The Geometric Protean Model for On-Line Social Networks

Anthony Bonato, Jeannette Janssen, and Pawel Pralat

- Introduce a new geometric, rank-based model for the link structure of on-line social networks (OSNs)
- With high probability, the geo-protean (GEO-P) model generates graph satisfying many observed properties of OSNs

Large-scale

- Complex network with large number of nodes

- Some nodes have very high degree.

Small world property and shrinking distance

- 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 average degree

Power law degree distribution

- 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.

Bad spectral expansion

- Social network often organize into separate clusters in which the intra-cluster links are significantly higher than the number of inter-cluster links

The GEO-P Model for OSNs

- Produces a sequence (G_t: $t >= 0$) of undirected graph on n nodes, where t denotes time. $G_t=(V_t, E_t)$.

- 4 parameters: the attachment strength $\alpha \in (0,1)$, the density parameter $\beta \in (0, 1-\alpha)$, the dimension $m \in N$, and the link probability $p \in (0,1]$

-Each nodes $v \in V_t$ has rank $r(v,t) \in [n]$

-The rank function $r(\cdot, t): V_t \rightarrow [n]$ is a bijection for allt , so every node has a unique rank.

-Highest ranked node has rank equal to 1, the lowest ranked node has rank n.

-For each $v \in V_{t-1} \quad V_t$, $r(v,t) = r(v, t-1) + \delta - \gamma$

where $\delta = 1$ if $r(v, t-1) > Rt$ 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 Rm, with the torus metric $d(\cdot,\cdot)$ derived from the $L\infty$ metric. In particular, for any two points x and y in Rm,

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

-To initialize the model, let G0 = (V0, E0) 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 $r(\cdot, t) : V_t \to [n]$.

The following theorem demonstrates that the geo-protean model generates power law graphs with exponent

$$b = 1+1/\alpha$$

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*

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

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*

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

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*

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

Dimension of OSNs

-Given an OSN, 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.

-An estimate for m is given by:

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

The following chart gives the predicted dimension m for each network. We round m up to the nearest integer

n is total number of users for Cyworld, Flickr,Twitter, and YouTube

b is the power law exponent for the in-degree distribution

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