# Complex Networks and Social Networks

Anthony Bonato and Amanda Tian

## 1 Introduction

Complex networks arise in many diverse contexts, ranging from web pages and their links, protein-protein interaction networks, and social networks. The modelling and mining of these large-scale, self-organizing systems is a broad effort spanning many disciplines. A number of common properties have been observed in complex networks, such as power law degree distributions and the small world property (see Section 2 for further background on these properties).

While classical binomial random graphs form a well studied field in their own right, in the last decade we have seen a wealth of new random graph models modelling complex networks. These stochastic graph models simulate properties of complex networks, but also expanding our theoretical understanding of random graphs. Models for complex networks also give insight into the underlying generative properties of complex networks, and can serve as a predictive tool in their evolution.

With the current popularity of *on-line social networks* (or *OSNs*) such as Facebook, LinkedIn, and Twitter, there is an increasing interest in their measurement and modelling. In addition to other complex networks properties, OSNs exhibit shrinking distances over time, increasing average degree, and bad spectral expansion. Unlike other complex networks such as the web graph, models for OSNs are relatively new and lesser known. In on-line social networks, models may help detect and classify communities, and better clarify how news and gossip is spread in social networks.

Anthony Bonato
Department of Mathematics, Ryerson University, Toronto, ON, Canada, M5B 2K3, e-mail: abonato@ryerson.ca

Amanda Tian
Department of Mathematics, Ryerson University, Toronto, ON, Canada, M5B 2K3, e-mail: yanhua.tian@ryerson.ca

We will survey the properties of complex networks and their models, focusing on the case of OSNs. A more detailed survey of complex networks (without the focus on OSNs) may be found in the book [4]. In Section 2 we give a brief overview of some the main observed properties of OSNs. These are the key properties that network models attempt to simulate. We study various complex network models in Section 3. Our focus is on five models, each rigorously formulated and analyzed: Kronecker graphs [25, 26], the MAG model [20], the $G(Q,E)$ affiliation graph model [24], the ILT model [5], and the GEO-P model [7]. These models focus primarily on simulating properties of social networks, and are relatively recent. We finish with a list of open problems surrounding the modelling of OSNs.

## 2 Properties of Complex Networks

Researchers are now in the enviable position of observing how OSNs evolve over time, and as such, network analysis and models of OSNs typically incorporate time as a parameter. Unlike in traditional social network analysis, we can now mine the social interactions of millions of people from across the globe. While by no means exhaustive, some of the main observed properties of OSNs include the following. For definitions of the terms used below (such as diameter, clustering coefficient, etc), see [4].

(i) *Large-scale.* OSNs are examples of complex networks with number of nodes (which we write as $n$) often in the millions; further, some users have disproportionately high degrees. About half a billion users are registered on Facebook [12]. Some of the nodes of Twitter corresponding to well-known celebrities including Lady Gaga and Justin Bieber have degree over ten million [36].

(ii) *Small world property and shrinking distances.* The small world property, introduced by Watts and Strogatz [38], is a central notion in the study of complex networks (see also [21]). The small world property demands a low diameter of $O(\log n)$, and a higher clustering coefficient than found in a binomial random graph with the same number of nodes and same average degree. Adamic et al. [1] provided an early study of an OSN at Stanford University, and found that the network has the small world property. Similar results were found in [2] which studied Cyworld, MySpace, and Orkut, and in [32] which examined data collected from Flickr, YouTube, LiveJournal, and Orkut. Low diameter (of 6) and high clustering coefficient were reported in the Twitter by both Java et al. [19] and Kwak et al. [23]. Kumar et al. [22] reported that in Flickr and Yahoo!360 the diameter actually decreases over time. Similar results were reported for Cyworld in [2]. Well-known models for complex networks such as preferential attachment or copying models have logarithmically growing diameters with time.

(iii) *Power law degree distributions.* In a graph $G$ of order $n$, let $N_k = N_k(n)$ be the number of nodes of degree $k$. The degree distribution of $G$ follows a *power law* if $N_k$ is proportional to $k^{-b}$, for a fixed exponent $b > 2$ and some range of $k$. Power laws were observed over a decade ago in subgraphs sampled from the web graph, and are ubiquitous properties of complex networks (see Chapter 2 of [4]). Kumar, Novak, and Tomkins [22] studied the evolution of Flickr and Yahoo!360, and found that these networks exhibit power-law degree distributions. Power law degree distributions for both the in- and out-degree distributions were documented in Flickr, YouTube, LiveJournal, and Orkut [32], as well as in Twitter [19, 23].

(iv) *Bad spectral expansion.* 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 [33]. As a result, it is reported in [11] that social networks, unlike other complex networks, possess bad spectral expansion properties realized by small gaps between the first and second eigenvalues of their adjacency matrices.

(v) *Bad compressibility.* A recent study of [8] contrasts the compressibility of OSNs with the web graph. Assume that the vertex set of the digraph $G$ is given by $[n] = \{1,2,\ldots,n\}$. The so-called *minimum logarithmic arrangement* or *MLOGA problem*, is to find a permutation $\pi : V(G) \to [n]$ such that the term

$$\sum_{(u,v)\in E} \log|\pi(u) - \pi(v)| \qquad (1)$$

is minimized. The motivating idea is minimize the sum of the edge lengths according to the ordering of vertices. The cost (1) represents the compression size in an encoding that is nearly informational-theoretically optimal. While MLOGA is **NP**-hard [8], the authors of [8] introduce heuristics for its computation. Using data from LiveJournal and Flickr, it was found in [8] that the compression performance with different orderings were worse than that found in web graph samples. The lack of a natural ordering of social networks when compared say to the URL ordering of web pages may be the cause of the poor incompressibility. Nevertheless, bad compressibility appears to be another feature peculiar to OSNs when, say, contrasted with the web graph.

(vi) *Densification power law.* Let $(G_t : t \geq 0)$ be sequence of graphs such that $G_t$ is an induced subgraph of $G_{t+1}$ for all $t \geq 0$, and suppose that $G_t$ has $e_t$ edges and $n_t$ nodes. The graph sequence satisfies a *densification power law* if there is a constant $a \in (1,2)$ such that for sufficiently large $t$, $e_t$ is proportional to $n_t^a$. We call $a$ the *exponent* of the densification power law. In particular, the average degree of the network grows to infinity with the order of the network . In [25], 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. Densification power laws were found in Flickr and Yahoo!360 in [22].

## 3 Models of Complex Networks

In this chapter, we do not give an exhaustive overview of models for complex networks. A survey of such models may be found in Chapter 4 of the book [4]. We focus, rather, on the relatively new models for OSNs introduced over the last few years. We are content to survey the results here, pointing the reader to further details and proofs in the papers cited. Many properties of the models hold with probability tending to 1 as time (or the order of the graphs considered) tends to infinity; we say such properties hold *asymptotically almost surely*, or *aas*.

### *3.1 Kronecker graphs*

Kronecker graphs [25, 26] were one of the early successful models for complex networks with densification. Their definition relies on a certain well known graph product. Given graphs $G$ and $H$, form the *categorical* (or *Kronecker*) *product* $G \times H$ by setting vertices to be pairs $(a,b)$ with $a \in V(G)$ and $b \in V(H)$, and $(a,b)$ joined to $(c,d)$ if and only $a$ is joined to $c$ in $G$, and $b$ is joined to $d$ in $H$. See Figure 1.

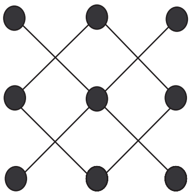

**Fig. 1** The Kronecker product of the path with three vertices with itself.

Let $A$ and $B$ be two real matrices, with sizes $n \times m$ and $n' \times m'$, respectively. The *Kronecker* (or *tensor*) *product* of $A$ and $B$, is the matrix $A \otimes B$ with size $nn' \times mm'$ given by

$$\begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}B & a_{n2}B & \cdots & a_{nm}B \end{pmatrix}.$$

If $A(G)$ is the adjacency matrix of $G$, then note that

$$A(G \times H) = A(G) \otimes A(H).$$

The Kronecker graphs are formed by forming the $k$th power $G^k$ of $G$ with respect to this product; we call $G$ here the *initiator graph*. The motivation behind this definition is that to produce $G^k$ from $G^{k-1}$, nodes of a community expand to copies of the community. Note that the Kronecker model is a deterministic one.

The authors prove the following theorem, which leads to power law graphs by the choice of the initiator graph.

**Theorem 1.** *[25, 26] Kronecker graphs have multinomial degree distributions.*

Further, Kronecker graphs satisfy a densification power law and constant diameter.

**Theorem 2.** *[25, 26] Kronecker graphs $G^k$ satisfy a densification power law with exponent*

$$\log(|E(G)|)/\log(|V(G)|).$$

*Further, if $G$ is reflexive, then the diameter of $G^k$ is the diameter of $G$.*

The model is made tuneable by allowing the initial adjacency matrix $A(G)$ to have real entries in $[0,1]$. Hence, we may think of the initiator graph $G$ as a probability space, where the probability there is an edge between $i$ and $j$ is the $ij$ entry of $A(G)$ (although this is not exactly the case as the row (or column) sums may add up to a quantity greater than $1$). Such *stochastic Kronecker graphs* with certain initiator graphs of order 2, were studied by Mahdian and Xu [31]. They studied the giant component of graphs generated by the model, and proved it *aas* has a constant diameter beyond the connectivity threshold.

It is shown in [27] that the Kronecker graph model with certain $2\times 2$ initiator matrices is useful in simulating complex networks. Their work shows that certain stochastic Kronecker matrices fit samples of the web graph, the internet AS graph, Flickr, and certain biological networks. A fast, scalable algorithm KRONFIT was introduced to fit real network data to Kronecker graphs.

## 3.2 The ILT model

The *Iterated Local Transitivity* (ILT) model [5], simulates OSNs and other complex networks. 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, [15, 34, 39]). 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 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 $|V(G_t)|$. As with Kronecker model, the ILT model is deterministic.

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. Define the *volume* of $G_t$ by

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

**Theorem 3.** *[5] 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 3 supplies a densification power law with exponent $a = \frac{\log 3}{\log 2} \approx 1.58$.

Define the *Wiener index* of $G_t$ as

$$W(G_t) = \frac{1}{2} \sum_{x,y \in V(G_t)} d(x,y).$$

The Wiener index may be used to define the *average distance* of $G_t$ as

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

**Theorem 4.** *[5] For $t > 0$,*

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

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, for many initial graphs $G_0$ (such as large cycles) the average distance decreases.

The clustering coefficient of the graph at time $t$ generated by the ILT model is estimated as follows.

**Theorem 5.** *[5]*

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

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

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

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 $C(G_t)$. (For a discussion of the clustering coefficient of $G(n,p)$, see Chapter 2 of [4].)

Let $A$ denote the adjacency matrix and $D$ denote the diagonal adjacency matrix of a graph $G$ of order $n$. Then the normalized Laplacian of $G$ is

$$\mathscr{L} = I - D^{-1/2}AD^{-1/2},$$

where $I$ is the $n \times n$ identity matrix. Let $0 = \lambda_0 \leq \lambda_1 \cdots \leq \lambda_{n-1} \leq 2$ denote the eigenvalues of $\mathscr{L}$. The *spectral gap* of the normalized Laplacian is

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

The following theorem suggests a significant spectral difference between graphs generated by the ILT model and random graphs. Define $\lambda(G_t)$ to be the spectral gap of the normalized Laplacian of $G_t$.

**Theorem 6.** *[5] For $t \geq 1$, $\lambda(G_t) > \frac{1}{2}$.*

Theorem 6 represents a drastic departure from the good expansion found in random graphs, where $\lambda = o(1)$ [9].

Let $\rho_0(t) \geq |\rho_1(t)| \geq \dots$ denote the eigenvalues of the adjacency matrix of $G_t$. 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 $I$ is the identity matrix of order $n_t$. We note the following recurrence for the eigenvalues of the adjacency matrix of $G_t$. As in the Laplacian case, there is a small spectral gap of the adjacency matrix.

**Theorem 7.** *[5] Let $\rho_0(t) \geq |\rho_1(t)| \geq \cdots \geq |\rho_{n-1}(t)|$ denote the eigenvalues of the adjacency matrix of $G_t$. Then*

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

That is, $\rho_1(t) \geq c|\rho_0(t)|$ for some constant $c > 0$. Theorem 7 is in contrast to the fact that in $G(n,p)$ random graphs, $|\rho_1| = o(\rho_0)$ (see [9]).

As shown in Theorem 3, the ILT model has a fixed densification exponent equalling $\log 3/\log 2$. A randomized version of the model, where edges are randomly added between new nodes, is presented in [5]. In the randomized model, the densification exponent is tuneable, and with high probability it generates graphs with the small world property and bad spectral expansion.

### 3.3 Affiliation networks

In [24], a model for social networks was given by first introducing a bipartite model called *affiliation graphs*. Paths of length two in the affiliation graphs are *folded* onto edges to derive a model for social networks. The central thesis behind using a folded affiliation network is that friendships between members of social networks arise from common shared affiliations, such as sharing the same hobby or profession.

More precisely, the model evolves in two stages. First, a bipartite random graph model $B(Q,U)$ is introduced, with colours $Q$ and $U$. For instance, $Q$ represents a set of users, while $U$ represents a set of groups of users. The parameters of the models are positive integers $c_q$ and $c_u$, along with a probability $p \in (0,1)$. The model evolves over discrete time-steps. At time $t = 0$, the graph $B_0(Q,U)$ is a (deterministic) bipartite graph with at least $c_q c_u$ edges, so that each node in $Q$ has degree at least $c_q$, while each node of $U$ has degree at least $c_u$. At time $t > 0$, a new node $q$ is added to $Q$. A node $q'$ from $Q$ is chosen proportional to its degree, and $c_q$ neighbours of $q'$ chosen uniformly at random (without replacement) become neighbours of $q$. Similarly, a node $u$ is added to $U$ with a similar copying process.

Now to define the (multi)graph $G(Q,U)$, the parameters of the models are positive integers $c_q$, $c_u$, and $s$ along with a probability $p \in (0,1)$. At $t = 0$, $G_0(Q,U)$ is the set $Q$ in $B_0(Q,U)$, and two nodes of $Q$ have an edge between them for each common neighbour they share in $U$. At time $t > 0$ we do the following. With probability $p$ a new node $q$ is added to $Q$. The edges of $q$ are determined in $B(Q,U)$, and edges are added between $q$ and other nodes if they share common neighbours in $U$. With probability $1 - p$, an edge is added between existing nodes $q_1$ and $q_2$ if they share as a common neighbour the new vertex $u$ in $U$. A set of $s$ nodes are chosen independently of each of other, proportionally by degree, and are joined to $q$.

It is proved in [24] that *aas* the degree distributions of the graphs generated by $G(Q,U)$ follows a power law. Further, if $c_u < \frac{p}{1-p} c_q$, then *aas* the graph $G(Q,U)$ is dense with $\omega(|Q|)$ many edges.

For a graph $G$, let $R$ be the set of node pairs which are connected by a path. For $0 < q < 1$, define the *q-effective diameter of* $G$ to be the minimum $d$ such that, for at least $q|R|$ of node pairs in $R$, their distance is at most $d$. The $G(Q,U)$ exhibits low distances between nodes as made precise by the following theorem.

**Theorem 8.** *[24] For constants $m, q \in (0,1)$, if $c_u < \frac{p}{1-p} c_q$, then* aas *the q-effective diameter of the graph $G(Q,U)$ is non-increasing.*

### 3.4 The MAG model

In the *Multiplicative Attribute Graph* (or *MAG*) model [20], nodes are assigned a set of attributes represented by a binary vectors. These could be viewed as answer to yes or no questions about the users interests or background. The MAG model accounts for *heterophily* (that is, love of the same) and *homophily* (that is, love of the different). More precisely, the MAG model $M(n,r,\mu,\Theta)$ has parameters equalling $n$ the number of nodes, $r$ the number of attributes of each node, $\mu$ the probability that an attribute takes the value of 1, and $\Theta$ the attribute affinity matrix

$$\begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix},$$

where $\alpha > \beta > \gamma$ are fixed probabilities in $(0,1)$. We use the notation $\Theta_{00} = \alpha$, $\Theta_{01} = \Theta_{10} = \beta$, and $\Theta_{11} = \gamma$. Each node $u$ is assigned a binary *attribute vector* $a(u)$ of length $r$; we denote the $i$th entry of $a(u)$ by $a_i(u)$. An independent and identically distributed Bernouilli distribution parameterized by $\mu$ is used to model the attribute vectors, where the probability that the $i$th attribute of a node is 1 is given by $\mu$. The probability that nodes $u$ and $v$ are joined is given, independently, by

$$\prod_{i=1}^{r} \Theta_{a_i(u)a_i(v)}.$$

In particular, the $i$th entry of attribute vectors $a_i(u)$ and $a_i(v)$ selects the entry of the matrix $\Theta$; for example, if $a_i(u) = 0$ and $a_i(v) = 0$, then $\Theta_{a_i(u)a_i(u)} = \alpha$. The product is then taken of all these entries. If the values on the diagonal of $\Theta$ are large, then the link probability is high when nodes share the same attributes. For instance, the matrix

$$\begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.8 \end{pmatrix}$$

represents homophily, while

$$\begin{pmatrix} 0.2 & 0.9 \\ 0.9 & 0.1 \end{pmatrix}$$

represents heterophily. It is assumed that $r = d \log n$ for some constant $d$ (see also the Logarithmic Dimension Hypothesis in item of (1) of Section 4).

The MAG model generates graphs which satisfy a densification power law.

**Theorem 9.** *[20] The expected number of edges of graphs generated by $M(n,r,\mu,\Theta)$ is*

$$\frac{n(n-1)}{2}(\mu^2\alpha + 2\mu(1-\mu)\beta + (1-\mu)^2\gamma)^r + n(\mu\alpha + (1-\mu)\gamma)^r.$$

The diameter of MAG graphs is also low.

**Theorem 10.** *[20] If $(\mu\beta + (1-\mu)\gamma)^d > 1/2$, then* aas *$M(n,r,\mu,\Theta)$ has constant diameter.*

Under certain assumptions, the MAG model follows a log-normal degree distribution. With more parameters, a variation of the model *aas* generates graphs whose degree distribution follows a power law.

### *3.5 The GEO-P model*

Our next and final model [7] uses both the notions of embedding the nodes in a metric space (geometric), and a link probability based on a ranking of the nodes (protean). We identify the users of an OSN with points in $m$-dimensional Euclidean space. Each node has a region of influence, and nodes may be joined with a certain probability if they land within each others region of influence. Nodes are ranked by their popularity from 1 to $n$, where $n$ is the number of nodes, and 1 is the highest ranked node. Nodes that are ranked higher have larger regions of influence, and so are more likely to acquire links over time. For simplicity, we consider only undirected graphs. The number of nodes $n$ is fixed but the model is dynamic: at each time-step, a node is born and one dies. A static number of nodes is more representative of the reality of OSNs, as the number of users in an OSN would typically have a maximum (an absolute maximum arises from roughly the number of users on the internet, not counting multiple accounts). For a discussion of ranking models for complex networks, see [14, 16, 18, 30].

We now formally define the GEO-P model. The model produces a sequence $(G_t : t \geq 0)$ of undirected graphs on $n$ nodes, where $t$ denotes time. We write $G_t = (V_t, E_t)$. There are four parameters: the *attachment strength* $\alpha \in (0,1)$, the *density parameter* $\beta \in (0, 1-\alpha)$, the *dimension* $m \in \mathbb{N}$, and the *link probability* $p \in (0,1]$. Each node $v \in V_t$ has rank $r(v,t) \in [n]$ (we use $[n]$ to denote the set $\{1,2,\ldots,n\}$). The rank function $r(\cdot,t) : V_t \rightarrow [n]$ is a bijection for all $t$, so every node has a unique rank. The highest ranked node has rank equal to 1; the lowest ranked node has rank $n$. The initialization and update of the ranking is done by *random initial rank* (Other ranking schemes may also be used. We use random initial rank for its simplicity.) In particular, the node added at time $t$ obtains an initial rank $R_t$ which is randomly chosen from $[n]$ according to a prescribed distribution. Ranks of all nodes are adjusted accordingly. Formally, for each $v \in V_{t-1}$ that is not deleted at time $t$,

$$r(v,t) = r(v,t-1) + \delta - \gamma,$$

where $\delta = 1$ if $r(v,t-1) > R_t$ and 0 otherwise, and $\gamma = 1$ if the rank of the node deleted in step $t$ is smaller than $r(v,t-1)$, and 0 otherwise.

Let $S$ be the unit hypercube in $\mathbb{R}^m$, with the torus metric $d(\cdot,\cdot)$ derived from the $L_\infty$ metric. More precisely, for any two points $x$ and $y$ in $\mathbb{R}^m$, their distance is given by

$$d(x,y) = \min\{||x-y+u||_\infty : u \in \{-1,0,1\}^m\}.$$

The torus metric is chosen so that there are no boundary effects.

To initialize the model, let $G_0 = (V_0, E_0)$ be any graph on $n$ nodes that are chosen from $S$. We define the *influence region* of node $v$ at time $t \geq 0$, written $R(v,t)$, to be the ball around $v$ with volume

$$|R(v,t)| = r(v,t)^{-\alpha} n^{-\beta} .$$

For $t \geq 1$, we form $G_t$ from $G_{t-1}$ according to the following rules.

1. Add a new node $v$ that is chosen *uniformly at random* from $S$. Next, independently, for each node $u \in V_{t-1}$ such that $v \in R(u,t-1)$, an edge $vu$ is created with probability $p$. Note that the probability that $u$ receives an edge is proportional to $p\,r(u,t-1)^{-\alpha}$. The negative exponent guarantees that nodes with higher ranks ($r(u,t-1)$ close to 1) are more likely to receive new edges than lower ranks.
2. Choose uniformly at random a node $u \in V_{t-1}$, delete $u$ and all edges incident to $u$.
3. Node $v$ obtains an initial rank $r(v,t) = R_t$ which is randomly chosen from $[n]$ according to a prescribed distribution.
4. Update the ranking function $r(\cdot,t) : V_t \to [n]$.

Since the process is an ergodic Markov chain, it will converge to a stationary distribution. (See [28] for more on Markov chains.) The random graph corresponding to this distribution with given parameters $\alpha,\beta,m,p$ is called the *geo-protean* graph (or *GEO-P* model), and is written GEO-P$(\alpha,\beta,m,p)$.

Let $N_k = N_k(n,p,\alpha,\beta)$ denote the number of nodes of degree $k$, and $N_{\geq k} = \sum_{l \geq k} N_l$. The following theorem demonstrates that the geo-protean model generates power law graphs with exponent

$$b = 1 + 1/\alpha. \tag{2}$$

Note that the variables $N_{\geq k}$ represent the cumulative degree distribution, so the degree distribution of these variables has power law exponent $1/\alpha$.

**Theorem 11.** *[7] Let $\alpha \in (0,1)$, $\beta \in (0,1-\alpha)$, $m \in \mathbb{N}$, $p \in (0,1]$, and*

$$n^{1-\alpha-\beta}\log^{1/2} n \leq k \leq n^{1-\alpha/2-\beta}\log^{-2\alpha-1} n.$$

*Then* aas *GEO-P$(\alpha,\beta,m,p)$ satisfies*

$$N_{\geq k} = \left(1 + O(\log^{-1/3} n)\right)\frac{\alpha}{\alpha+1} p^{1/\alpha} n^{(1-\beta)/\alpha} k^{-1/\alpha}.$$

Geo-protean graphs are relatively dense.

**Theorem 12.** *[7]* Aas *the average degree of GEO-P$(\alpha,\beta,m,p)$ is*

$$d = (1+o(1))\frac{p}{1-\alpha} n^{1-\alpha-\beta}. \tag{3}$$

Note that the average degree tends to infinity with $n$; that is, the model generates graphs satisfying a *densification power law*. While the diameter is not shrinking, it can be made constant by allowing the dimension to grow as a logarithmic function of $n$.

**Theorem 13.** *[7] Let $\alpha \in (0,1)$, $\beta \in (0,1-\alpha)$, $m \in \mathbb{N}$, and $p \in (0,1]$. Then* aas *the diameter $D$ of GEO-P$(\alpha,\beta,m,p)$ satisfies*

$$D = \Omega(n^{\frac{\beta}{(1-\alpha)m}} \log^{\frac{-\alpha}{m}} n), \text{ and } D = O(n^{\frac{\beta}{(1-\alpha)m}} \log^{\frac{2\alpha}{(1-\alpha)m}} n). \tag{4}$$

*In particular,* aas *the order of the diameter can be expressed as:*

$$\log D = \frac{\beta}{(1-\alpha)m} \log n + O\left(\frac{\log\log n}{m}\right).$$

If $m = C\log n$, for some constant $C > 0$, then *aas* we obtain a diameter bounded above by a constant.

Aas the GEO-P model, for some values of $m$, generates graphs with higher clustering coefficient than in a random graph $G(n,d/n)$ with the same expected average degree. We use the notation $\lfloor x \rfloor_2$ to denote the largest *even* integer smaller than or equal to $x$.

**Theorem 14.** *[7]* Aas *the clustering coefficient of G sampled from GEO-P*$(\alpha,\beta,m,p)$ *satisfies the following inequality*

$$\begin{aligned} c(G) &\geq (1+o(1))\left(\frac{3}{4}\left(1-\frac{2}{3K}\right)\right)^m \left(\frac{1-\alpha}{1+\alpha}\right)p \\ &= (1+o(1))\exp\left(-f\left(\frac{m}{K}\right)\right)\left(\frac{3}{4}\right)^m\left(\frac{1-\alpha}{1+\alpha}\right)p, \end{aligned}$$

*where* $f(\frac{m}{K}) = \Theta(\frac{m}{K})$, *and*

$$K = \left\lfloor \left(\frac{n^{1-\alpha-\beta}}{\log^3 n}\right)^{1/m} \right\rfloor_2 .$$

Note that if

$$m \leq (1-\alpha-\beta)\frac{\log n}{\log\log n}\left(1-\frac{1}{\log\log n}\right) = (1+o(1))(1-\alpha-\beta)\frac{\log n}{\log\log n},$$

then $K \gg m$, and the clustering coefficient of GEO-P$(\alpha,\beta,m,p)$ is *aas* at least

$$(1+o(1))\left(\frac{3}{4}\right)^m\left(\frac{1-\alpha}{1+\alpha}\right)p = n^{o(1)} \gg (1+o(1))\frac{p}{1-\alpha}n^{-\alpha-\beta} = c(G(n,d/n)).$$

Hence, the clustering coefficient is larger than that of a comparable random graph.

The next theorem represents a drastic departure from the good expansion found in binomial random graphs, where $\lambda = o(1)$ [9, 10].

**Theorem 15.** *[7] Let* $\alpha \in (0,1)$, $\beta \in (0,1-\alpha)$, $m \in \mathbb{N}$, *and* $p \in (0,1]$. *Let* $\lambda(n)$ *be the spectral gap of the normalized Laplacian of GEO-P*$(\alpha,\beta,m,p)$. *Then* aas

1. *If* $m = m(n) = o(\log n)$, *then* $\lambda(n) = 1 + o(1)$.
2. *If* $m = m(n) = C\log n$ *for some* $C > 0$, *then*

$$\lambda(n) \geq 1 - \exp\left(-\frac{\alpha+\beta}{C}\right).$$

Given an OSN, we describe how we may estimate the corresponding dimension parameter $m$ if we assume the GEO-P model. In particular, if we know the order $n$, power law exponent $b$, average degree $d$, and diameter $D$ of an OSN, then we can calculate $m$ using our theoretical results. Formula (2) gives an estimate for $\alpha$ based on the power law exponent $b$. If $d^* = \log d / \log n$, then equation (3) implies that, asymptotically, $1 - \alpha - \beta = d^*$. If $D^* = \log D / \log n$, then formula (4) about the diameter implies that, asymptotically, $D^* = \frac{\beta}{(1-\alpha)m}$. Thus, an estimate for $m$ is given by:

$$m = \frac{1}{D^*}\left(1 - \left(\frac{b-1}{b-2}\right)d^*\right) = \frac{\log n}{\log D}\left(1 - \left(\frac{b-1}{b-2}\right)\frac{\log d}{\log n}\right). \tag{5}$$

This estimate suggests that the dimension is proportional to $\log n / \log D$. If $D$ is constant, then this means that $m$ grows logarithmically with $n$. Recall that the dimension of an OSN may be roughly defined as the least integer $m$ such that we can accurately embed the OSN in $m$-dimensional Euclidean space. Based on our model we conjecture that the dimension of an OSN is best fit by approximately $\log n$.

The parameters $b$, $d$, and $D$ have been determined for samples from OSNs in various studies such as [2, 19, 23, 32]. The following chart summarizes this data and gives the predicted dimension for each network. We round $m$ up to the nearest integer. Estimates of the total number of users $n$ for Cyworld, Flickr, and Twitter come from Wikipedia [40], and those from YouTube comes from their website [41]. When the data consisted of directed graphs, we took $b$ to be the power law exponent for the in-degree distribution. As noted in [2], the power law exponent of $b = 5$ for Cyworld holds only for users whose degree is at most approximately 100. When taking a sample, we assume that some of the neighbours of each node will be missing. Hence, when computing $d^*$, we used $n$ equalling the number of users in the sample. As we assume that the diameter of the OSN is constant, we compute $D^*$ with $n$ equalling the total number of users.

| **Parameter** | **OSN** | | | |
|---|---|---|---|---|
| | Cyworld | Flickr | Twitter | YouTube |
| $n$ | $2.4 \times 10^7$ | $3.2 \times 10^7$ | $7.5 \times 10^7$ | $3 \times 10^8$ |
| $b$ | 5 | 2.78 | 2.4 | 2.99 |
| $d^*$ | 0.22 | 0.17 | 0.17 | 0.1 |
| $D^*$ | 0.11 | 0.19 | 0.1 | 0.16 |
| $m$ | 7 | 4 | 5 | 6 |

### 3.5.1 The GEO-P Tension model

A variant of the GEO-P model was presented in [35] that warrants further study. In the GEO-P model, if a node $v$ falls in an influence region of of two nodes $u_1$ and $u_2$, then $v$ can join to $u_1$ and $u_2$ with equal probability. We consider a variant where the probability depends on the volume of the corresponding influence regions. Consider a fixed *tension parameter* $h \in (-\infty, 0)$. For a given $t \geq 0$ and vertex $u$, define

$$T(u,t) = r(u,t)^h.$$

Given two nodes $u$ and $v$, define

$$T(u,v,t) = \frac{T(u,t) + T(v,t)}{2}.$$

The definition of the GEO-P Tension model is analogous to the GEO-P model, but at time $t > 0$, independently, for each node $u \in V_{t-1}$ such that $v \in R(u,t-1)$, an edge $vu$ is created with probability $pT(u,v,t)$. Hence, if $v$ is in the influence region of both $u_1$ and $u_2$ with the rank of $u_1$ higher than $u_2$, then it is more likely to join to $u_1$.

Preliminary simulation results indicate that the GEO-P Tension model captures many of the properties of OSNs described in Section 2. Figures 2, 3 and 4 display the log-log plots of the degree distribution of graphs of order 7115 simulated by the GEO-P Tension model in dimensions 1 to 5 inclusive. We set the tension parameter $h = -0.1, -0.3$ and $-0.7$, respectively. Tables 1 and 2 list the diameters (where $|V(C)|$ is the order of the largest connected component) and spectral gaps (with respect to the adjacency matrix) of the corresponding graphs.

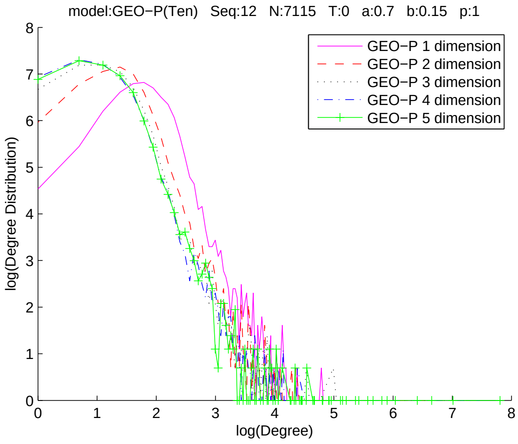

**Fig. 2** Degree distribution of graph generated by the GEO-P Tension model, $h = -0.1$.

For the GEO-P(Ten) model, it remains to rigorously prove the properties outline in Section 2.

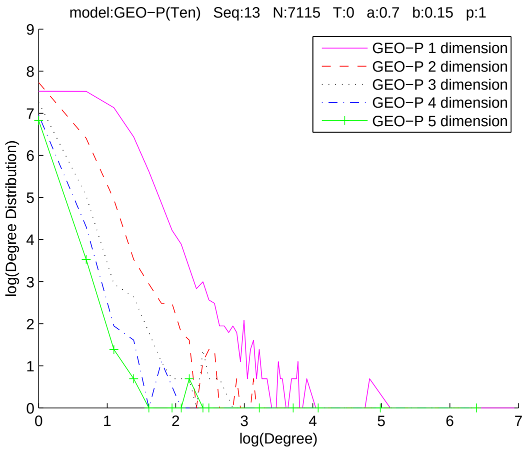

**Fig. 3** Degree distribution of graph generated by the GEO-P Tension model, $h = -0.3$.

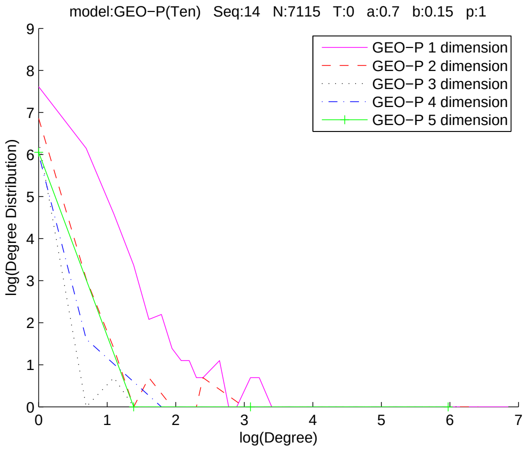

**Fig. 4** Degree distribution of graph generated by the GEO-P Tension model, $h = -0.7$.

**Table 1** Spectral gaps of graphs generated by the GEO-P(Ten) model, in dimensions 1 to 5 inclusive.

| **GEO-P(Ten)** $N = 7115$, $\alpha = 0.7$, $\beta = 0.15$, $p = 1$ | | | | | |
|---|---|---|---|---|---|
| $h = -0.1$ | | $h = -0.3$ | | $h = -0.7$ | |
| **dim** | **gap** | **dim** | **gap** | **dim** | **gap** |
| 1 | 14.170589 | 1 | 9.048678 | 1 | 13.048242 |
| 2 | 0 | 2 | 9.286437 | 2 | 12.406163 |
| 3 | 11.410871 | 3 | 10.354512 | 3 | 13.150765 |
| 4 | 15.54475 | 4 | 10.186001 | 4 | 14.226928 |
| 5 | 20.845548 | 5 | 12.422439 | 5 | 15.133812 |

**Table 2** Diameters of graphs generated by the GEO-P(Ten) model.

| **GEO-P(Ten)** $N = 7115$, $\alpha = 0.7$, $\beta = 0.15$, $p = 1$ | | | | | | |
|---|---|---|---|---|---|---|
| | $h = -0.1$ | | $h = -0.3$ | | $h = -0.7$ | |
| **dimension** | **diam** | $\lvert V(C) \rvert$ | **diam** | $\lvert V(C) \rvert$ | **diam** | $\lvert V(C) \rvert$ |
| 1 | 20 | 7098 | 6 | 6300 | 4 | 2673 |
| 2 | 20 | 6953 | 7 | 3149 | 4 | 990 |
| 3 | 20 | 6847 | 7 | 1648 | 3 | 571 |
| 4 | 15 | 6744 | 6 | 1152 | 3 | 415 |
| 5 | 12 | 6747 | 5 | 977 | 2 | 428 |

## 4 Open problems

Many questions remain in modelling OSNs and other complex networks. We collect these here for future reference.

1. The *Logarithmic Dimension Hypothesis* (or *LDH*) [7] conjectures that the dimension of an OSN is best fit by about $\log n$, where $n$ is the number of users in the OSN. The motivation for the conjecture comes from both the GEO-P and MAG models. Both models posit $\log n$ attributes for each user so as to provably ensure that certain properties found in OSNs (such as constant diameter and bad spectral expansion) are satisfied. Given the availability of OSN data, it may be possible to fit the data to the model to determine the dimension of a given OSN. Initial estimates in [35] from sampled OSN data indicate that the spectral gap found in OSNs correlates with the spectral gap found in the GEO-P model when the dimension is approximately $\log n$, giving some additional credence to the LDH. See also the MAG model as discussed in Subsection 3.4.
2. Another interesting direction would be to generalize the GEO-P to a wider array of ranking schemes (such as ranking by age or degree), and determine when similar properties (such as power laws and bad spectral expansion) provably *aas* hold. Simulations with the GEO-P Tension model show promising data [35], but the rich dependence structure of this model may make rigorous analysis a challenge.
3. As discussed in Section 2, the recent work [8] indicates that social networks lack high compressibility, especially in contrast to the web graph. Note that property (v) bad compressibility has not been explicitly studied in any of the models presented here. It would be interesting to study compressibility in these models, and to devise a model which provably has all five properties.
4. Anecdotal evidence from everyday experience with Twitter and Facebook shows that news and gossip spread quickly in such networks. An epidemiological model, such as SIS or SIRS, or even a deterministic model such as firefighting and seepage [6] would be worth exploring in real OSN data and in the models.

## 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. W. Aiello, A. Bonato, C. Cooper, J. Janssen, P. Prałat, A spatial web graph model with local influence regions, *Internet Mathematics* **5** (2009), 175–196.
4. A. Bonato, *A Course on the Web Graph*, American Mathematical Society Graduate Studies Series in Mathematics, Providence, Rhode Island, 2008.
5. A. Bonato, N. Hadi, P. Horn, P. Prałat, C. Wang, Models of on-line social networks, *Internet Mathematics* **6** (2011) 285-313.
6. A. Bonato, R.J. Nowakowski, *The Game of Cops and Robbers on Graphs*, American Mathematical Society, Providence, Rhode Island, 2011.
7. A. Bonato, J. Janssen, and P. Prałat, The geometric protean model for on-line social networks, In: *Proceedings of the 7th Workshop on Algorithms and Models for the Web-Graph (WAW2010)*, Lecture Notes in Computer Science 6516, Springer, 2010, 110–121.
8. F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, P. Raghavan, On compressing social networks, In: *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09)*, 2009.
9. F.R.K. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, Rhode Island, 1997.
10. F.R.K. Chung, L. Lu, *Complex Graphs and Networks*, American Mathematical Society, U.S.A., 2004.
11. E. Estrada, Spectral scaling and good expansion properties in complex networks, *Europhys. Lett.* **73** (2006) 649–655.
12. Facebook: statistics. Accessed September 1, 2011.
`http://www.facebook.com/press/info.php?statistics`.
13. A. Flaxman, A. Frieze, J. Vera, A geometric preferential attachment model of networks, *Internet Mathematics* **3** (2007) 187-205.
14. S. Fortunato, A. Flammini, F. Menczer, Scale-free network growth by ranking, *Phys. Rev. Lett.* **96** 218701 (2006).
15. O. Frank, Transitivity in stochastic graphs and digraphs, *Journal of Mathematical Sociology* **7** (1980) 199-213.
16. A. Henry, P. Prałat, Rank-Based Models of Network Structure and the Discovery of Content, In: *Proceedings of the 8th Workshop on Algorithms and Models for the Web Graph (WAW 2011)*, 2011.
17. S. Janson, T. Łuczak, A. Ruciński, *Random Graphs*, Wiley, NewYork, 2000.
18. J. Janssen, P. Prałat, Protean graphs with a variety of ranking schemes, *Theoretical Computer Science* **410** (2009), 5491–5504.
19. A. Java, X. Song, T. Finin, B. Tseng, Why we twitter: understanding microblogging usage and communities, In: *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*, 2007.
20. M. Kim, J. Leskovec, Multiplicative attribute graph model of real-world networks, In: *Proceedings of the 7th Workshop on Algorithms and Models for the Web Graph (WAW 2010)*, 2010.
21. J. Kleinberg, The small-world phenomenon: An algorithmic perspective, In: *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.
22. 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.
23. H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media?, In: *Proceedings of the 19th International World Wide Web Conference*, 2010.

24. S. Lattanzi, D. Sivakumar, Affiliation networks, In: *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, 2009.
25. 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.
26. 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.
27. J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, Z. Ghahramani, Kronecker Graphs: An approach to modeling networks, *Journal of Machine Learning Research* **11** (2010) 985-1042.
28. D.A. Levin, Y. Peres, E.L. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, 2009.
29. D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins, Geographic routing in social networks, In: *Proceedings of the National Academy of Sciences* **102** (2005) 11623–11628.
30. T. Łuczak, P. Prałat, Protean graphs, *Internet Mathematics* **3** (2006), 21–40.
31. M. Mahdian, Y. Xu, Stochastic Kronecker graphs, In: *Proceedings of the 5th Workshop on Algorithms and Models for the Web-Graph, 2007*
32. 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.
33. M.E.J. Newman, J. Park, Why social networks are different from other types of networks, *Phys. Rev. E* **68** 036122 (2003).
34. J.P. Scott, *Social Network Analysis: A Handbook*, Sage Publications Ltd, London, 2000.
35. Yanhua Tian, Models and Mining of On-line Social Networks, M.Sc. Thesis, Ryerson University, 2011.
36. Twitaholic. Accessed September 1, 2011. `http://twitaholic.com/`.
37. D.J. Watts, P.S. Dodds, M.E.J. Newman. Identity and search in social networks, *Science* **296** (2002) 1302–1305.
38. D.J. Watts, S.H. Strogatz, Collective dynamics of ‘small-world’ networks, *Nature* **393** (1998) 440–442.
39. 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.
40. Wikipedia: List of social networking websites. Accessed September 1, 2011. `http://en.wikipedia.org/wiki/List of social networking websites`.
41. YouTube, Advertising and Targeting. Accessed September 1, 2011. `http://www.youtube.com/t/advertising_targeting`.