Internet Mathematics Vol. 8, Nos. 1–2: 2–28

Geometric Protean Graphs

Anthony Bonato, Jeannette Janssen, and Paweł Prałat

Abstract. We study the link structure of online social networks (OSNs) and introduce a new model for such networks that may help in inferring their hidden underlying reality. 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. We discuss how the geo-protean model may eventually be used as a tool to group users with similar attributes using only the link structure of the network.

1. Introduction

Online 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 at least half of all Internet users [Ahn et al. 07]. We may model an OSN by a graph with nodes representing users and edges corresponding to friendship links. As OSNs gain increasing popularity among the general public, there is a parallel increase in interest in the cataloguing and modeling of their structure, function, and evolution. OSNs supply a vast and

ISSN: 1542-7951 print

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 have revealed emergent topological properties of OSNs. For example, the recent study [Kwak et al. 10] crawled the entire Twitter site and studied properties found among the 41.7 million user profiles 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 [Bonato 08, Chung 04] 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 as to which properties such models should simulate. Notable recent models are those of [Kumar et al. 06], [Lattanzi and Sivakumar 09], and the iterated local transitivity model of [Bonato et al. 11].

Researchers are now in the enviable position of observing how OSNs evolve over time, and therefore, 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:

1. *Large scale.* OSNs are examples of complex networks with numbers of nodes (which we write as n) often in the millions; further, some users have disproportionately high degrees. For example, some of the nodes of Twitter corresponding to well-known celebrities have degree over five million.

2. *Small-world property and shrinking distances.*The small-world property, introduced in [Watts and Strogatz 98], is a central notion in the study of complex networks (see also [Kleinberg 00]). 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 the same average degree. An early study of an OSN at Stanford University appeared in [Adamic et al. 03], in which the authors found that the network has the small-world property. Similar results were found in [Ahn et al. 07], which studied Cyworld, MySpace, and Orkut, and in [Mislove et al. 07], which examined data collected from Flickr, YouTube, LiveJournal, and Orkut. Low diameter (of 6) and high clustering coefficient were reported in Twitter in [Java et al. 07] and [Kwak et al. 10]. In [Kumar et al. 06], the authors reported that in Flickr and Yahoo!360, the diameter actually decreases over time. Similar results were reported for Cyworld in [Ahn et al. 07]. Well-known models for complex networks such as preferential attachment and copying models have diameters that grow logarithmically with time. Various

models (see [Leskovec et al. 05a, Leskovec et al. 05b]) have been proposed that simulate power-law degree distributions and decreasing distances.

3. *Densification power law.* 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 [Leskovec et al. 05a], 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. Densification was reported in Cyworld [Ahn et al. 07] and has been detected in other OSNs.

4. *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 $b > 2$. Power laws were observed over a decade ago in subgraphs sampled from the web graph, and are ubiquitous properties of complex networks [Bonato 08, Chapter 2]. In [Kumar et al. 06], the authors 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 [Mislove et al. 07], as well as in Twitter [Java et al. 07, Kwak et al. 10].

5. *Bad spectral expansion.* Social networks often organize into separate clusters in which the number of intracluster links is significantly higher than the number of intercluster links. In particular, social networks contain communities (characteristic of social organization), where tightly knit groups correspond to the clusters [Newman and Park 03]. As a result, it is reported in [Estrada 06] 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—that provably satisfies all six properties above (see Section 3; note that while 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 sets 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 and 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 [Liben-Nowell et al. 05], which introduced a rank-based model in an m-dimensional grid for social networks (see also the notion of *social distance* provided in [Watts et al. 02]). Such an approach was taken in geometric preferential-attachment models of [Flaxman et al. 07] and in the SPA geometric model for the web graph [Aiello et al. 09].

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 (3.4). 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. Full proofs of our results are presented in Section 4. In the final discussion, we summarize our findings and conjecture the correct diameter for OSNs.

2. The GEO-P Model for OSNs

We now present our model for OSNs, which is based on the notion of embedding the nodes in a metric space (geometric) and on 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 other's 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, one node is born and one dies. A static number of nodes is more representative of the reality of OSNs, since the number of users in an OSN will 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 [Fortunato et al. 06, Henry and Prałat 11, Janssen and Prałat 09, Łuczak and Prałat 06].

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. 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_∞ 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 thus "wraps around" the boundaries of the unit cube, so all points in S are 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 proportional to $pr(u, t-1)^{-\alpha}$. The negative exponent guarantees that nodes with higher ranks ($r(u, t-1)$ close to 1) are more likely to receive new edges than nodes with lower ranks.
2. Choose uniformly at random a node $u \in V_{t-1}$, and delete u and all edges incident to u.
3. Vertex 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 [Levin et al. 09] for more on Markov chains.) The random graph corresponding to this distribution with given parameters α, β, m, p is called the *geo-protean* graph (or *GEO-P* model), and is written GEO-P(α, β, m, p). The coupon-collector problem can give us insight into when the stationary state will be reached. Namely, let $L = n(\log n + O(\omega(n)))$, where $\omega(n)$ is any function tending to infinity with n. It is a well-known result that with probability tending to 1 as n tends to infinity, after L steps all original vertices will be deleted. See Figure 1 for a simulation of the model in the unit square.

3. Results and Dimension

3.1. Results

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 1 to 4 that we discussed in the introduction. Proofs are presented in Section 4. Throughout, we will use the stronger notion of *wep* in favor 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

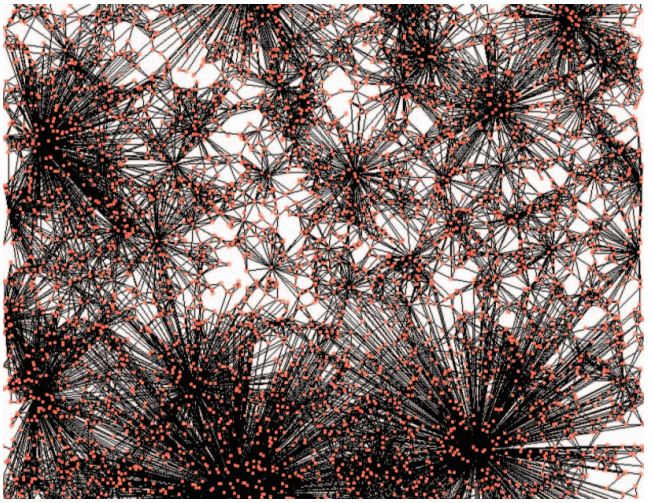

Figure 1. A simulation of the GEO-P model, with $n = 5000$, $\alpha = 0.7$, $\beta = 0.15$, $m = 2$, and $p = 0.9$ (color figure available online).

generates power-law graphs with exponent

$$b = 1 + 1/\alpha. \tag{3.1}$$

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 3.1. *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*

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

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

Theorem 3.2. *The average degree of* GEO-P(α, β, m, p) *is wep*

$$d = (1 + o(1)) \frac{p}{1-\alpha} n^{1-\alpha-\beta}. \tag{3.2}$$

Note that the average degree tends to infinity with n; that is, the model generates graphs satisfying a *densification power law*. In [Leskovec et al. 05a], 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 3.3. *Let $\alpha \in (0,1)$, $\beta \in (0, 1-\alpha)$, $m \in \mathbb{N}$, and $p \in (0,1]$. Then wep the diameter D of* GEO-P(α, β, m, p) *satisfies*

$$D = \Omega\left(n^{\frac{\beta}{(1-\alpha)m}} \log^{\frac{-\alpha}{m}} n\right) \text{ and } D = O\left(n^{\frac{\beta}{(1-\alpha)m}} \log^{\frac{2\alpha}{(1-\alpha)m}} n\right). \tag{3.3}$$

In particular, wep the order of the diameter can be expressed as

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

We note that in a geometric model where regions of influence have constant volume and possess 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 above by a constant.

Let $G = (V, E)$ be a graph. For sets of vertices $X, Y \subseteq V$, define $e(X, Y)$ to be the set of edges with one endpoint in X and the other in Y. For simplicity, we write $e(X) = e(X, X)$. Let $N(v)$ be the neighbor set of the vertex v. The *clustering coefficient* of vertex $v \in V$ is defined as follows:

$$c(v) = \frac{e(N(v))}{\binom{\deg(v)}{2}}.$$

(Note that formally, we need to assume that $\deg(v) \geq 2$ above, but this is *aas* the case in our model. One can define $c(v) = 0$ when $\deg(v) \leq 1$.) In other words, $c(v) \in [0,1]$ is the probability that two different neighbors of v, chosen uniformly at random, are adjacent. In the random graph $G(n,p)$, the expected value of $c(v)$ is p for any vertex v. The *clustering coefficient* of G is defined as

$$c(G) = \frac{1}{|V|} \sum_{v \in V} c(v).$$

Hence $\mathbb{E}(c(G(n,p))) = p$. We prove that for some values of m, *wep* the GEO-P model generates graphs with higher clustering coefficient than in a random graph

$G(n, d/n)$ with the same expected average degree. It follows from Theorem 3.2 that

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

We use the notation $\lfloor x \rfloor_2$ to denote the largest *even* integer less than or equal to x.

Theorem 3.4. *The clustering coefficient of G sampled from* GEO-P(α, β, m, p) *satisfies wep 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(α, β, m, p) is *wep* 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.

If $m = o(\log n)$ but large enough that the condition $K \gg m$ does not hold, a similar result holds, but the error term is no longer $(1 + o(1))$. In this case, *wep*

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

Finally, if

$$m = (1 + o(1))b(1-\alpha-\beta)\log n$$

for some constant $b \in (0, 1)$, then $K \geq \lfloor e^{1/b} \rfloor_2 \geq e^{1/b} - 2$, and so *wep*

$$\begin{aligned} c(G) &\geq \left(\frac{3}{4}\left(1 - \frac{2}{3(e^{1/b} - 2)}\right)(1 + o(1))\right)^m \\ &= \exp\left(b(1-\alpha-\beta)\log\left(\frac{3}{4} - \frac{1}{2(e^{1/b} - 2)}\right)(1 + o(1))\log n\right). \end{aligned}$$

For the random graph counterpart we have that *wep*

$$c(G(n,d/n)) = (1+o(1))\frac{p}{1-\alpha}n^{-\alpha-\beta} = \exp\Big((-\alpha-\beta)(1+o(1))\log n\Big).$$

Thus, we get a larger clustering coefficient for b small enough, that is, for $b < b_0$, where $b_0 = b_0(\alpha,\beta)$ satisfies the following equation:

$$b(1-\alpha-\beta)\log\left(\frac{3}{4}-\frac{1}{2(e^{1/b}-2)}\right) = -\alpha-\beta.$$

(Note that the function on the left-hand side is increasing and tends to 0 as $b \to 0$.) For $b > b_0$ we get the opposite behavior; that is, the clustering coefficient is smaller compared to the binomial random graph counterpart.

The normalized Laplacian of a graph relates to important graph properties; see [Chung 97]. Let A denote the adjacency matrix and D 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 $0 = \lambda_0 \le \lambda_1 \le \cdots \le \lambda_{n-1} \le 2$ 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 zero is an indication of bad expansion properties. Bad expansion is characteristic of OSNs: see property 4 in the introduction. The next theorem represents a drastic departure from the good expansion found in binomial random graphs, where $\lambda = o(1)$ [Chung 97, Chung 04].

Theorem 3.5. *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*

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

3.2. 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. Formula (3.1) gives an estimate for α based on the power-law exponent b. If $d^* = \log d/\log n$, then equation (3.2) implies that asymptotically, $1-\alpha-\beta = d^*$. If $D^* = \log D/\log n$, then formula

	OSN			
Parameter	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

Table 1. Estimated dimensions of various OSNs.

(3.3) about the diameter implies that asymptotically, $D^* = \frac{\beta}{(1-\alpha)m}$. Thus, an estimate for m is given by

$$m = \frac{1}{D^*}\left(1 - \left(\frac{b-1}{b-2}\right)d^*\right) = \frac{\log n}{\log D}\left(1 - \left(\frac{b-1}{b-2}\right)\frac{\log d}{\log n}\right). \tag{3.4}$$

This estimate suggests that the dimension is proportional to $\log n / \log D$. 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 [Ahn et al. 07, Java et al. 07, Kwak et al. 10, Mislove et al. 07]. Table 1 summarizes these 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,[1] and those from YouTube come from their website.[2] When the data consisted of directed graphs, we took b to be the power-law exponent for the in-degree distribution. As noted in [Ahn et al. 07], 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 neighbors of each node will be missing. Hence, in computing d^*, we used n equal to the number of users in the sample. Since we assume that the diameter of the OSN is constant, we compute D^* with n the total number of users.

[1] Wikipedia: "List of Social Networking Websites." Accessed April 1, 2011 (http://en.wikipedia.org/wiki/List_of_social_networking_websites).

[2] YouTube, "Advertising and Targeting." Accessed April 1, 2011 (www.youtube.com/t/advertising_targeting).

4. Proofs of Results

We will make frequent use of the following standard result about the sum of independent random variables, known as the *Chernoff bound*; for a proof, see [Janson et al. 00, Theorem 2.8].

Theorem 4.1. *Let X be a random variable that can be expressed as a sum $X = \sum_{i=1}^{n} X_i$ of independent random indicator variables, where $X_i \in \mathrm{Be}(p_i)$ with (possibly) different $p_i = \mathbb{P}(X_i = 1) = \mathbb{E}X_i$. Then the following holds for $t \geq 0$:*

$$\mathbb{P}(X \geq \mathbb{E}X + t) \leq \exp\left(-\frac{t^2}{2(\mathbb{E}X + t/3)}\right),$$
$$\mathbb{P}(X \leq \mathbb{E}X - t) \leq \exp\left(-\frac{t^2}{2\mathbb{E}X}\right).$$

In particular, if $\varepsilon \leq 3/2$, then

$$\mathbb{P}(|X - \mathbb{E}X| \geq \varepsilon \mathbb{E}X) \leq 2\exp\left(-\frac{\varepsilon^2 \mathbb{E}X}{3}\right).$$

Moreover, if $\mathbb{E}X \leq \log^2 n$, then wep $X = O(\log^2 n)$.

Before we prove the theorems discussed in the previous sections, we first give a lemma that shows that if the initial rank is large enough, the rank of a vertex maintains a value close to its initial value until its death. The first lemma is proved in [Janssen and Prałat 09].

Lemma 4.2. [Janssen and Prałat 09] *Suppose that vertex v obtained an initial rank $R \geq \sqrt{n}\log^2 n$ in the ranking by the random initial rank scheme. Then wep*

$$r(v,t) = R(1 + O(\log^{-1/2} n))$$

to the end of its life.

As we will see, the degree of a vertex depends on its rank and its age. The measure used to quantify the age of a vertex is its *age rank*. The age rank $a(v,t)$ of vertex v at time t is a number between 1 and n that represents the rank of v if all vertices alive at time t are ranked according to age, oldest first. So $a(v,t) = 1$ means that v is the oldest vertex alive at time t, while $a(v,t) = n$ implies that v is the youngest.

Lemma 4.3. *Suppose that vertex v obtained an initial rank $R \geq \log^3 n$. Moreover, v has age rank $a(v, L) \geq Cn$ for some constant $C \in (0, 1)$. Then wep*

$$r(v, t) \leq R \cdot 3^{2\log(1/C)} = O(R)$$

during the whole process up to time L.

Proof. If follows from [Janssen and Prałat 09, Theorem 5.5] that *wep* the age rank of a vertex v after

$$t \leq n \log n/2 - 2n \log \log n$$

steps is $(1 + o(1))n \exp(-t/n)$. Since v has age rank at least Cn at time L, this implies that $C \leq (1 + o(1)) \exp(-t_v/n)$, where t_v is the age of v at time L (so v was born at time $L - t_v$). Thus, $t_v \leq (1 + o(1))n \log(1/C)$. Recall that the rank of v at the time it is born is R; we wish to find an upper bound on $r(v, L)$ by considering how the rank can change in t_v steps of the process. Consider the following random variable X_t: $X_0 = R$, $X_{t+1} = X_t + 1$, with probability X_t/n, with $X_{t+1} = X_t$ otherwise. The variable X_t is an upper bound on the rank of v after t steps. Note that this upper bound considers only those changes to the rank due to a vertex of higher rank being inserted, not the changes in rank due to vertices of lower rank being deleted.

We introduce the stopping time

$$T = \min\{t \geq 0 : X_t > 3R \text{ or } t > n/2\}.$$

For any time $t \leq T$, X_t increases with probability at most $3R/n$. After $n/2$ steps, our random variable increases by at most

$$(1 + o(1)) \left(\frac{3R}{n}\right)(n) = (1 + o(1)) \left(\frac{3R}{2}\right)$$

wep, provided that $n/2 \leq T$. Hence *wep* $T \geq n/2$. Thus *wep* in the first $n/2$ steps after its birth, the rank of vertex v can increase by at most a factor of 3. Dividing the total life span of v into at most $2\log(1/C)$ blocks of $n/2$ time steps each, we obtain the result. □

4.1. Proof of Theorems 3.1 and 3.2

Before we prove Theorems 3.1 and 3.2, we need one more result. The following theorem shows how the degree of a given vertex depends on its age rank and its initial rank.

Theorem 4.4. *Let $\alpha \in (0, 1)$, $\beta \in (0, 1 - \alpha)$, $m \in \mathbb{N}$, $p \in (0, 1]$, $i = i(n) \in [n]$. 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*

$$\deg(v_i, L) = (1 + O(\log^{-1/2} n))p\left(\frac{i}{(1-\alpha)n} + \left(\frac{R_i}{n}\right)^{-\alpha}\frac{n-i}{n}\right)n^{1-\alpha-\beta}. \quad (4.1)$$

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

$$\deg(v_i, L) \geq (1 + O(\log^{-1/2} n))p\left(\frac{i}{(1-\alpha)n} + (n^{\alpha/2}\log^{-2\alpha} n)\frac{n-i}{n}\right)n^{1-\alpha-\beta}. \quad (4.2)$$

Proof. Let $\deg^+(v_i, L)$ denote the number of neighbors of v_i with age rank smaller than i, and define

$$\deg^-(v_i, L) = \deg(v_i, L) - \deg^+(v_i, L).$$

Let us focus on these random variables independently.

Since vertices are distributed uniformly at random, the expected initial degree of v_i (at the time v_i was born) is

$$\sum_{r=1}^{n} pr^{-\alpha}n^{-\beta} = pn^{-\beta}\int_1^n x^{-\alpha}dx + O(1) = \frac{p}{1-\alpha}n^{1-\alpha-\beta} + O(1).$$

From the $n-1$ vertices that were older than v_i at the time it was born, only $i-1$ remain. Since vertices are deleted uniformly at random, the expected number of older neighbors remaining is

$$\mathbb{E}\deg^+(v_i, L) = \frac{p}{1-\alpha}\frac{i}{n}n^{1-\alpha-\beta} + O(1).$$

Since $\deg^+(v_i, L)$ can be expressed as a sum of independent random variables, it follows from Theorem 4.1 that *wep* this random variable is well concentrated around its expectation, provided that $\mathbb{E}\deg^+(v_i, L) = \Omega(\log^3 n)$. This condition holds if, for example, $i > n^{\alpha+\beta}\log^3 n$. In this case we have that $\deg^+(v_i, L) = (1 + O(\log^{-1/2} n))\mathbb{E}\deg^+(v_i, L)$.

Next we consider the contribution to the degree of v_i of vertices that are younger than v_i. Suppose first that $R_i \geq \sqrt{n}\log^2 n$. It follows from Lemma 4.2 that *wep* $r(v_i, t) = R_i(1 + O(\log^{-1/2} n))$ to the end of its life. Therefore,

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

Similarly as before, this is well concentrated around its expectation. More precisely, *wep*

$$\deg^-(v_i, L) = (1 + O(\log^{-1/2} n))\mathbb{E}\deg^-(v_i, L),$$

provided that $\mathbb{E}\deg^-(v_i, L) = \Omega(\log^3 n)$. Note that it is sufficient to have $i < n - n^{\alpha+\beta}\log^3 n$, even if $R_i = n$.

Let us mention that if one of $d^+(v_i, L)$, $d^-(v_i, L)$ is $O(\log^3 n)$ in expectation, then the other is expected to be $\Omega(n^{1-\alpha-\beta})$ and *wep* is concentrated around its expectation. This implies that *wep* their sum $\deg(v_i, L)$ is always concentrated for $R_i \geq \sqrt{n}\log^2 n$. Combining the number of older and younger neighbors, we obtain that *wep* (4.1) holds.

Finally, if $R_i < \sqrt{n}\log^2 n$, then *wep* $r(v_i, t) \leq \sqrt{n}\log^2 n(1 + O(\log^{-1/2} n))$ to the end of its life. The argument for $\deg^+(v_i)$ is analogous and so is omitted, but we obtain only a lower bound for $\deg^-(v_i)$. Thus we find that *wep* (4.2) holds. □

It follows from Theorem 4.4 that *wep* the minimum degree is

$$(1 + o(1))pn^{1-\alpha-\beta}.$$

Vertices of minimum degree are old and of low rank: they have age rank $i = o(n)$ and thus have lost most of their initial links, and since their initial rank was $R_i = n - o(n)$, they did not acquire many new links.

The maximum degree changes during the process, and this behavior is impossible to predict. However, in order to get an upper bound, suppose the extreme case in which the oldest vertex is ranked number one during its entire life. In such an extreme case, the degree would be $(1 + o(1))pn^{1-\beta}$ *wep*, which indicates that *wep* the maximum degree is at most $(1 + o(1))pn^{1-\beta}$.

We now turn to the average degree.

Proof of Theorem 3.2. By Theorem 4.4, the average degree is

$$\begin{aligned}\frac{2|E|}{n} &= \frac{2}{n}\sum_{i=1}^{n}\deg^+(v_i, L)\\ &= (1 + o(1))\ \frac{2}{n}\ \sum_{i=1}^{n}\frac{p}{1-\alpha}\ \frac{i-1}{n-1}n^{1-\alpha-\beta}\\ &= (1 + o(1))\frac{p}{1-\alpha}n^{1-\alpha-\beta}. \qquad (4.3)\end{aligned}$$

□

The proof of Theorem 3.1 is now a simple consequence of Theorem 4.4. Let k be such that

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

One can show that *wep* each vertex v_i that has 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 neighbors, 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 neighbors.

Let i_0 be the largest value of i such that

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

(Note that $i_0 = n - O(n/\log n)$, since $k \leq n^{1-\alpha/2-\beta}\log^{-2\alpha-1} n$.) Thus

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

and the assertion follows from the Chernoff bound, since $k \leq n^{1-\alpha/2-\beta}\log^{-2\alpha-1} n$, and so $\mathbb{E}Z_{\geq k} = \Omega(\sqrt{n}\log^{2+1/\alpha} n)$.

4.2. Proof of Theorem 3.3

We consider the upper and lower bounds in separate arguments.

Upper Bound. The idea of the proof is to show that we can construct a connected subgraph of vertices spread out over the hypercube with a small diameter. This subgraph will act as a *backbone*, and the next step in the proof shows that each vertex is connected to the backbone by a path of length at most 2.

We will fix values $A \in (0,1)$ and $R \in [n]$ to suit our needs later. To construct the backbone we partition the hypercube into $1/A$ subcubes, each of volume A. Consider all vertices with initial rank at most R and age rank between $n/4$ and $n/2$; we will call such vertices *eminent* vertices. We now choose A and R such that *wep* (1) the influence region of each eminent vertex contains the subcube

in which it is located as well as all neighboring subcubes, and (2) each subcube contains at least $\log^2 n$ eminent vertices.

Property (2) will be achieved if the sphere of influence of each eminent vertex has size at least $4^m A$ throughout the process. Note that the sphere of influence of an eminent vertex initially has volume at least $R^{-\alpha} n^{-\beta}$, and by Lemmas 4.2 and 4.3, *wep* it remains at least $3^{-2\alpha \log 4} R^{-\alpha} n^{-\beta} > (R/25)^{-\alpha} n^{-\beta}$ to the end of the process. We can thus achieve (1) by choosing R and A such that the initial influence region is sufficiently larger than $4^m A$: we choose $5^m A$. This leads to our first condition on R and A:

$$(R/25)^{-\alpha} n^{-\beta} = 5^m A. \tag{4.4}$$

It follows from the Chernoff bound that to guarantee that *wep* every subcube contains at least $\log^2 n$ eminent vertices, it is sufficient to choose A and R so that the expected number of eminent vertices in a subcube is at least $\frac{5}{2} \log^2 n$. Since the initial rank is independent of age, the expected number of eminent vertices in a subcube equals $(n/4)(R/n)A = (R/4)A$. Thus the following condition on A and R guarantees (2):

$$A \frac{R}{4} = \frac{5}{2} \log^2 n. \tag{4.5}$$

Combining (4.4) and (4.5), we obtain the following values for R and A:

$$R = (10 \cdot 25^{-\alpha} \cdot 5^m n^{\beta} \log^2 n)^{1/(1-\alpha)},$$
$$\frac{1}{A} = \frac{(10 \cdot 25^{-\alpha})^{1/(1-\alpha)}}{10} 5^{\frac{m}{1-\alpha}} n^{\frac{\beta}{1-\alpha}} \log^{\frac{2\alpha}{1-\alpha}} n.$$

The set of all eminent vertices forms the *backbone*. Since vertices in each subcube induce a random graph with a parameter p (constant, bounded away from zero), *wep* the induced subgraph is connected and of diameter two. For any two neighboring subcubes, the probability that there is no edge between them is equal to

$$(1-p)^{(\log_2^2 n)} = \exp(-\Omega(\log^4 n)),$$

so *wep* there is at least one edge connecting two neighboring subcubes. Since the largest metric distance in the hypercube with torus metric equals $1/2$, and the subcubes have diameter $2A^{1/m}$, the diameter of the backbone is *wep*

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

To finish the proof, we will show that *wep* any vertex v that is not in the backbone is within distance two from some vertex in the backbone. Consider any vertex v not part of the backbone. Consider S_v, the ball of volume $n^{-\alpha-\beta}$ centered at v. The volume of S_v is the minimum volume of a sphere of influence,

so for each vertex w in S_v, edge vw exists with probability p. Note also that every vertex of age rank greater than $n/2$ in any subcube links to the backbone vertex in the same subcube with probability p, since the sphere of influence of the backbone vertex includes all of the subcube. There are $(1+o(1))n^{1-\alpha-\beta}/2$ vertices of age rank greater than $n/2$ in S_v, and a path of length 2 from v to the backbone using that vertex exists with probability p^2. Thus *wep* such a path exists.

Lower Bound. Note that for $m = \Theta(\log n)$, the theorem states that $D \geq 1$, and the lower bound trivially holds. We can assume then that $m = o(\log n)$. Let $R = n^{\frac{\beta}{1-\alpha}} \log^{-1} n$. Note that at every point of the process, the union of influence regions of vertices with rank at most R has volume at most

$$\sum_{r=1}^{R} r^{-\alpha} n^{-\beta} = \Theta(R^{1-\alpha} n^{-\beta}) = o(1).$$

These vertices can generate *long* edges. We will call such vertices *hubs*. Other edges are *short*. The length of every short edge is *wep* at most

$$(1+o(1))(R^{-\alpha} n^{-\beta})^{1/m} = (1+o(1)) n^{\frac{-\beta}{(1-\alpha)m}} \log^{\frac{\alpha}{m}} n.$$

(This time, the length corresponds to the torus metric, not the graph distance.) Since the total volume of the influence regions of the hubs is $o(1)$, there must be a vertex v and a constant $c \in (0, 1/3)$ such that *wep* the ball around v with radius c does not intersect any hub of an influence region. Moreover, since vertices are uniformly distributed, *wep* there is a vertex u at (metric) distance greater than c from v. Any path from v to u must use short edges to bridge the distance from v to the edge of the circle with radius c. This implies that *wep* the path has (graph distance) length at least $(1+o(1))n^{\frac{\beta}{(1-\alpha)m}} \log^{\frac{-\alpha}{m}} n$, which finishes the proof.

4.3. Proof of Theorem 3.4

Consider G sampled from GEO-P(α, β, m, p), and define $V_{\min} = n^{-\alpha-\beta}$. Then in GEO-P$(\alpha, \beta, m, p)$, $V_{\min}$ corresponds to the lower bound for a volume of the influence region that is obtained for a vertex with rank n. Thus, if the distance between u and v is at most $r_{\min} = V_{\min}^{1/m}/2$, two vertices u and v are adjacent with probability p. (Of course, edges can also be created between vertices that are a larger distance apart, but this requires additional conditions on the initial rank of these vertices.) Recall that due to our choice of metric, the ball of radius $r_{\min}$ centered at v is a hypercube. We will call this the *minimal hypercube* of v. The minimal region of a vertex is always a subset of its region of influence.

Fix $v \in V(G)$. Without loss of generality, we can assume that v is the origin of the hypercube S; that is, $v = (0, 0, \ldots, 0) \in S$. In order to estimate $c(v)$ from below, we partition the minimal hypercube of v into K^m disjoint identical small hypercubes of volume $\log^3 n/n$ each. The expected number of neighbors of v in every ball is $p \log^3 n$, so *wep* the number of neighbors equals $(1 + O(\log^{-1/2} n))p \log^3 n$. Note that

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

Recall that $\lfloor x \rfloor_2$ is the largest even integer less than or equal to x.

We index the balls as follows:

$$b_{i_1, i_2, \ldots, i_m} = (a_{i_1 - 1}, a_i) \times \cdots \times (a_{i_m - 1}, a_i),$$

where

$$a_i = -\frac{V_{\min}^{1/m}}{2} + i \left(\frac{\log^3 n}{n} \right)^{1/m}.$$

For all $s \in [m]$, i_s takes values $i_s = 1, 2, \ldots, K$.

Note that every vertex in $b_{i_1, i_2, \ldots, i_m}$ falls into the influence region of every vertex in $b_{j_1, j_2, \ldots, j_m}$ if the following condition holds:

$(*)$ For all $s \in [m]$, $|i_s - j_s| \leq K/2 - 1$.

Using this observation, we can derive a lower bound on $e(N(v))$, the number of edges in the neighborhood of v. By symmetry, we have that *wep*

$$\begin{aligned} e(N(v)) &\geq \frac{1}{2} \sum_{i_1=1}^{K} \cdots \sum_{i_m=1}^{K} \sum_{j_1=1,(*)}^{K} \cdots \sum_{j_m=1,(*)}^{K} e(b_{i_1, i_2, \ldots, i_m}, b_{j_1, j_2, \ldots, j_m}) \\ &= \frac{1}{2} \sum_{i_1=1}^{K} \cdots \sum_{i_m=1}^{K} \sum_{j_1=1,(*)}^{K} \cdots \sum_{j_m=1,(*)}^{K} (1 + o(1))(p \log^3 n)^2 p \\ &\geq (1 + o(1)) \frac{2^m}{2} \sum_{i_1=1}^{K/2} \cdots \sum_{i_m=1}^{K/2} \sum_{j_1=1}^{K/2-1+i_1} \cdots \sum_{j_m=1}^{K/2-1+i_m} (p \log^3 n)^2 p; \end{aligned}$$

we use the notation $\sum_{j_s=1,(*)}^{K}$ to indicate that the sum is over all j_s between 1 and n such that $(*)$ holds.

By symmetry, we have that the case $i_s \leq K/2$ is symmetric to the case $K - i_s + 1 \leq K/2$. So we can count $e(N(v))$ by considering only the hypercubes $b_{i_1, \ldots, i_m}$ where $i_s < K/2$ for $1 \leq s \leq m$, and multiplying the result by 2^m. Using

this symmetry argument, we have that *wep*

$$e(N(v)) \geq \frac{2^m}{2} \sum_{i_1=1}^{K/2} \cdots \sum_{i_m=1}^{K/2} \sum_{j_1=1}^{K/2-1+i_1} \cdots \sum_{j_m=1}^{K/2-1+i_m} e(b_{i_1,i_2,\ldots,i_m}, b_{j_1,j_2,\ldots,j_m})$$
$$= (1+o(1))\frac{2^m}{2} \sum_{i_1=1}^{K/2} \cdots \sum_{i_m=1}^{K/2} \sum_{j_1=1}^{K/2-1+i_1} \cdots \sum_{j_m=1}^{K/2-1+i_m} (p\log^3 n)^2 p;$$

we use the notation $\sum_{j_s=1,(*)}^{K}$ to indicate that the sum is over all j_s between 1 and n such that $(*)$ holds.

Thus, *wep*

$$\begin{aligned} e(N(v)) &\geq (1+o(1))\frac{2^m}{2} \sum_{i_1=1}^{K/2} \cdots \sum_{i_m=1}^{K/2} \left(\frac{K}{2}-1+i_1\right) \cdots \left(\frac{K}{2}-1+i_m\right) (p\log^3 n)^2 p \\ &= (1+o(1))\frac{2^m}{2} \left(\left(\frac{K/2+(K-1)}{2}\right)\frac{K}{2}\right)^m (p\log^3 n)^2 p \\ &= (1+o(1))\frac{1}{2}\left(\frac{3}{4}\left(1-\frac{2}{3K}\right)K^2\right)^m (p\log^3 n)^2 p \\ &= (1+o(1))\exp\left(-O\left(\frac{m}{K}\right)\right)\left(\frac{3}{4}\right)^m \frac{\left(pn^{1-\alpha-\beta}\right)^2}{2} p. \end{aligned} \tag{4.6}$$

Fix a vertex $v_i \in V(G)$, $i = an$, for some $a \in [0,1]$. Suppose that the initial rank of v_i is $R_i = bn$ for some $b \in (0,1]$. It follows from Theorem 4.4 that *wep*

$$\deg(v_i) = (1+o(1))p\left(\frac{a}{1-\alpha} + b^{-\alpha}(1-a)\right) n^{1-\alpha-\beta}.$$

Hence by (4.6), *wep* $c(v_i)$ may be bounded from below as follows:

$$\begin{aligned} c(v_i) &\geq (1+o(1))\exp\left(O\left(\frac{m}{K}\right)\right)\left(\frac{3}{4}\right)^m \frac{(pn^{1-\alpha-\beta})^2/2}{\binom{\deg(v_i,L)}{2}} p \\ &= (1+o(1))\exp\left(O\left(\frac{m}{K}\right)\right)\left(\frac{3}{4}\right)^m \left(\frac{a}{1-\alpha} + b^{-\alpha}(1-a)\right)^{-2} p. \end{aligned}$$

Since the initial rank is distributed uniformly at random, we derive that the expected value of $c(v_i)$ can be bounded as follows:

$$\mathbb{E}(c(v_i)) \geq (1+o(1))\exp\left(O\left(\frac{m}{K}\right)\right)\left(\frac{3}{4}\right)^m p \int_0^1 \left(\frac{a}{1-\alpha} + b^{-\alpha}(1-a)\right)^{-2} db.$$

Therefore, we find that *wep*

$$c(G) \geq (1+o(1))\exp\left(O\left(\frac{m}{K}\right)\right)\left(\frac{3}{4}\right)^m pD(\alpha),$$

where

$$D(\alpha) = \int_0^1 \int_0^1 \left(\frac{a}{1-\alpha} + b^{-\alpha}(1-a) \right)^{-2} db\, da.$$

After changing the order of integration, it is straightforward to see that

$$\begin{aligned} D(\alpha) &= \int_0^1 \int_0^1 \left(a\left(\frac{1}{1-\alpha} - b^{-\alpha} \right) + b^{-\alpha} \right)^{-2} da\, db \\ &= -\int_0^1 \left[\left(a\left(\frac{1}{1-\alpha} - b^{-\alpha} \right) + b^{-\alpha} \right)^{-1} \left(\frac{1}{\frac{1}{1-\alpha} - b^{-\alpha}} \right) \right]_{a=0}^{a=1} \\ &= -\int_0^1 \left((1-\alpha) - b^{\alpha} \right) \left(\frac{1}{\frac{1}{1-\alpha} - b^{-\alpha}} \right) db \\ &= \left[\frac{1-\alpha}{1+\alpha}\, b^{1-\alpha} \right]_{b=0}^{b=1} = \frac{1-\alpha}{1+\alpha}, \end{aligned}$$

which finishes the proof of the theorem.

This proof shows that as the dimension m increases, this lower bound on the size of $e(N(v))$, and thus the number of triangles, decreases exponentially. This concurs with our interpretation of the dimension as the minimal number of attributes that characterizes a user in a social network. It makes sense that assortativity decreases as the dimension increases: if a user has a friend A with whom she shares interests in one dimension and a friend B with whom she connects in another dimensions, then the chances that A and B become friends may not be much larger than dictated by random chance.

4.4. Proof of Theorem 3.5

In order to prove Theorem 3.5, we show that there are sparse cuts in our model. As in the previous subsection, 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.

Theorem 4.5. *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*

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

Proof. Let us call a vertex v *dangerous* if at some point of the protean process, the influence region of v has nonempty intersection with both S_1 and S_2; that is, there exists t such that $R(v,t) \cap S_1 \neq \emptyset$ and $R(v,t) \cap S_2 \neq \emptyset$. It is easy to see that the older vertex of every edge between S_1 and S_2 must be dangerous. We are going to estimate $e(S_1, S_2)$ by investigating the number of dangerous vertices with the final rank from a given range. Since the number of edges adjacent to a given dangerous vertex in the cut can be estimated by its degree, the conclusion can be obtained.

First, let us consider vertices with small ranks, that is, vertices with $r(v, L) < \sqrt{n} \log^2 n$. Unfortunately, for these vertices we cannot control the behavior of the corresponding random variables $r(v,t)$ during the process. However, deterministically $|R(v,t)| = O(n^{-\beta})$ for all vertices at any point of the process. Since vertices are distributed in S uniformly at random, the expected number of dangerous vertices with $r(v, L) < \sqrt{n} \log^2 n$ can be estimated by $O(n^{-\beta/m})\sqrt{n} \log^2 n = O(n^{1/2-\beta/m} \log^2 n)$. Hence *wep* the number of dangerous vertices from this range is

$$O(n^{\max\{1/2-\beta/m, 0\}} \log^2 n)$$

by the Chernoff bound. As we already mentioned, it is impossible to estimate the number of edges adjacent to these vertices. However, by considering the extreme case, that is, the case in which these vertices have the smallest possible ranks during the whole time, we obtain that *wep* the number of edges between these vertices and older ones is at most

$$\sum_{r=1}^{O\left(n^{\max\{1/2-\beta/m,0\}} \log^2 n\right)} r^{-\alpha} n^{1-\beta} = O\left(n^{1-\beta+(1-\alpha)\max\{1/2-\beta/m,0\}} \log^{2(1-\alpha)} n\right) = o\left(n^{2-\alpha-\beta}\right).$$

Now, for a given $i = 1, 2, \ldots, \frac{1}{2}\log_2 n - 2\log_2 \log n$, consider vertices with

$$2^{i-1}\sqrt{n} \log^2 n \leq r(v, L) < 2^i \sqrt{n} \log^2 n. \tag{4.7}$$

By Lemma 4.2, *wep* $r(v,t) = (1 + O(\log^{-1/2} n))r(v,L)$ (which implies that $|R(v,t)| = (1 + O(\log^{-1/2} n))r(v,L)^{-\alpha} n^{-\beta}$) during its whole life. It follows from the Chernoff bound that *wep* the number of dangerous vertices that satisfy (4.7) is at most

$$O((2^i \sqrt{n} \log^2 n)^{1-\alpha/m} n^{-\beta/m}),$$

and the number of edges adjacent to these vertices can be estimated by

$$\begin{aligned} &O\left(\left(2^i \sqrt{n} \log^2 n\right)^{1-\alpha/m} n^{-\beta/m}\right) \cdot O\left(\left(2^i \sqrt{n} \log^2 n\right)^{-\alpha} n^{1-\beta}\right) \\ &\quad = O\left(\left(2^i \sqrt{n} \log^2 n\right)^{1-\frac{m+1}{m}\alpha} n^{1-\frac{m+1}{m}\beta}\right). \end{aligned}$$

Finally, *wep* the number of edges in the cut that are adjacent to vertices with final ranks at least $\sqrt{n} \log^2 n$ is

$$\begin{aligned} \sum_{i=1}^{\frac{1}{2}\log_2 n - 2\log_2 \log n} & O\left((2^i \sqrt{n} \log^2 n)^{1-\frac{m+1}{m}\alpha} n^{1-\frac{m+1}{m}\beta}\right) \\ &= \begin{cases} O(n^{2-\frac{m+1}{m}\alpha-\frac{m+1}{m}\beta}), & \text{if } \alpha < \frac{m}{m+1}, \\ O(n^{1-\frac{m+1}{m}\beta} \log n), & \text{if } \alpha = \frac{m}{m+1}, \\ O(n^{\frac{3}{2}-\frac{m+1}{m}\frac{\alpha}{2}-\frac{m+1}{m}\beta} \log^{2-2\frac{m+1}{m}\alpha} n), & \text{if } \alpha > \frac{m}{m+1}, \end{cases} \end{aligned}$$

which is $o(n^{2-\alpha-\beta})$, provided that $m = o(\log n)$. Hence, item (1) of the theorem follows.

For item (2) in the case $m = C \log n$ for some $C > 0$, we must be more precise and take care of constants hidden in the $O(\cdot)$ notation. It can be shown that *wep*

$$\begin{aligned} e(S_1, S_2) &\le (1 + o(1)) \frac{p}{4(1-\alpha)} n^{2-\frac{m+1}{m}\alpha-\frac{m+1}{m}\beta} \\ &= (1 + o(1)) \frac{p}{4(1-\alpha)} n^{2-\alpha-\beta} \exp\left(-(\alpha+\beta)\frac{\log n}{m}\right) \\ &= (1 + o(1)) \frac{p}{4(1-\alpha)} n^{2-\alpha-\beta} \exp\left(-\frac{\alpha+\beta}{C}\right), \end{aligned}$$

and the proof is complete. □

To finish the proof of Theorem 3.5, we use the expander mixing lemma for the normalized Laplacian (see [Chung 97] 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 $X \setminus Y$ and Y plus twice the number of edges that contain only vertices of $X \cap Y$.)

Lemma 4.6. *For all sets $X \subseteq G$,*

$$\left| e(X,X) - \frac{(\mathrm{vol}(X))^2}{\mathrm{vol}(G)} \right| \leq \lambda \frac{\mathrm{vol}(X)\mathrm{vol}(\bar{X})}{\mathrm{vol}(G)}. \tag{4.8}$$

Proof of Theorem 3.5. It follows from (4.3) and the Chernoff bound that *wep*

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

Suppose first that $m = o(\log n)$. From Theorem 4.5 we derive that *wep*

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

and Lemma 4.6 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 $C > 0$. Using Theorem 4.5 one more time, we find that *wep*

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

As before, the assertion follows directly from Lemma 4.6. □

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 1 to 4 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 4 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 that 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 [Chierichetti et al. 09] 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.

Acknowledgments. The authors gratefully acknowledge support from NSERC and MITACS. The present article is the full version of an article that appeared in [Bonato et al. 10].

References

[Adamic et al. 03] L. A. Adamic, O. Buyukkokten, and E. Adar. "A Social Network Caught in the Web." *First Monday* 8:6 (2003).

[Ahn et al. 07] Y. Ahn, S. Han, H. Kwak, S. Moon, and 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.

[Aiello et al. 09] W. Aiello, A. Bonato, C. Cooper, J. Janssen, and P. Prałat. "A Spatial Web Graph Model with Local Influence Regions." *Internet Mathematics* 5 (2009), 175–196.

[Bonato 08] A. Bonato. *A Course on the Web Graph*, American Mathematical Society Graduate Studies Series in Mathematics. Providence: AMS, 2008.

[Bonato et al. 10] 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, pp. 110–121. New York: Springer, 2010.

[Bonato et al. 11] A. Bonato, N. Hadi, P. Horn, P. Prałat, and C. Wang. "Models of On-Line Social Networks." *Internet Mathematics* 6 (2011), 285–313.

[Chierichetti et al. 09] F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, and P. Raghavan. "On Compressing Social Networks." In *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09)*, 2009.

[Chung 97] F. R. K. Chung, *Spectral Graph Theory*. Providence: American Mathematical Society, 1997.

[Chung 04] F. R. K. Chung, L. Lu, *Complex Graphs and Networks*, Providence: American Mathematical Society, 2004.

[Estrada 06] E. Estrada. "Spectral Scaling and Good Expansion Properties in Complex Networks." *Europhys. Lett.* 73 (2006), 649–655.

[Flaxman et al. 07] A. Flaxman, A. Frieze, and J. Vera. "A Geometric Preferential Attachment Model of Networks" *Internet Mathematics* 3 (2007), 187–205.

[Fortunato et al. 06] S. Fortunato, A. Flammini, and F. Menczer. "Scale-Free Network Growth by Ranking." *Phys. Rev. Lett.* 96:21 (2006), 218701.

[Henry and Prałat 11] A. Henry and 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.

[Janson et al. 00] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*. New York: Wiley, 2000.

[Janssen and Prałat 09] J. Janssen and P. Prałat. "Protean Graphs with a Variety of ranking Schemes" *Theoretical Computer Science* 410 (2009), 5491–5504.

[Java et al. 07] A. Java, X. Song, T. Finin, and B. Tseng. "Why We Twitter: Understanding Microblogging Usage and Communities." In *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*, 2007.

[Kleinberg 00] J. Kleinberg. "The Small-World Phenomenon: An Algorithmic Perspective." In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[Kumar et al. 06] R. Kumar, J. Novak, and 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.

[Kwak et al. 10] H. Kwak, C. Lee, H. Park, and S. Moon. "What Is Twitter, a Social Network or a News Media?" In *Proceedings of the 19th International World Wide Web Conference*, 2010.

[Lattanzi and Sivakumar 09] S. Lattanzi and D. Sivakumar. "Affiliation Networks." In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, 2009.

[Leskovec et al. 05a] J. Leskovec, J. Kleinberg, and 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.

[Leskovec et al. 05b] J. Leskovec, D. Chakrabarti, J. Kleinberg, and 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.

[Levin et al. 09] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. Providence: American Mathematical Society, 2009.

[Liben-Nowell et al. 05] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. "Geographic Routing in Social Networks." *Proceedings of the National Academy of Sciences* 102 (2005) 11623–11628.

[Łuczak and Prałat 06] T. Łuczak and P. Prałat. "Protean Graphs." *Internet Mathematics* 3 (2006), 21–40.

[Mislove et al. 07] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. "Measurement and Analysis of On-Line Social Networks." In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 2007.

[Newman and Park 03] M. E. J. Newman and J. Park. "Why Social Networks Are Different from Other Types of Networks." *Phys. Rev. E* 68:3 (2003), 036122.

[Watts and Strogatz 98] D. J. Watts and S. H. Strogatz. "Collective Dynamics of 'Small-World' Networks." *Nature* 393 (1998) 440–442.

[Watts et al. 02] D .J. Watts, P. S. Dodds, and M. E. J. Newman. "Identity and Search in Social Networks." *Science* 296 (2002) 1302–1305.

Anthony Bonato, Department of Mathematics, Ryerson University, Toronto, ON, Canada M5B 2K3 (abonato@ryerson.ca)

Jeannette Janssen, Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada B3H 3J5 (janssen@mathstat.dal.ca)

Paweł Prałat, Department of Mathematics, West Virginia University, Morgantown, WV 26506-6310, USA (pralat@math.wvu.edu)

Received April 2, 2011; accepted June 21, 2011.