The geometric protean model for on-line social networks

A. Bonato[1], J. Janssen[2], and P. Prałat[3]

[1] Department of Mathematics
Ryerson University
Toronto, Canada
abonato@ryerson.ca
[2] Department of Mathematics and Statistics
Dalhousie University
Halifax, Canada
janssen@mathstat.dal.ca
[3] Department of Mathematics
West Virginia University
Morgantown, USA
pralat@math.wvu.edu *

Abstract. We introduce a new geometric, rank-based model for the link structure of on-line social networks (OSNs). In the geo-protean (GEO-P) model for OSNs nodes are identified with points in Euclidean space, and edges are stochastically generated by a mixture of the relative distance of nodes and a ranking function. With high probability, the GEO-P model generates graphs satisfying many observed properties of OSNs, such as power law degree distributions, the small world property, densification power law, and bad spectral expansion. We introduce the dimension of an OSN based on our model, and examine this new parameter using actual OSN data.

1 Introduction

On-line social networking sites such as Facebook, Flickr, LinkedIn, MySpace, and Twitter are examples of large-scale, complex, real-world networks, with an estimated total number of users that equals half of all Internet users [2]. We may model an OSN by a graph with nodes representing users and edges corresponding to friendship links. While OSNs gain increasing popularity among the general public, there is a parallel increase in interest in the cataloguing and modelling of their structure, function, and evolution. OSNs supply a vast and historically unprecedented record of large-scale human social interactions over time.

The availability of large-scale social network data has led to numerous studies that revealed emergent topological properties of OSNs. For example, the recent study [15] crawled the entire Twitter site and obtained 41.7 million user profiles

* The authors gratefully acknowledge support from MITACS, NSERC, Ryerson, and WVU.

and 1.47 billion social relations. The next challenge is the design and rigorous analysis of models simulating these properties. Graph models were successful in simulating properties of other complex networks such as the web graph (see the books [4, 8] for surveys of such models), and it is thus natural to propose models for OSNs. Few rigorous models for OSNs have been posed and analyzed, and there is no universal consensus of which properties such models should simulate. Notable recent models are those of Kumar et al. [14], Lattanzi and Sivakumar [16], and the Iterated Local Transitivity model [5].

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. While by no means exhaustive, some of the main observed properties of OSNs include the following.

(i) *Large-scale.* OSNs are examples of complex networks with number nodes (which we write as n) often in the millions; further, some users have disproportionately high degrees. For example, each of the nodes of Twitter corresponding to celebrities Ashton Kutcher, Ellen Degeneres, and Britney Spears have degree over five million [23].

(ii) *Small world property and shrinking distances.* The small world property, introduced by Watts and Strogatz [25], is a central notion in the study of complex networks (see also [13]). 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 [21] 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. [12] and Kwak et al. [15]. Kumar et al. [14] 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. Various models (see [17, 18]) were proposed simulating power law degree distributions and decreasing distances.

(iii) *Power law degree distributions.* In a graph G of order n, let N_k 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 $b > 2$. 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 [14] 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 [21], as well as in Twitter [12, 15].

(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 (char-

acteristic of social organization), where tightly knit groups correspond to the clusters [22]. As a result, it is reported in [9] 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.

Our main contributions in the present work are twofold: to provide a model—the geo-protean (GEO-P) model—which provably satisfies all five properties above (see Section 3; note that the model does not generate graphs with shrinking distances, the parameters can be adjusted to give constant diameter), and second, to suggest a reverse engineering approach to OSNs. Given only the link structure of OSNs, we ask whether it is possible to infer the hidden reality of such networks. Can we group users with similar attributes from only the link structure? For instance, a reasonable assumption is that out of the millions of users on a typical OSN, if we could assign the users various attributes such as age, sex, religion, geography, and so on, then we should be able to identify individuals or at least small sets of users by their set of attributes. Thus, if we can infer a set of identifying attributes for each node from the link structure, then we can use this information to recognize communities and understand connections between users.

Characterizing users by a set of attributes leads naturally to a vector-based or geometric approach to OSNs. In geometric graph models, nodes are identified with points in a metric space, and edges are introduced by probabilistic rules that depend on the proximity of the nodes in the space. We envision OSNs as embedded in a *social space*, whose dimensions quantify user traits such as interests or geography; for instance, nodes representing users from the same city or in the same profession would likely be closer in social space. A first step in this direction was given in [19], which introduced a rank-based model in an m-dimensional grid for social networks (see also the notion of *social distance* provided in [24]). Such an approach was taken in geometric preferential attachment models of Flaxman et al. [10], and in the SPA geometric model for the web graph [3].

The geo-protean model incorporates a geometric view of OSNs, and also exploits ranking to determine the link structure. Higher ranked nodes are more likely to receive links. A formal description of the model is given in Section 2. Results on the model are summarized in Section 3. We present a novel approach to OSNs by assigning them a dimension; see the formula (4). Given certain OSN statistics (order, power law exponent, average degree, and diameter), we can assign each OSN a dimension based on our model. 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. Proofs of some of our results are presented in Section 4; the full version of the paper will contain proofs of all the results.

2 The GEO-P Model for OSNs

We now present our model for OSNs, which is based on 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 [11, 20].

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 $r(\cdot, t) : V_t \to [n]$ 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.) 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 $v \in V_{t-1} \cap V_t$,

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

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_∞ metric. In particular, for any two points x and y in $\mathbb{R}^m$,

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

The torus metric thus "wraps around" the boundaries of the unit cube, so every point in S is equivalent. The torus metric is chosen so that there are no boundary effects, and altering the metric will not significantly affect the main results.

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 equal to $p\, r(u, t-1)^{-\alpha} n^{-\beta}$. The negative exponent $(-\alpha)$ 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. 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. The random graph corresponding to this distribution with given parameters α, β, m, p is called the *geo-protean* (or *GEO-P* model) graph, and is written GEO-P(α, β, m, p).

3 Results and Dimension

We now state the main theoretical results we discovered for the geo-protean model, with proofs supplied in the next section. The model generates with high probability graphs satisfying each of the properties (i) to (iv) in the introduction. Proofs are presented in Section 4. Throughout, we will use the stronger notion of *wep* in favour of the more commonly used *aas*, since it simplifies some of our proofs. We say that an event holds *with extreme probability* (*wep*), if it holds with probability at least $1 - \exp(-\Theta(\log^2 n))$ as $n \to \infty$. Thus, if we consider a polynomial number of events that each holds *wep*, then *wep* all events hold.

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

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

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 1. *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 wep *GEO-P*(α, β, m, p) *satisfies*

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

For a graph $G = (V, E)$ of order n, define the *average degree of* G by $d = \frac{2|E|}{n}$. Our next results shows that geo-protean graphs are dense.

Theorem 2. Wep *the average degree of GEO-P*(α, β, m, p) *is*

$$d = (1 + o(1)) \frac{p}{1 - \alpha} n^{1-\alpha-\beta}. \tag{2}$$

Note that the average degree tends to infinity with n; that is, the model generates graphs satisfying a *densification power law*. In [17], densification power laws were reported in several real-world networks such as the physics citation graph and the internet graph at the level of autonomous systems.

Our next result describes the diameter of graphs sampled from the GEO-P model. While the diameter is not shrinking, it can be made constant by allowing the dimension to grow as a logarithmic function of n.

Theorem 3. *Let $\alpha \in (0,1)$, $\beta \in (0, 1-\alpha)$, $m \in \mathbb{N}$, and $p \in (0,1]$. Then* wep *the diameter of GEO-P(α, β, m, p) is*

$$O(n^{\frac{\beta}{(1-\alpha)m}} \log^{\frac{2\alpha}{(1-\alpha)m}} n). \tag{3}$$

We note that in a geometric model where regions of influence have constant volume and possessing the same average degree as the geo-protean model, the diameter is $\Theta(n^{\frac{\alpha+\beta}{m}})$. This is a larger diameter than in the GEO-P model. If $m = C \log n$, for some constant $C > 0$, then *wep* we obtain a diameter bounded by a constant. We conjecture that *wep* the diameter is of order $n^{\frac{\beta}{(1-\alpha)m}+o(1)}$. In the full version of the paper, we prove that *wep* the GEO-P model generates graph with constant clustering coefficient.

The normalized Laplacian of a graph relates to important graph properties; see [7]. Let A denote the adjacency matrix and D denote the diagonal degree matrix of a graph G. Then the normalized Laplacian of G is $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$. Let $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1} \leq 2$ denote the eigenvalues of $\mathcal{L}$. The *spectral gap* of the normalized Laplacian is

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

A spectral gap bounded away from 0 is an indication of bad expansion properties, which are characteristic of OSNs (see property (iv) in the introduction). The next theorem represents a drastic departure from the good expansion found in binomial random graphs, where $\lambda = o(1)$ [7, 8].

Theorem 4. *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(α, β, m, p). Then* wep

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*

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

3.1 Dimension of OSNs

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. The formulas (1) gives an estimate for α based on the power law exponent b. If $d^* = \log d / \log n$, then equation (2) implies that, asymptotically, $1 - \alpha - \beta = d^*$. If $D^* = \log D / \log n$, then (3) and our conjecture about the diameter implies that, asymptotically, $D^* = \frac{\beta}{(1-\alpha)m}$. Thus, an estimate for m is given by:

$$m = \frac{1}{D^*}\left(1 - \left(\frac{b-1}{b-2}\right)d^*\right). \tag{4}$$

Note that (4) suggests that the dimension depends on $\log n / \log D$. If D is constant, 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, 12, 15, 21]. 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 [26], and those from YouTube comes from their website [27]. 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	2.4×10^7	3.2×10^7	7.5×10^7	3×10^8
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

4 Proofs of results

We sketch the proofs of our results here, emphasizing those parts that give insight into the model. Detailed proofs of all our results will appear in a full paper.

4.1 Degree distribution; proof of Theorem 1

Theorem 1 follows immediately from the following theorem which shows how the degree of a given vertex depends precisely on its *age rank* and prestige label.

A vertex v has age rank $a(v,t) = i$ at time t if it is the i-th oldest vertex of all vertices existing at time t. The result below refers to the degree of a vertex at a time L, when the steady state of the GEO-P model has been reached.

The proof of the theorem follows standard methods, and is omitted here.

Theorem 5. *Let $i = i(n) \in [n]$. Let v_i be the vertex in GEO-P(α, β, m, p) whose age rank at time L equals $a(v_i, L) = i$, and let R_i be the initial rank of v_i.*

If $R_i \geq \sqrt{n}\log^2 n$, then wep

$$\deg(v_i, L) = (1 + O(\log^{-1/2} n))p\left(\frac{i}{(1-\alpha)n} + \left(\frac{R_i}{n}\right)^{-\alpha}\frac{n-i}{n}\right)n^{1-\alpha-\beta}.$$

Otherwise, that is if $R_i < \sqrt{n}\log^2 n$, wep

$$\deg(v_i, L) \geq (1 + O(\log^{-1/2} n))p\left(\frac{i}{(1-\alpha)n} + n^{\alpha/2}\log^{-2\alpha} n\frac{n-i}{n}\right)n^{1-\alpha-\beta}.$$

The proof of Theorem 1 is now a consequence of Theorem 5. One can show by an omitted calculation that *wep* each vertex v_i that has the initial rank $R_i \geq \sqrt{n}\log^2 n$ such that

$$\frac{R_i}{n} \geq \left(1 + \log^{-1/3} n\right)\left(pn^{1-\alpha-\beta}\frac{n-i}{n}k^{-1}\right)^{1/\alpha}$$

has fewer than k neighbours, and each vertex v_i for which

$$\frac{R_i}{n} \leq \left(1 - \log^{-1/3} n\right)\left(pn^{1-\alpha-\beta}\frac{n-i}{n}k^{-1}\right)^{1/\alpha}$$

has more than k neighbours.

Let i_0 be the largest value of i such that

$$\left(pn^{1-\alpha-\beta}\frac{n-i}{n}k^{-1}\right)^{1/\alpha} \geq \frac{2\log^2 n}{\sqrt{n}}.$$

This guarantees that the equations above do not contradict the requirement that $R_i \geq \log^2 n\sqrt{n}$. Note that $i_0 = n - O(n/\log n)$, since $k \leq n^{1-\alpha/2-\beta}\log^{-2\alpha-1} n$.

Using this result, we can compute the expected value of $N_{\geq k}$.

$$\begin{aligned}\mathbb{E}N_{\geq k} &= \sum_{i=1}^{i_0}\left(1 + O(\log^{-1/3} n)\right)\left(pn^{1-\alpha-\beta}\frac{n-i}{n}k^{-1}\right)^{1/\alpha} + O\left(\sum_{i=i_0+1}^{n}\frac{\log^2 n}{\sqrt{n}}\right)\\ &= \left(1 + O(\log^{-1/3} n)\right)\frac{\alpha}{\alpha+1}p^{1/\alpha}n^{(1-\beta)/\alpha}k^{-1/\alpha}.\end{aligned}$$

The concentration follows from the well-known Chernoff bound. □

4.2 Bad expansion: proof of Theorem 4

For the proof of Theorem 4 we show that there are sparse cuts in the GEO-P model. For sets X and Y we use the notation $e(X,Y)$ for the number of edges with one end in each of X and Y. Suppose that the unit hypercube $S = [0,1]^m$ is partitioned into two sets of the same volume,

$$S_1 = \{x = (x_1, x_2, \ldots, x_m) \in S : x_1 \leq 1/2\},$$

and $S_2 = S \setminus S_1$. Both S_1 and S_2 contain $(1+o(1))n/2$ vertices *wep*. In a good expander (for instance, the binomial random graph $G(n,p)$), *wep* there would be

$$(1+o(1))\frac{|E|}{2} = (1+o(1))\frac{p}{4(1-\alpha)}n^{2-\alpha-\beta}$$

edges between S_1 and S_2. Below we show that it is not the case in our model. The proof of the following theorem is omitted.

Theorem 6. *Let $\alpha \in (0,1)$, $\beta \in (0, 1-\alpha)$, $m \in \mathbb{N}$, and $p \in (0,1]$. Then* wep *GEO-P(α, β, m, p) has the following properties.*

1. *If $m = m(n) = o(\log n)$, then $e(S_1, S_2) = o(n^{2-\alpha-\beta})$.*
2. *If $m = m(n) = C \log n$ for some $C > 0$, then*

$$e(S_1, S_2) \leq (1+o(1))\frac{p}{4(1-\alpha)}n^{2-\alpha-\beta}\exp\left(-\frac{\alpha+\beta}{C}\right).$$

To finish the proof of Theorem 4, we use the expander mixing lemma for the normalized Laplacian (see [7] for its proof). For sets of nodes X and Y we use the notation $\mathrm{vol}(X)$ for the volume of the subgraph induced by X, $\bar{X}$ for the complement of X, and, as introduced before, $e(X,Y)$ for the number of edges with one end in each of X and Y. (Note that $X \cap Y$ does not have to be empty; in general, $e(X,Y)$ is defined to be the number of edges between $X \setminus Y$ to Y plus twice the number of edges that contain only vertices of $X \cap Y$.)

Lemma 1. *For all sets $X \subseteq G$,*

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

It follows from (2) and the Chernoff bound that *wep*

$$\begin{aligned} \mathrm{vol}(G_L) &= (1+o(1))\frac{p}{1-\alpha}n^{2-\alpha-\beta} \\ \mathrm{vol}(S_1) &= (1+o(1))\frac{p}{2(1-\alpha)}n^{2-\alpha-\beta} \quad = \quad (1+o(1))\mathrm{vol}(S_2). \end{aligned}$$

Suppose first that $m = o(\log n)$. From Theorem 6 we get that *wep*

$$\begin{aligned} e(S_1, S_1) = \mathrm{vol}(S_1) - e(S_1, S_2) \quad &= \quad (1+o(1))\mathrm{vol}(S_1) \\ &= (1+o(1))\frac{p}{2(1-\alpha)}n^{2-\alpha-\beta}, \end{aligned}$$

and Lemma 1 implies that *wep* $\lambda_n \geq 1 + o(1)$. By definition, $\lambda_n \leq 1$ so $\lambda_n = 1 + o(1)$.

Suppose now that $m = C \log n$ for some constant $C > 0$. By Theorem 6, we obtain that *wep*

$$e(S_1, S_1) = (1 + o(1)) \frac{p}{1-\alpha} n^{2-\alpha-\beta} \left(\frac{1}{2} - \frac{\exp\left(-\frac{\alpha+\beta}{C}\right)}{4} \right).$$

The assertion follows directly from Lemma 1. □

4.3 Diameter; proof of Theorem 3.

In order to show that the graph has a relatively small diameter, we will first show that *wep* there exists a "backbone" of vertices with a large influence region (which allow for long links), and that all vertices are within at most graph distance two from this backbone.

To find the backbone, fix A, and partition the hypercube into $1/A$ hypercubes. Fix R, and consider nodes with initial rank at most R and age at most $n/2$; we call these the *influential nodes*. We now choose A and R so that (i) in each small hypercube, *wep* there are $\log^2 n$ influential nodes, and (ii) the influence region of each influential node from its birth until the end of the process contains the whole hypercube in which it is located, and also all neighbouring hypercubes.

It can be shown that (ii) holds *wep* if the initial influence region of each influential node is at least $5^m A$. Therefore, we obtain that

$$R^{-\alpha} n^{-\beta} = 5^m A. \tag{5}$$

Property (i) holds if the expected number of influential nodes in each hypercube is at least $2 \log^2 n$ (Chernoff bound). Hence, we require that

$$\frac{n}{2} A \frac{R}{n} = 2 \log^2 n. \tag{6}$$

Combining (5) and (6) we obtain that the number of hypercubes is equal to

$$\frac{1}{A} = 5^{\frac{m+\alpha}{1-\alpha}} n^{\frac{\beta}{1-\alpha}} \log^{\frac{2\alpha}{1-\alpha}} n.$$

Now, since *wep* there are $\log^2 n$ nodes in each hypercube to choose from, *wep* we can select exactly one node from each hypercube so that each node is adjacent to the chosen nodes from all neighbouring hypercubes (the younger node falls into the region of influence of the older neighbours, and creates an edge with probability p). This subgraph then forms the backbone. It is clear that the diameter of the backbone is

$$\left(\frac{1}{A}\right)^{1/m} = O(n^{\frac{\beta}{(1-\alpha)m}} \log^{\frac{2\alpha}{(1-\alpha)m}} n)$$

We now show that *wep* a node v that is not in the backbone is distance at most two from some node in the backbone. Since *wep* the minimum degree is $\Omega(n^{1-\alpha-\beta})$, *wep* $\Omega(n^{1-\alpha-\beta})$ neighbours of v have age rank at least $n/2$. Since each such neighbour falls into the region of influence of some node in the backbone, *wep* at least one neighbour of v must be connected the backbone. □

5 Conclusion and Discussion

We introduced the geo-protean (GEO-P) geometric model for OSNs, and showed that with high probability, the model generates graphs satisfying each of the properties (i) to (iv) in the introduction. We introduce the dimension of an OSN based on our model, and examine this new parameter using actual OSN data. We observed that the dimension of various OSNs ranges from four to 7. It may therefore, be possible to group users via a relatively small number of attributes, although this remains unproven. The Logarithmic Dimension Hypothesis (or LDH) conjectures the dimension of an OSN is best fit by $\log n$, where n is the number of users in the OSN.

The ideas of using geometry and dimension to explore OSNs deserves to be more thoroughly investigated. 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 from actual 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 credence to the LDH. 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 hold.

We finish by mentioning that recent work [6] indicates that social networks lack high compressibility, especially in contrast to the web graph. We propose to study the relationship between the GEO-P model and the incompressibility of OSNs in future work.

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, accepted to *Internet Mathematics*, 2010.
6. 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.

7. F.R.K. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, Rhode Island, 1997.
8. F.R.K. Chung, L. Lu, *Complex Graphs and Networks*, American Mathematical Society, U.S.A., 2004.
9. E. Estrada, Spectral scaling and good expansion properties in complex networks, *Europhys. Lett.* **73** (2006) 649–655.
10. A. Flaxman, A. Frieze, J. Vera, A geometric preferential attachment model of networks, *Internet Mathematics* **3** (2007) 187-205.
11. J. Janssen, P. Prałat, Protean graphs with a variety of ranking schemes, *Theoretical Computer Science* **410** (2009), 5491–5504.
12. 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.
13. J. Kleinberg, The small-world phenomenon: An algorithmic perspective, In: *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.
14. 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.
15. 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.
16. S. Lattanzi, D. Sivakumar, Affiliation Networks, In: *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, 2009.
17. 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.
18. 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.
19. 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.
20. T. Łuczak, P. Prałat, Protean graphs, *Internet Mathematics* **3** (2006), 21–40.
21. 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.
22. M.E.J. Newman, J. Park, Why social networks are different from other types of networks, *Phys. Rev. E* **68**(3) 036122 (2003).
23. Twitterholic. Accessed September 12, 2010. `http://twitterholic.com/`
24. D.J. Watts, P.S. Dodds, M.E.J. Newman. Identity and search in social networks, *Science* **296** (2002) 1302–1305.
25. D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393** (1998) 440–442.
26. Wikipedia: List of social networking websites. Accessed September 12, 2010. `http://en.wikipedia.org/wiki/List of social networking websites`
27. YouTube, Advertising and Targeting. Accessed September 12, 2010. `http://www.youtube.com/t/advertising_targeting`