflexive symmetric graph with 9043 nodes and 110939 edges. For sake of convenience, we only consider the main connected component G_c of DicoSyn, which admits 8835 nodes and 110533 edges. With an average degree of 12.5, G_c is sparse. Other parameters of G_c are $\ell \approx 4.17$ (to compare with $\ell_{\mathrm{rand}} = 3.71$) and $C \approx 0.39$ (to compare with $C_{\mathrm{rand}} = p = 0.0013$). The degree distribution is heavy-tailed, as shown by Figure 1 (a least-square method gives a slope of -2.01 with a confidence 0.96). Therefore G_c verifies the four properties of a small world.

Note, that the degree distribution for random Erdös-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 2, where the degree distribution of a Erdös-Rényi graph with same number of nodes and average degree than G_c is plotted. This illustrates how a small world compares to a $\mathcal{G}$ graph with same number of nodes and expected degree:

- Same sparsity (by construction),

- Similar average path length,

- Higher clustering,

---

[1]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.Verbe 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.verbe is a symmetric graph, made reflexive for convenience. A visual representation based on random walks [15] can be consulted on `http://Prox.irit.fr`.

(a) linear scales

(b) logarithmic scales

Figure 1: Degree distribution of G_c

- Heavy-tailed distribution (instead of Poisson distribution)

(a) linear scales

(b) logarithmic scales
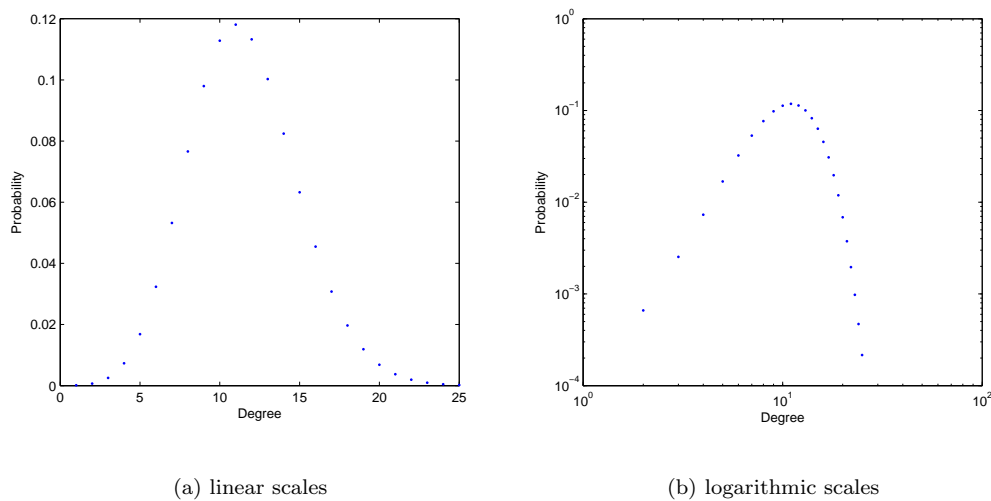
Figure 2: Degree distribution of a typical $\mathcal{G}(n, p)$ graph

In [3], Albert and Barabasi have made a survey on existing complex networks studies, including [4, 20, 9, 1, 13, 17, 6, 25, 5, 23, 14, 28]. Some of their findings are presented in Table 1 along G_c's properties.

3   Generating Small Worlds: State of Art

Small-world networks have been studied intensely since they were first described in Watts and Strogatz [28]. Researchs have been done in order to be able to generate random datasets with well-known characteristics shared by social networks. Most papers focus on either the clustering and diameter, or on the power-law.

3.1   Clustering and diameter property

Watts and Strogatz [28], and Kleinberg [19] have studied families of random graphs that share the clustering and diameter properties of small worlds. Watts and Strogatz model consist in altering a regular ring lattice by

Name	n	$<k>$	ℓ	C	γ	r_2
DicoSyn.Verbes	8835	11.51	4.17	0.39	2.01	0.96
Internet routers	150000	2.66	11		2.4	
Movie actors	212250	28.78	4.54	0.79	2.3	
Co-authorship, SPIRES	56627	173	4.0	0.726	1.2	
Co-authorship, math.	70975	3.9	9.5	0.59	2.5	
Co-authorship, neuro.	209293	11.5	6	0.76	2.1	
Ythan estuary food web	134	8.7	2.43	0.22	1.05	
Silwood Park food web	154	4.75	3.40	0.15	1.13	
Words, synonyms	22311	13.48	4.5	0.7	2.8	

Table 1: Main properties of some complex networks

rewiring randomly some links. In Kleinberg's model, a d-dimensional grid is extended by adding extra-links of which the range follows a d-harmonic distribution.

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

3.2 Heavy-tail property

There is a lot of research devoted on the production of random graphs that follow a given degree distribution [8, 21, 22, 26]. Such generic models easily produce heavy-tailed random graphs if we give them a power law distribution.

On the field of specific heavy-tailed models, there is Albert and Barabasi preferential attachment's model [3, 6], in which links are added one by one, and where the probability that an existing node receives a new link is proportional to its degree. 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.

3.3 Others models

Other models of graph generation are Guillaume and Latapy's *All Shortest Paths* [18], where one construct a graph by extracting the shortest paths of a random graph, and the Dorogovtsev-Mendes model [11]. Note, that the latter captures all desired properties, but is not realistic.

4 Confluence & Random Walk in Networks

4.1 Random Walk in Networks

Just like Section 2, $G = (V, E)$ is a reflexive, symmetric graph with $n = |V|$ nodes and $m = |E|$ edges. We assume that a particle wanders randomly on the graph:

- 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 an homogeneous Markov chain for on V. A classical way to represent this chain is a $n \times n$ stochastic matrix $[G]$:

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

Because G is reflexive, no node has null degree, so the underlying Markov chain $[G]$ is well defined. 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 $[G]$. More precisely, 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. One can demonstrate, by the dint of the Perron-Frobenius theorem [24], that if $G = (V, E)$ is a connected, reflexive and symmetric graph, then:

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

In other words, given than t is large enough, the probability of being on node v at time t is proportional to the degree of V, and no longer depends on the departure node u.

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. For instance, assume the existence of three nodes u, v_1 and v_2 such that

- u, v_1 and v_2 belong to the same connected component,

- 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,

- v_1 and v_2 have the same degree.

From (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, that is $\deg(v_1)/\sum_{x \in V} \deg(x) = \deg(v_2)/\sum_{x \in V} \deg(x)$.

However these two sequences are not identical. Starting from u, the dynamic of the particle's trajectory on its random walk is completely determined by the graph's topological structure, and 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 verified on the graph of french verbs $\mathcal{G}_c$, with:

- $u = déshabiller$ ("to undress"),

- $v_1 = effeuiller$ ("to thin out"),

- $v_2 = rugir$ ("to roar"),

Intuitively, *effeuiller* should be closer (in G_c) to *déshabiller* than *rugir*, because this is the case semantically. Also *effeuiller* and *rugir* have the same degree (11).

The values of $(([G]^t)_{u,v_1})$ and $(([G]^t)_{u,v_2})$ with respect to t are shown in Figure 3(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*).

One could think that the existence of strong and weak confluences is typical of graphs with high clustering, because the notion of *closeness* sounds like belonging to a same community. However, strong and weak confluences also occur in graphs with low clustering coefficients, such as Erdös-Rényi random graphs. For example, Figure 3(b) 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 Erdös-Rényi graph with same number of nodes and average degree than G_c.

Figure 3(b) is very similar to Figure 3(a). This points out that the concept of confluence exists in random graphs like it does in small worlds. In the following Section, we will use this to turn random graphs into small-worlds.

5   From Random Graph to Small World by Wandering

Now we want to use the concept of confluence to provide a way to construct small-world like graphs. In order to do that we introduce the mutual confluence conf between two nodes of a graph G at a time t:

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

For *not too large* values of t, a strong mutual confluence between two nodes may indicate that those nodes are close. We claim that a good way to obtain a small world from a random graph is to set edges between the pairs of nodes with the highest confluence.

(a) French verbs graph G_c
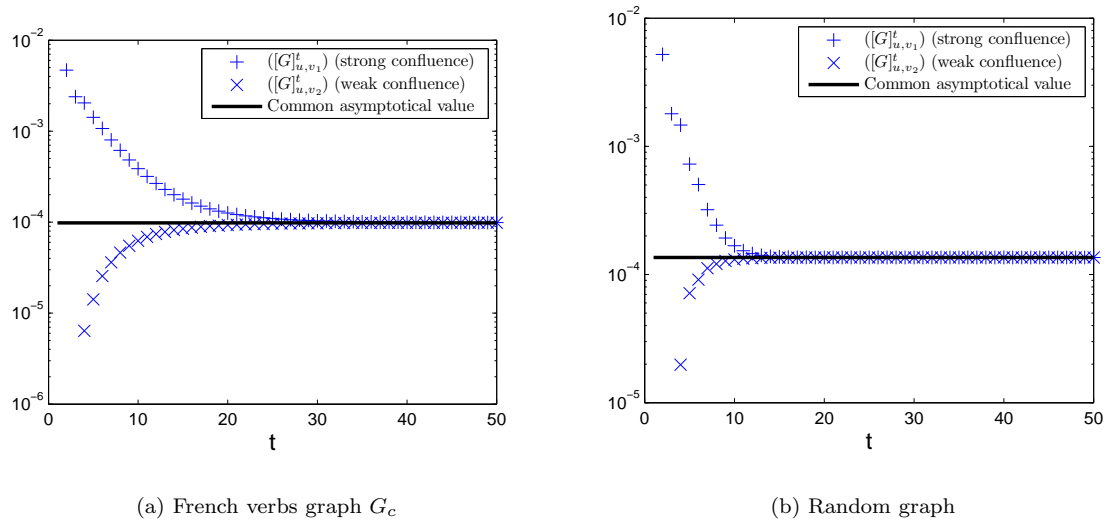
(b) Random graph

Figure 3: $(([G]^t)_{u,v_1})$ and $(([G]^t)_{u,v_1})$ for G_c and a random graph

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 $G = \text{scg}(G_{in}, t, m)$ defined by:

- G a symmetric, reflexive graph with the same nodes than G_{in} and m edges,

- $\forall r \neq s, u \neq v \in V$, if $(r, s) \in E$ and $(u, v) \notin E$, then $\text{conf}_{G_{in}}(r, s, t) \geq \text{conf}_{G_{in}}(u, v, t)$.

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
    $M \longleftarrow n$
    while $M < m$ **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 */
        $M \longleftarrow M + 2$
    end
end

Algorithm 1 proposes a way to construct $\text{scg}(G, t, m)$. 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[2]

---

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

5.2   Making Small-Worlds

Algorithm 2: makesw, Making a small world

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
$\quad$ $G_{in} \longleftarrow$ a symmetric, reflexive, Erdös-Rényi Random Graph with n nodes and m_{in} edges
$\quad$ $G \longleftarrow \text{scg}(G_{in}, t, m)$
$\quad$ $G \longleftarrow$ largest connected component of G
end

We propose to construct graphs with a small-world structure by extracting the confluences of Erdös-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 diameter. However, our experiments show that the size of the main connected component is always more than 80%, so this is not such a big issue.

5.3   Validation

In order to obtain good small-worlds, the values n, m_{in}, m and t must be carefully selected. In the following, we set $n = 1000$, $m_{in} = 4000$, and $m = 10000$, and we focus on the importance of the parameter t.

Like stated in Section 2, there is no strict definition of a small-world, but typical values for diameter, clustering and degree distribution. We arbitrary propose to say that $G = \text{makesw}(n, m_{in}, t, m)$ is small-world shaped if it verifies:

- $m \leq 10n \log(n)$ (verified for $n = 1000$, $m = 10000$),

- its clustering coefficient C_G is greater than $\frac{10m}{n^2}$,

- its diameter is lower 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.

Remark   The power law estimation we give is not very accurate (see for instance [27]). However, giving a correct estimation of the odds that a given discrete distribution is heavy-tailed is a difficult issue ([16, 10]), and refining the power-law estimation is beyond the scope of this paper.

It is is easy to verify that with those requirements, a random Erdös-Rényi graph with 1000 nodes and 10000 edges is not a small world with high probability (for instance because of the clustering coefficient). On the other hand, $G = \text{makesw}(n, m_{in}, t, m)$ verifies small-world properties for some values of t, as shown in Figure4:

- The upper curve shows the diameter L (remember that we only consider the main connected component, therefore the diameter is always well defined). The diameter is always low and consistent with a small-world 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 $28 \leq t \leq 50$.

- Lastly, the lower curve summarizes the values of t that verify the small-world requirements (mainly $28 \leq t \leq 40$).
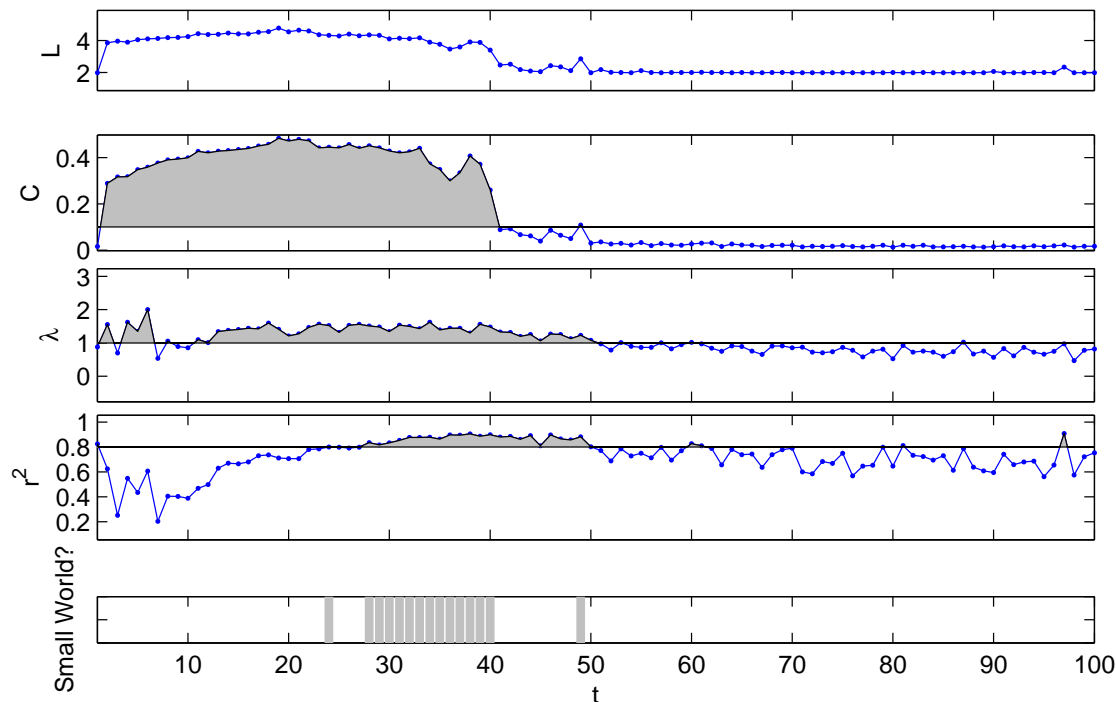
Figure 4: Small-world properties of $G = \mathrm{makesw}(n, m_{in}, t, m)$ with respect to t.

6 Conclusion

We proposed in this article a method to turn random graphs into Small-World graphs by the dint of random walks. This simple and intuitive method allow to set a target number of nodes and edges. The resulting graphs possess all desired properties: low diameter, low edge density with a high local clustering, and a heavy-tailed degree distribution. This method is suitable for generating random small-world graphs, but it is only a first step for answering the question: *why are most of real graphs small-worlds, despite the fact that the small-world structure is very unlikely among possible graphs?*

In order to be eligible for explaining small-world effects, a small-world generator should be based on local interactions. Therefore it should be decentralized, which is not the case of Algorithm 2. However, there exists variations of Algorithm 2 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 has the same behavior that Algorithm 2 (but the number of edges m is then indirectly set by the parameter s). Understanding the relationship between m and s is part of our future work.

Also note, that the random walks we used in this first algorithm may be too long: for instance, Figure 4 shows that a length between 28 and 40 is needed to achieve small-world properties for a 1000 nodes graph, which is much larger than the expected diameter of a small-world graphs of that size. We are currently working on a way to shorten the random walks by embedding a preferential attachment scheme [3] into our algorithm.

References

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

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180, New York, NY, USA, 2000. ACM.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

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

[5] A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations, April 2001.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.

[7] G. Bianconi and 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] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Comput. Networks*, 33(1-6):309–320, 2000.

[10] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, Jun 2007.

[11] S. Dorogovtsev and J. Mendes. Evolution of networks. *Advances in Physics*, 51 (4):1079–1187, 2002.

[12] P. Erdos and A. Rényi. On random graphs. *Publicationes Mathemticae (Debrecen)*, 6:290–297, 1959.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[14] R. Ferrer-i-Cancho and R. 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.

[15] B. Gaume. Balades aléatoire dans les petits mondes lexicaux. *Information Engineering Sciences*, 4(2), 2004.

[16] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution, August 2004.

[17] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *IEEE INFOCOM 2000*, pages 1371–1380, Tel Aviv, Israel, March 2000. IEEE.

[18] J.-L. Guillaume and M. Latapy. Complex network metrology. *Complex Systems*, 16:83–94, 2005.

[19] J. Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[20] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. The web as a graph. In *PODS*, pages 1–10, 2000.

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

[22] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, pages 161–179, 1995.

[23] J. M. Montoya and R. V. Solé. Small world patterns in food webs. *J. Theor. Biol.*, 214:405–412, February 2002.

[24] S. J. Ne. Perron–frobenius theory: A new proof of the basics, 1994.

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

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

[27] M. E. J. Newman. Power laws, pareto distributions and zipf's law, December 2004.

[28] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.