A dynamic model for on-line social networks

A. Bonato[1], N. Hadi[2], P. Horn[3], P. Prałat[4], and C. Wang[1]

[1] Ryerson University, Toronto, Canada
abonato@ryerson.ca, cpwang@ryerson.ca
[2] Wilfrid Laurier University, Waterloo, Canada
hadi4130@wlu.ca
[3] University of California, San Diego, U.S.A.
phorn@math.ucsd.edu
[4] Dalhousie University, Halifax, Canada
pralat@mathstat.dal.ca *

Abstract. We present a deterministic model for on-line social networks based on transitivity and local knowledge in social interactions. In the Iterated Local Transitivity (ILT) model, at each time-step and for every existing node x, a new node appears which joins to the closed neighbour set of x. The ILT model provably satisfies a number of both local and global properties that were observed in real-world on-line social and other complex networks, such as a densification power law, decreasing average distance, and higher clustering than in random graphs with the same average degree. Experimental studies of social networks demonstrate poor expansion properties as a consequence of the existence of communities with low number of inter-community links. A spectral gap for both the adjacency and normalized Laplacian matrices is proved for graphs arising from the ILT model, thereby simulating such bad expansion properties.

1 Introduction

On-line social networks such as Facebook, MySpace, and Flickr have become increasingly popular in recent years. In such networks, nodes represent people on-line, and edges correspond to a friendship relation between them. In these complex real-world networks with sometimes millions of nodes and edges, new nodes and edges dynamically appear over time. Parallel with their popularity among the general public is an increasing interest in the mathematical and general scientific community on the properties on-line social networks, in both gathering data and statistics about the networks, and in finding models simulating their evolution. Data about social interactions on-line networks is more readily accessible and measurable than in off-line social networks, which suggests a need for rigorous models capturing their evolutionary properties.

The small world property of social networks, introduced by Watts and Strogatz [29], is a central notion in the study of complex networks, and has roots in the work of Milgram [25] on short paths of friends connecting strangers. The small world property posits low average distance (or diameter) and high clustering, and has been observed in a wide variety of complex networks.

An increasing number of studies have focused on the small world and other complex network properties in on-line social networks. Adamic et al. [1] provided an early study an on-line social network at Stanford University, and found that the network has the small world property. Correlation between friendship and geographic location was found by Liben-Nowell et al. [24] using data from LiveJournal. Kumar et al. [21] studied the evolution of the on-line networks Flickr and Yahoo!360. They found (among other things) that the average distance between users actually decreases over time, and that these networks exhibit power-law degree distributions. Golder et al. [19] analyzed the Facebook network by studying the messaging pattern between friends with a sample of 4.2 million users. They also found a power law degree distribution and the small world property. Similar results were found in [2] which studied Cyworld, MySpace, and Orkut, and in [26] which examined data collected from four on-line social networks: Flickr, YouTube, LiveJournal, and Orkut. For further background on complex networks and their models, see the books [5, 7, 10, 13].

* The authors gratefully acknowledge support from NSERC and MITACS grants.

Recent work by Leskovec et al. [22] underscores the importance of two additional properties of complex networks above and beyond more traditionally studied phenomena such as the small world property. A graph G with e_t edges and n_t nodes satisfies a *densification power law* if there is a constant $a \in (1,2)$ such that e_t is proportional to n_t^a. In particular, the average degree grows to infinity with the order of the network (in contrast to say the preferential attachment model, which generates graphs with constant average degree). In [22], densification power laws were reported in several real-world networks such as a physics citation graph and the internet graph at the level of autonomous systems. Another striking property found in such networks (and also in on-line social networks; see [21]) is that distances in the networks (measured by either diameter or average distance) decreases with time. The usual models such as preferential attachment or copying models have logarithmically or sublogarithmically growing diameters and average distances with time. Various models (such as the Forest Fire [22] and Kronecker multiplication [23] models) were proposed simulating power law degree distribution, densification power laws, and decreasing distances.

We present a new model, called *Iterated Local Transitivity* (ILT), for on-line social and other complex networks which dynamically simulates many of their properties. Although modelling has been done extensively for other complex networks such as the web graph (see [5]), models of on-line social networks have only recently been introduced (such as those in [12, 21, 24]). The central idea behind the ILT model is what sociologists call *transitivity*: if u is a friend of v, and v is a friend of w, then u is a friend of w (see, for example, [16, 28, 30]). In its simplest form, transitivity gives rise to the notion of *cloning*, where u is joined to all of the neighbours of v. In the ILT model, given some initial graph as a starting point, nodes are repeatedly added over time which clone *each* node, so that the new nodes form an independent set. The ILT model not only incorporates transitivity, but uses only local knowledge in its evolution, in that a new node only joins to neighbours of an existing node. Local knowledge is an important feature of social and complex networks, where nodes have only limited influence on the network topology. We stress that our approach is mathematical rather than empirical; indeed, the ILT model (apart from its potential use by computer and social scientists as a simplified model for on-line social networks) should be of theoretical interest in its own right.

Variants of cloning were considered earlier in duplication models for protein-protein interactions [3, 4, 9, 27], and in copying models for the web graph [6, 20]. There are several differences between the duplication and copying models and the ILT model. For one, duplication models are difficult to analyze due to their rich dependence structure. While the ILT model displays a dependency structure, determinism makes it more amenable to analysis. The ILT model may be viewed as simplified snapshot of the duplication model, where *all* nodes are cloned in a given time-step, rather than duplicating nodes one-by-one over time. Cloning all nodes at each time-step as in the ILT model leads to densification and high clustering, along with bad expansion properties (as we describe in the next paragraph).

We prove that the model exhibits a densification power law with exponent $a = \frac{\log 3}{\log 2}$; see Theorem 2. We study the average distances and clustering coefficient of the model as time tends to infinity. In particular, we show that the average distance of the model of time t converges to a function dependent on the Wiener index of the initial graph; see Theorem 2. For many initial graphs, the average distance decreases, and the diameter does not change over time. In Theorem 3, the clustering coefficient of the graph at time t is estimated and shown to tend to 0 slower than a $G(n,p)$ random graph with the same average degree. Experimental studies of social networks (see Estrada [15]) demonstrate smaller expansion properties than in other complex networks as a consequence of the existence of communities with low number of inter-community links. Interestingly, this phenomena is found in the ILT model, where a smaller spectral gap than in random graphs is found for both the normalized Laplacian (see Theorem 5) and adjacency (see Theorem 7) matrices.

2 The ILT Model

We first give a precise formulation of the model. The ILT model generates simple, undirected graphs $(G_t : t \geq 0)$ over a countably infinite sequence of discrete time-steps. The only parameter of the model is the initial graph G_0, which is any fixed finite *connected* graph. Assume that for a fixed $t \geq 0$, the graph G_t has been constructed. To form G_{t+1}, for each node $x \in V(G_t)$, add its *clone* x', such that x' is joined to x and all of its neighbours at time t. Note that the set of new nodes at time $t+1$ form an independent set of cardinality $|V(G_t)|$.

We write $\deg_t(x)$ for the degree of a node at time t, n_t for the order of G_t, and e_t for its number of edges. It is straightforward to see that $n_t = 2^t n_0$. Given a node x at time t, let x' be its clone. The simple but important recurrences governing the degrees of nodes are given as

$$\deg_{t+1}(x) = 2\deg_t(x) + 1, \tag{1}$$
$$\deg_{t+1}(x') = \deg_t(x) + 1. \tag{2}$$

2.1 Average Degree and Densification

We now consider the number of edges and average degree of G_t, and prove the following densification power law for the ILT model. Define the *volume* of G_t by

$$\mathrm{vol}(G_t) = \sum_{x \in V(G_t)} \deg_t(x) = 2e_t.$$

Theorem 1. *For $t > 0$, the average degree of G_t equals*

$$\left(\frac{3}{2}\right)^t \left(\frac{\mathrm{vol}(G_0)}{n_0} + 2\right) - 2.$$

Note that Theorem 1 supplies a densification power law with exponent $a = \frac{\log 3}{\log 2} \approx 1.58$. We think that the densification power law makes the ILT model realistic, especially in light of real-world data mined from complex networks (see [22]). Theorem 1 follows immediately from Lemma 1, since the average degree of G_t is $\mathrm{vol}(G_t)/n_t$.

Lemma 1. *For $t > 0$,*

$$\mathrm{vol}(G_t) = 3^t \mathrm{vol}(G_0) + 2n_0(3^t - 2^t).$$

In particular,

$$e_t = 3^t(e_0 + n_0) - n_t.$$

Proof. By (1) and (2) we have that

$$\begin{aligned}
\mathrm{vol}(G_{t+1}) &= \sum_{x \in V(G_t)} \deg_{t+1}(x) + \sum_{x' \in V(G_{t+1}) \setminus V(G_t)} \deg_{t+1}(x') \\
&= \sum_{x \in V(G_t)} (2\deg_t(x) + 1) + \sum_{x \in V(G_t)} (\deg_t(x) + 1) \\
&= 3\mathrm{vol}(G_t) + n_{t+1}.
\end{aligned} \tag{3}$$

Hence by (3) for $t > 0$,

$$\begin{aligned}
\mathrm{vol}(G_t) &= 3\mathrm{vol}(G_{t-1}) + n_t \\
&= 3^t \mathrm{vol}(G_0) + n_0 \left(\sum_{i=0}^{t-1} 3^i 2^{t-i} \right) \\
&= 3^t \mathrm{vol}(G_0) + 2n_0(3^t - 2^t),
\end{aligned}$$

where the third equality follows by summing a geometric series.

2.2 Average Distance

Define the *Wiener index* of G_t as

$$W(G_t) = \sum_{x,y \in V(G_t)} d_t(x,y).$$

The Wiener index arises in applications of graph theory to chemistry, and may be used to define the *average distance* of G_t as

$$L(G_t) = \frac{W(G_t)}{\binom{n_t}{2}}.$$

We will compute the average distance by deriving first the Wiener index. Define the *ultimate average distance of* G_0, as

$$UL(G_0) = \lim_{t \to \infty} L(G_t)$$

assuming the limit exists. We provide an exact value for $L(G_t)$ and compute the ultimate average distance for any initial graph G_0.

Theorem 2. *1. For* $t > 0$,

$$W(G_t) = 4^t \left(W(G_0) + (e_0 + n_0) \left(1 - \left(\frac{3}{4} \right)^t \right) \right).$$

2. For $t > 0$,

$$L(G_t) = 2 \left(\frac{4^t \left(W(G_0) + (e_0 + n_0) \left(1 - \left(\frac{3}{4} \right)^t \right) \right)}{4^t n_0^2 - 2^t n_0} \right).$$

3. For all graphs G_0,

$$UL(G_0) = \frac{2(W(G_0) + e_0 + n_0)}{n_0^2}.$$

Further, $UL(G_0) \leq L(G_0)$ *if and only if* $W(G_0) \geq (n_0 - 1)(e_0 + n_0)$.

Note that the average distance of G_t is bounded above by $\mathrm{diam}(G_0) + 1$ (in fact, by $\mathrm{diam}(G_0)$ in all cases except cliques). Further, the condition in (3) for $UL(G_0) < L(G_0)$ holds for large cycles and paths. Hence, for many initial graphs G_0, the average distance decreases, a property observed in on-line social and other networks (see [21, 22]).

When computing distances in the model, the following lemma is helpful. As its proof is elementary, we omit it.

Lemma 2. *Let* x *and* y *be nodes in* G_t *with* $t > 0$. *Then*

$$d_{t+1}(x', y) = d_{t+1}(x, y') = d_{t+1}(x, y) = d_t(x, y),$$

and

$$d_{t+1}(x', y') = \begin{cases} d_t(x,y) & \text{if } xy \notin E(G_t), \\ d_t(x,y) + 1 = 2 & \text{if } xy \in E(G_t). \end{cases}$$

Proof of Theorem 2. We only prove item (1), noting that items (2) and (3) follow from (1) by computation. We derive a recurrence for $W(G_t)$ as follows. To compute $W(G_{t+1})$, there are five cases to consider: distances within G_t, and distances of the forms: $d_{t+1}(x, y')$, $d_{t+1}(x', y)$, $d_{t+1}(x, x')$, and $d_{t+1}(x', y')$. The first three cases contribute $3W(G_t)$ by Lemma 2. The 4th case contributes n_t. The final case contributes $W(G_t) + e_t$ (the term e_t comes from the fact that each edge xy contributes $d_t(x, y) + 1$).

Thus,

$$\begin{aligned} W(G_{t+1}) &= 4W(G_t) + e_t + n_t \\ &= 4W(G_t) + 3^t(e_0 + n_0). \end{aligned}$$

Hence,

$$\begin{aligned} W(G_t) &= 4^t W(G_0) + \left(\sum_{i=0}^{t-1} 4^i \left(3^{t-1-i}\right) (e_0 + n_0) \right) \\ &= 4^t W(G_0) + 4^t (e_0 + n_0) \left(1 - \left(\frac{3}{4} \right)^t \right). \end{aligned}$$ □

Diameters are constant in the ILT model. We record this as a strong indication of the (ultra) small world property in the model.

Lemma 3. *For all graphs G_0 different than a clique,*

$$\operatorname{diam}(G_t) = \operatorname{diam}(G_0),$$

and $\operatorname{diam}(G_t) = \operatorname{diam}(G_0) + 1 = 2$ *when G_0 is a clique.*

2.3 The Clustering Coefficient and Degrees

The purpose of this subsection is to estimate the clustering coefficient of G_t. Let $N_t(x)$ be the neighbour set of x at time t, and let $e(x,t)$ be the number of edges in the subgraph of G_t induced by $N_t(x)$. For a node $x \in V(G_t)$ with degree at least 2 define

$$c_t(x) = \frac{e(x,t)}{\binom{\deg_t(x)}{2}}.$$

By convention $c_t(x) = 0$ if the degree of x is at most 1. The *clustering coefficient* of G_t is

$$C(G_t) = \frac{\sum_{x \in V(G_t)} c_t(x)}{n_t}.$$

Our main result is the following.

Theorem 3.

$$\Omega\left(\left(\frac{7}{8} \right)^t t^{-2} \right) = C(G_t) = O\left(\left(\frac{7}{8} \right)^t t^2 \right).$$

Observe that $C(G_t)$ tends to 0 as $t \to \infty$. If we let $n_t = n$ (so $t \sim \log_2 n$), then this gives that

$$C(G_t) = n^{\log_2(7/8) + o(1)}. \tag{4}$$

In contrast, for a random graph $G(n,p)$ with comparable average degree $pn = \Theta((3/2)^{\log_2 n}) = \Theta(n^{\log_2(3/2)})$ as G_t, the clustering coefficient is $p = \Theta(n^{\log_2(3/4)})$ which tends to zero much faster than $C(G_t)$.

We introduce the following dependency structure that will help us classify the degrees of nodes. Given a node $x \in V(G_0)$ we define its *descendant tree at time t*, written $T(x,t)$, to be a rooted binary tree with root x, and whose leaves are all of the nodes at time t. To define the $(k+1)$th row of $T(x,t)$, let y be a node in the kth row (y corresponds to a node in G_k). Then y has exactly two descendants on row $k+1$: y itself and y'. In this way, we may identify the nodes of G_t with a length t binary sequence corresponding to the descendants of x, using the convention that a clone is labelled 1. We refer to such a sequence as the *binary sequence for x* at time t. We need the following technical lemma whose proof is omitted.

Lemma 4. *Let $S(x,k,t)$ be the nodes of $T(x,t)$ with exactly k many 0's in their binary sequence at time t. Then for all $y \in S(x,k,t)$*

$$2^k(\deg_0(x)+1)+t-k-1 \;\leq\; \deg_t(y) \;\leq\; 2^k(\deg_0(x)+t-k+1)-1.$$

It follows from Lemma 4 that the number of nodes of degree at least k at time t, denoted by $N_{(\geq k)}$, satisfies

$$\sum_{i=\log_2 k}^{t} \binom{t}{i} \leq N_{(\geq k)} \leq \sum_{i=\log_2 k-\log_2 t}^{t} \binom{t}{i}.$$

In particular, $N_{(\geq k)} = \Theta(n_t)$ for $k \leq \sqrt{n_t}$, and therefore, the degree distribution of G_t does not follow a power law. Since $\binom{t}{k}$ nodes have degree around 2^k, the degree distribution has 'binomial type' behaviour. We now prove the following lemma.

Lemma 5. *For all $x \in V(G_t)$ with k 0's in their binary sequence, we have that*

$$\Omega(3^k) = e(x,t) = O(3^k t^2).$$

Proof. For $x \in V(G_t)$ we have that

$$\begin{aligned} e(x,t+1) &= e(x,t) + \deg_t(x) + \sum_{i=1}^{\deg_t(x)} (1 + \deg_{G_t \restriction N_t(x)}(x)) \\ &= 3e(x,t) + 2\deg_t(x), \end{aligned}$$

where $G_t \restriction N_t(x)$ is the subgraph induced by $N_t(x)$ in G_t. For x', we have that

$$e(x', t+1) = e(x,t) + \deg_t(x).$$

Since there are k many 0's and $e(x,2)$ is always positive for all initial graphs G_0, $e(x,t) \geq 3^{k-2}e(x,2) = \Omega(3^k)$ and the lower bound follows.

For the upper bound, a general binary sequence corresponding to x is of the form

$$(1,\ldots,1,0,1,\ldots,1,0,1,...,1,0,1,\ldots,1,0,1,\ldots,1)$$

with the 0's in positions i_k $(1 \leq i \leq k)$. Consider a path in the descendant tree from the root of the tree to node x. By Lemma 4, the node on the path in the ith row $(i < i_j)$ has (at time i) degree $O(2^{j-1}t)$.

Hence, the number of edges we estimate is $O(t^2)$ until the (i_1-1)th row, increases to $3O(t^2)+O(2^1t)$ in the next row, and increases to $3O(t^2)+O(2^1t^2)$ in the (i_2-1)th row. By induction, we have that

$$\begin{aligned} e(x,t) &= 3(\ldots(3(3O(t^2)+O(2^1t^2))+O(2^2t^2))\ldots)+O(2^kt^2) \\ &= O(t^2)3^k\sum_{i=0}^{k}\left(\frac{2}{3}\right)^j \\ &= O(3^kt^2). \end{aligned}$$

□

We now prove our result on clustering coefficients.

Proof of Theorem 3. For $x \in V(G_t)$ with k many 0's in its binary sequence, by Lemmas 4 and 5 we have that

$$c(x) = \Omega\left(\frac{3^k}{(2^kt)^2}\right) = \Omega\left(\left(\frac{3}{4}\right)^k t^{-2}\right),$$

and

$$c(x) = O\left(\frac{3^kt^2}{(2^k)^2}\right) = O\left(\left(\frac{3}{4}\right)^k t^2\right).$$

Hence, since we have $n_0\binom{t}{k}$ nodes with k many 0's in its binary sequence,

$$C(G_t) = \frac{\sum_{k=0}^{t} n_0\binom{t}{k}\Omega\left(\left(\frac{3}{4}\right)^k t^{-2}\right)}{n_0 2^t} = \Omega\left(\frac{t^{-2}\left(1+\frac{3}{4}\right)^t}{2^t}\right) = \Omega\left(\left(\frac{7}{8}\right)^t t^{-2}\right).$$

In a similar fashion, it follows that

$$C(G_t) = \frac{\sum_{k=0}^{t} n_0\binom{t}{k}O\left(\left(\frac{3}{4}\right)^k t^{2}\right)}{n_0 2^t} = O\left(\left(\frac{7}{8}\right)^t t^2\right). \qquad \square$$

3 Spectral Properties of the ILT Model

Social networks often organize into separate clusters in which the intra-cluster links are significantly higher than the number of inter-cluster links. In particular, social networks contain communities (characteristic of social organization), where tightly knit groups correspond to the clusters [17]. As a result, social networks possess bad expansion properties realized by small gaps between their first and second eigenvalues [15]. In this section, we find that the ILT model has such bad expansion properties for both its normalized Laplacian and adjacency matrices.

3.1 The Spectral Gap of the Normalized Laplacian

For regular graphs, the eigenvalues of the adjacency matrix are related to several important graph properties, such as in the expander mixing lemma. The normalized Laplacian of a graph, introduced by Chung [8], relates to important graph properties even in the case where the underlying graph is not regular (as is the case in the ILT model). Let A denote the adjacency matrix and D denote the diagonal adjacency matrix of a graph G. Then the normalized Laplacian of G is

$$\mathcal{L} = I - D^{-1/2}AD^{-1/2}.$$

Let $0 = \lambda_0 \leq \lambda_1 \leq \lambda_{n-1} \leq 2$ denote the eigenvalues of $\mathcal{L}$. The *spectral gap* of the normalized Laplacian is

$$\lambda = \max\{|\lambda_1 - 1|, |\lambda_{n-1} - 1|\}.$$

Chung, Lu, and Vu [11] observe that, for random power law graphs with some parameters (effectively in the case that $d_{min} \gg \log^2 n$), that $\lambda \leq (1 + o(1))\frac{4}{\sqrt{d}}$, where d is the average degree.

For the graphs G_t studied here, we observe that the spectra behaves quite differently and, in fact, the spectral gap has a constant order. The following theorem suggests a significant spectral difference between graphs generated by the ILT model and random graphs. Define $\lambda(t)$ to be the spectral gap of the normalized Laplacian of G_t.

Theorem 4. *For $t \geq 1$, $\lambda(t) > \frac{1}{2}$.*

Theorem 4 represents a drastic departure from the good expansion found in random graphs, where $\lambda = o(1)$ [8, 11], and from the preferential attachment model [18]. We use the expander mixing lemma for the normalized Laplacian (see [8]). For sets of vertices X and Y we use the notation $\mathrm{vol}(X)$ for the volume of the subgraph induced by X, and $e(X, Y)$ for the number of edges with one end in each of X and Y.

Lemma 6. *For all sets $X \subseteq G$,*

$$\left| e(X, X) - \frac{(\mathrm{vol}(X))^2}{\mathrm{vol}(G)} \right| \leq \lambda \frac{\mathrm{vol}(X)\mathrm{vol}(\bar{X})}{\mathrm{vol}(G)}.$$

Proof of Theorem 4. We observe that G_t contains an independent set (that is, a set of vertices with no edges) with volume $\mathrm{vol}(G_{t-1}) + n_{t-1}$. Let X denote this set, that is, the new nodes added at time t. Then by (3) it follows that

$$\mathrm{vol}(\bar{X}) = \mathrm{vol}(G_t) - \mathrm{vol}(X) = 2\mathrm{vol}(G_{t-1}) + n_{t-1}.$$

Since X is independent, Lemma 6 implies that

$$\lambda(t) \geq \frac{\mathrm{vol}(X)}{\mathrm{vol}(\bar{X})} = \frac{\mathrm{vol}(G_{t-1}) + n_{t-1}}{2\mathrm{vol}(G_{t-1}) + n_{t-1}} > \frac{1}{2}. \qquad \square$$

If G_0 has bad expansion properties, and has $\lambda_1 < 1/2$ (and thus, $\lambda > 1/2$) then, in fact, this trend of bad expansion continues as shown by the following theorem.

Theorem 5. *Suppose G_0 has at least two nodes, and for $t > 0$ let $\lambda_1(t)$ be the second eigenvalue of G_t. Then we have that*

$$\lambda_1(t) < \lambda_1(0).$$

Note that Theorem 5 implies that $\lambda_1(1) < \lambda_1(0)$ and this implies that the sequence $\{\lambda_1(t) : t \geq 0\}$ is strictly decreasing. This follows since G_t is constructed from G_{t-1} in the same manner as G_1 is constructed from G_0. If G_0 is K_1, then there is no second eigenvalue, but G_1 is K_2. Hence, in this case, the theorem implies that $\{\lambda_1(t) : t \geq 1\}$ is strictly decreasing.

Before we proceed with the proof of Theorem 5, we begin by stating some notation and a lemma. For a given node $u \in V(G_t)$, we let $\tilde{u} \in V(G_0)$ denote the node in G_0 that u is a descendant of. Given $uv \in E(G_0)$, we define

$$\mathcal{A}_{uv}(t) = \{xy \in E(G_t) : \tilde{x} = u, \tilde{y} = v\},$$

and for $v \in E(G_0)$, we set

$$\mathcal{A}_v(t) = \{xy \in E(G_t) : \tilde{x} = \tilde{y} = v\}.$$

We use the following lemma, for which the proof of items (1) and (2) follow from Lemma 1. The final item contains a standard form of the Raleigh quotient characterization of the second eigenvalue; see [8].

Lemma 7. *1. For $uv \in E(G_0)$,*

$$|\mathcal{A}_{uv}(t)| = 3^t.$$

2. For $v \in V(G_0)$,

$$|\mathcal{A}_v(t)| = 3^t - 2^t.$$

3. Define

$$\bar{d} = \frac{\sum_{v \in V(G_t)} f(v) \deg_t(v)}{\mathrm{vol}(G_t)}.$$

Then

$$\lambda_1(t) = \inf_{\substack{f : V(G_t) \to \mathbb{R}, \\ f \neq 0}} \frac{\sum\limits_{uv \in E(G_t)} (f(u) - f(v))^2}{\sum\limits_{v} f^2(v) \deg_t(v) - \bar{d}^2 \mathrm{vol}(G_t)}. \tag{5}$$

Note that in item (3), $\bar{d}$ is a function of f.

Proof of Theorem 5. Let $g : V(G_0) \to \mathbb{R}$ be the harmonic eigenvector for $\lambda_1(0)$ so that

$$\sum_{v \in V(G_0)} g(v) \deg_0(v) = 0,$$

and

$$\lambda_1(0) = \frac{\sum\limits_{uv \in E(G_0)} (g(u) - g(v))^2}{\sum\limits_{v \in V(G_0)} g^2(v) \deg_0(v)}.$$

Furthermore, we choose g scaled so that $\sum_{v\in V(G_0)} g^2(v)\deg_0(v) = 1$. This is the standard version of the Raleigh quotient for the normalized Laplacian from [8], so such a g exists so long as G_0 has at least two eigenvalues, which it does by our assumption that $G_0 \ncong K_1$. Our strategy in proving the theorem is to show that lifting g to G_1 provides an effective bound on the second eigenvalue of G_1 using the form of the Raleigh quotient given in (5).

Define $f : G_t \to \mathbb{R}$ by $f(x) = g(\tilde{x})$. Then note that

$$\begin{aligned}\sum_{xy\in E(G_t)} (f(x)-f(y))^2 &= \sum_{\substack{xy\in E(G_t),\\ \tilde{x}=\tilde{y}}} (f(x)-f(y))^2 + \sum_{\substack{xy\in E(G_t)\\ \tilde{x}\neq\tilde{y}}} (f(x)-f(y))^2\\ &= \sum_{uv\in E(G_0)} \sum_{xy\in\mathcal{A}_{uv}} (g(u)-g(v))^2\\ &= 3^t \sum_{uv\in E(G_0)} (g(u)-g(v))^2.\end{aligned}$$

By Lemma 7 (1) and (2) it follows that

$$\begin{aligned}\sum_{x\in V(G_t)} f^2(x)\deg_t(x) = \sum_{x\in V(G_t)}\sum_{xy\in E(G_t)} f^2(x) &= \sum_{u\in V(G_0)} \sum_{\substack{xy\in E(G_t),\\ \tilde{x}=u}} g^2(u)\\ &= \sum_{u\in V(G_0)} g^2(u)\left(\sum_{vu\in E(G_0)}\sum_{xy\in\mathcal{A}_{uv}} 1+2|\mathcal{A}_u|\right)\\ &= 3^t \sum_{u\in V(G_0)} g^2(u)\deg_0(u) + 2(3^t-2^t)\sum_{u\in V(G_0)} g^2(u)\\ &= 3^t + 2(3^t-2^t)\sum_{u\in G_0} g^2(u).\end{aligned}$$

By Lemma 1 and proceeding as above, noting that $\sum_{v\in V(G_0)} g(v)\deg_0(v) = 0$, we have that

$$\begin{aligned}\bar{d}^2\mathrm{vol}(G_t) &= \frac{\left(\sum_{x\in V(G_t)} f(x)\deg_t(x)\right)^2}{\mathrm{vol}(G_t)}\\ &= \frac{\left(2(3^t-2^t)\sum_{u\in V(G_0)} g(u)\right)^2}{\mathrm{vol}(G_t)}\\ &= \frac{4\cdot 3^{2t}\left(1-\left(\frac{2}{3}\right)^t\right)^2\left(\sum_{u\in V(G_0)} g(u)\right)^2}{3^t\left(\mathrm{vol}(G_0)+2n_0\left(1-\left(\frac{2}{3}\right)^t\right)\right)}\\ &\leq \frac{4\cdot 3^t\left(1-\left(\frac{2}{3}\right)^t\right)^2\sum_{u\in V(G_0)} g^2(u)}{\bar{D}+2\left(1-\left(\frac{2}{3}\right)^t\right)},\end{aligned}$$

where $\bar{D}$ is the average degree of G_0, and the last inequality follows from the Cauchy-Schwarz inequality.

By (5) we have that

$$\begin{aligned}
\lambda_1(t) &\le \frac{\sum\limits_{xy\in E(G_t)}(f(x)-f(y))^2}{\sum\limits_{x\in V(G_t)} f^2(x)\deg_t(x)+\bar{d}^2\mathrm{vol}(G_t)} \\
&\le \frac{3^t\sum\limits_{uv\in E(G_0)}(g(u)-g(v))^2}{3^t+2\cdot 3^t\left(1-\left(\frac{2}{3}\right)^t\right)\left(\sum_{u\in V(G_0)}g^2(u)\right)-\frac{4\cdot 3^t\left(1-\left(\frac{2}{3}\right)^t\right)^2\sum\limits_{u\in V(G_0)}g^2(u)}{\bar{D}+2\left(1-\left(\frac{2}{3}\right)^t\right)}} \\
&= \frac{\lambda_1(0)}{1+2\left(1-\left(\frac{2}{3}\right)^t\right)\left(\sum\limits_{u\in V(G_0)}g^2(u)\right)\left(1-\frac{2\left(1-\left(\frac{2}{3}\right)^t\right)}{\bar{D}+2\left(1-\left(\frac{2}{3}\right)^t\right)}\right)} \\
&< \lambda_1(0),
\end{aligned}$$

where the strict inequality follows from the fact that $\bar{D}\ge 1$ since G_0 is connected and $G_0\ncong K_1$. □

3.2 The Spectral Gap of the Adjacency Matrix

Let $\rho_0(t)\ge|\rho_1(t)|\ge\dots$ denote the eigenvalues of the adjacency matrix G_t. As in the Laplacian case, we can show that there is a small spectral gap of the adjacency matrix. If A is the adjacency matrix of G_t, then the adjacency matrix of G_{t+1} is

$$M=\begin{pmatrix} A & A+I \\ A+I & 0\end{pmatrix},$$

where is I is the identity matrix of order n_t. We note the following recurrence for the eigenvalues of the adjacency matrix of G_t, whose proof is omitted.

Theorem 6. *If ρ is an eigenvalue of the adjacency matrix of G_t, then*

$$\frac{\rho\pm\sqrt{\rho^2+4(\rho+1)^2}}{2},$$

are eigenvalues of the adjacency matrix of G_{t+1}.

Indeed, one can check that the eigenvectors of G_t can be written in terms of the eigenvalues of G_{t-1}. We prove the following theorem.

Theorem 7. *Let $\rho_0(t)\ge|\rho_1(t)|\ge\cdots\ge|\rho_n(t)|$ denote the eigenvalues of the adjacency matrix of G_t. Then*

$$\frac{\rho_0(t)}{|\rho_1(t)|}=\Theta(1).$$

That is, $\rho_0(t)\le c|\rho_1(t)|$ for some constant c. Theorem 7 is in contrast to fact that in $G(n,p)$ random graphs, $|\rho_0|=o(\rho_1)$.

Proof of Theorem 7. Without loss of generality, we assume that $G_0\ncong K_1$; otherwise, G_1 is K_2, and we may start from there. Thus, in particular, we can assume $\rho_0(0)\ge 1$.

We first observe that by Theorem 6

$$\rho_0(t)\ge\left(\frac{1+\sqrt{5}}{2}\right)^t\rho_0(0). \tag{6}$$

By Theorem 6 and by taking a branch of descendants from the largest eigenvalue it follows that

$$|\rho_1(t)| \geq \frac{2(\sqrt{5}-1)}{(1+\sqrt{5})^2}\left(\frac{1+\sqrt{5}}{2}\right)^t \rho_0(0).$$

Hence, to prove the theorem, it suffices to show that

$$\rho_0(t) \leq c\left(\frac{1+\sqrt{5}}{2}\right)^t \rho_0(0).$$

Observe that, also by Theorem 6 and taking the largest branch of descendants from the largest eigenvalues,

$$\rho_0(t) = \rho_0(0)\prod_{i=0}^{t-1}\left(\frac{1+\sqrt{5+\frac{8}{\rho_0(i)}+\frac{4}{\rho_0^2(i)}}}{2}\right) \leq \rho_0(0)\prod_{i=0}^{t-1}\left(\frac{1+\sqrt{5+\frac{6}{\rho_0(i)}}}{2}\right).$$

Thus,

$$\begin{aligned}
\frac{2^t\rho_0(t)}{(1+\sqrt{5})^t} &\leq \rho_0(0)\prod_{i=0}^{t-1}\frac{1+\sqrt{5+\frac{6}{\rho_0(i)}}}{1+\sqrt{5}} \\
&\leq \rho_0(0)\prod_{i=0}^{t-1}\left(1+\frac{\sqrt{5}}{1+\sqrt{5}}\frac{6}{5\rho_0(i)}\right) \\
&\leq \rho_0(0)\exp\left(\frac{6\sqrt{5}}{5(1+\sqrt{5})}\sum_{i=0}^{t-1}\rho_0(i)^{-1}\right) \\
&\leq \rho_0(0)\exp\left(\frac{6\sqrt{5}}{5(1+\sqrt{5})\rho_0(0)}\sum_{i=0}^{\infty}\left(\frac{2}{1+\sqrt{5}}\right)^{-i}\right) = \rho_0(0)c.
\end{aligned}$$

In all we have proved that for constants c and d that

$$c\left(\frac{1+\sqrt{5}}{2}\right)^t \rho_0(0) \geq \rho_0(t) \geq |\rho_1(t)| \geq d\left(\frac{1+\sqrt{5}}{2}\right)^t \rho_0(t). \qquad \square$$

4 Conclusion and further work

We introduced the ILT model for on-line social and other complex networks, where the network is cloned at each time-step. We proved that the ILT model generates graphs with a densification power law, in many cases decreasing average distance (and in all cases, the diameter is bounded above by a constant independent of time), have higher clustering than random graphs with the same average degree, and have smaller spectral gaps for both their normalized Laplacian and adjacency matrices than in random graphs.

Much more can be said about the ILT model than space permits here; for example, many graph properties at time t are strongly related to properties from time 0. For example, the cop and domination numbers of the graphs G_t equal those of G_0 (see [5] for definitions of these parameters). In addition, the automorphism group (endomorphism monoid) of G_0 embeds as a subgroup (submonoid) in the automorphism group (endomorphism monoid) of G_t. A discussion of these and other properties of the ILT model will appear in the full version of this paper.

In the duplication and copying models and in the model [14] of social networks, transitivity is modelled so that neighbours are copied with some fixed probability. The ILT model may be randomized, so that x' clones x with a fixed probability. We will study this randomized ILT model in future work. As we

noted after the statement of Lemma 4 the ILT model does not generate graphs with a power law degree distribution. An interesting problem that we will address in the full version of this paper is to design and analyze a randomized version of the ILT model satisfying the properties displayed in the deterministic ILT model (and make them tuneable; for example, the densification power law exponent should vary with the choice of parameters) as well as generating power law graphs. Such a randomized ILT model should with high probability generate power law graphs with topological properties similar to graphs from the deterministic ILT model.

References

1. L.A. Adamic, O. Buyukkokten, E. Adar, A social network caught in the web, *First Monday* **8** (2003).
2. Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, Analysis of topological characteristics of huge on-line social networking services, In: *Proceedings of the 16th International Conference on World Wide Web*, 2007.
3. G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau, S.C. Sahinalp, The degree distribution of the generalized duplication model, *Theoretical Computer Science* **369** (2006) 234-249.
4. A. Bhan, D.J. Galas, T.G. Dewey, A duplication growth model of gene expression networks, *Bioinformatics* **18** (2002) 1486-1493.
5. A. Bonato, *A Course on the Web Graph*, American Mathematical Society Graduate Studies Series in Mathematics, Providence, Rhode Island, 2008.
6. A. Bonato, J. Janssen, Infinite limits and adjacency properties of a generalized copying model, accepted to *Internet Mathematics*.
7. G. Caldarelli, *Scale-Free Networks*, Oxford University Press, Oxford, 2007.
8. F. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, Rhode Island, 1997.
9. F. Chung, L. Lu, T. Dewey, D. Galas, Duplication models for biological networks, *Journal of Computational Biology* **10** (2003) 677-687.
10. F. Chung, L. Lu, *Complex graphs and networks*, American Mathematical Society, Providence, Rhode Island, 2006.
11. F. Chung, L. Lu, V. Vu, The spectra of random graph with given expected degrees, *Internet Mathematics* **1** (2004) 257–275.
12. D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, S. Suri, Feedback effects between similarity and social influence in on-line communities, In: *Proceedings of the 14th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2008.
13. R. Durrett, *Random Graph Dynamics*, Cambridge University Press, New York, 2006.
14. H. Ebel, J. Davidsen, S. Bornholdt, Dynamics of social networks, *Complexity* **8** (2003) 24-27.
15. E. Estrada, Spectral scaling and good expansion properties in complex networks *Europhys. Lett.* **73** (2006) 649-655.
16. O. Frank, Transitivity in stochastic graphs and digraphs, *Journal of Mathematical Sociology* **7** (1980) 199-213.
17. M. Girvan, M.E.J. Newman. Community structure in social and biological networks, *Proceedings of the National Academy of Sciences* **99** (2002) 7821-7826.
18. C. Gkantsidis, M. Mihail, A. Saberi, Throughput and congestion in power-law graphs, In: *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement Modeling of Computer Systems*, 2003.
19. S. Golder, D. Wilkinson, B. Huberman, Rhythms of social interaction: messaging within a massive on-line network, In: *3rd International Conference on Communities and Technologies*, 2007.
20. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, Stochastic models for the web graph, In: *Proceedings of the 41th IEEE Symposium on Foundations of Computer Science*, 2000.
21. R. Kumar, J. Novak, A. Tomkins, Structure and evolution of on-line social networks, In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.
22. J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification Laws, shrinking diameters and possible explanations, In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.
23. J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication, In: *Proceedings of European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2005.
24. D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins, Geographic routing in social networks, *Proceedings of the National Academy of Sciences* **102** (2005) 11623-11628.
25. S. Milgram, The small world problem, *Psychology Today* **2** (1967) 60-67.
26. A. Mislove, M. Marcon, K. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of on-line social networks, In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 2007.
27. R. Pastor-Satorras, E. Smith, R.V. Sole, Evolving protein interaction networks through gene duplication, *J. Theor. Biol.* **222** (2003) 199-210.
28. J.P. Scott, *Social Network Analysis: A Handbook*, Sage Publications Ltd, London, 2000.
29. D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393** (1998) 440-442.
30. H. White, S. Harrison, R. Breiger, Social structure from multiple networks, I: Blockmodels of roles and positions, *American Journal of Sociology* **81** (1976) 730-780.