ON THE VARIANCE OF THE RANDOM SPHERE OF INFLUENCE GRAPH[§]

P. Hitczenko[*†], S. Janson[†], and J.E. Yukich[‡]

Abstract.
We show that the variance of the number of edges in the random sphere of influence graph built on n i.i.d. sites which are uniformly distributed over the unit cube in $\mathbf{R}^d$, grows linearly with n. This is then used to establish a central limit theorem for the number of edges in the random sphere of influence graph built on a Poisson number of sites. Some related proximity graphs are discussed as well.

1. Introduction.

The main focus of this paper is to find a growth rate on the variance of the number of edges in the sphere of influence graph. These graphs have been introduced by Toussaint [15] and, according to specialists in pattern recognition, perform better than previously used proximity graphs. Let $X_1, \ldots, X_n$ be i.i.d. random variables uniformly distributed on the unit cube in $\mathbf{R}^d$, $d \geq 2$. We will call these points sites. The *random sphere of influence graph* is constructed as follows: for each i, let B_i be a ball around X_i with radius equal to $\min\{|X_i - X_j| : i \neq j\}$ (i.e. the distance from X_i to its closest neighbor). This ball is often called the sphere of influence of X_i. We draw an edge between X_j and X_k if and only if the balls B_j and B_k overlap. The quantity of interest is the total number of edges in the graph. We call this the size of the graph and denote it by $\mathbf{e}$, $\mathbf{e}_n$, or $\mathbf{e}(X_1, \ldots, X_n)$. It is known that we always have $c_d n \leq \mathbf{E}\mathbf{e} \leq C_d n$ for some absolute positive constants c_d and C_d (we refer the reader to the survey paper [10] for detailed references on this result). Füredi [7] showed that $\mathbf{E}\mathbf{e}/n$ has a limit as $n \to \infty$ and identified the value of that limit (in the case $d = 2$, the limiting value is $1 + \pi/4$.) The same result was later obtained by Chalker *et al.* in [3]. The authors of the latter paper also found a bound on the tail probability of the deviation of the size of sphere of influence graph from its mean. Fluctuations results are a bit more difficult because the sphere of influence graph does not have very good regularity properties: in certain configurations, relocating just one site can lead to a significant change in the number of edges. (This is seen by considering the following situation: if $n-1$ sites are regularly spaced on a circle, and the last site is near the center of that circle, then it is incident to all $n-1$ remaining sites; if it is moved to the boundary of the circle its degree becomes bounded independently of n. Changes in the degrees of other sites are insignificant.)

---

* Supported in part by NSF grant DMS 9401345

† part of the research of these authors was done while they were in residence at the meeting Random Graphs and Combinatorial Structures in Oberwolfach, October 1997

‡ Supported in part by NSA grant MDA 904–97–1–0053

AMS 1991 Subject Classification: 60C05, 60D05, 60E15, 05C80, 60G42

This paper concentrates mainly on finding the upper and lower bounds on the variance of the size of the sphere of influence graph. We will show that the variance grows linearly in n, the total number of sites.

Theorem 1. *Let $\mathbf{e}_n$ be the number of edges in the sphere of influence graph built on n i.i.d. sites uniformly distributed over $[0,1]^d$. Then there exist absolute positive constants, c_d and C_d such that for any $n \geq 4$*

$$c_d n \leq \operatorname{var}(\mathbf{e}_n) \leq C_d n.$$

The proof of the upper bound is based on the Efron – Stein inequality and will be given in the next section. The lower bound follows a technique developed by Avram and Bertsimas [1], and we will give the details in section 3.

One consequence of our variance bound is the fact that, at least if we consider a Poisson distributed number of points, the total number of edges, normalized in the usual way, satisfies the central limit theorem. As it turns out, our methods give upper bounds on higher moments, and work for some related proximity graphs as well. We will discuss some of these results briefly in the last section. Throughout the paper, the constants (denoted by various letters) are always absolute. They can possibly depend on d which is considered arbitrary, but fixed. The value of a constant may change from line to line. The volume of the unit ball in $\mathbf{R}^d$ is denoted by v_d.

2. Upper bound for the variance.

Given the sites $X_1, \ldots, X_n$, let $D(X_i)$ be the degree of a vertex X_i (i.e. the number of edges incident to X_i) in the sphere of influence graph. The general approach uses the Efron – Stein inequality, along the same lines as in Steele [14], Section 6. To this end, let us denote by $\mathbf{e}_i^*$ the size of the graph with the ith observation withheld, i.e.

$$\mathbf{e}_i^* = \mathbf{e}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

We wish to apply the Efron – Stein inequality [6], which says that

$$\operatorname{var}(\mathbf{e}_{n-1}) = \operatorname{var}(\mathbf{e}_n^*) \leq \mathbf{E} \sum_{i=1}^{n} \left(\mathbf{e}_i^* - \frac{1}{n} \sum_{j=1}^{n} \mathbf{e}_j^* \right)^2 .$$

Since the average minimizes the sum under the above expectation, it can be replaced by any other quantity, for example by $\mathbf{e}$. Thus, we will need an upper bound on $\mathbf{E} \sum_{i=1}^{n} (\mathbf{e}_i^* - \mathbf{e})^2$. To this end, first observe that adding a new site creates new edges (incident to that site) and may remove edges between old sites, but cannot create any new edges between two old sites. Therefore,

$$\mathbf{e} - \mathbf{e}_i^* \leq D(X_i).$$

We will now estimate $\mathbf{e}_i^* - \mathbf{e}$. Let

$$N(X_i) = \{ j \leq n, j \neq i : \ |X_j - X_i| \leq |X_j - X_k|, \ k \neq i \}$$

be the (possibly empty) set of (indices) of sites for which X_i is the nearest neighbor. Consider $\mathbf{e}$ and suppose that X_i is removed. Then, every site for which X_i was the nearest neighbor will have to find a new nearest neighbor. As a result, its new sphere of influence will have a larger radius, and may therefore intersect other spheres, causing an increase in the total number of edges. However, for each such site, say X_j, the increase cannot be more than the degree of X_j in the random sphere of influence graph with vertex i withheld. Let us denote this degree by $D_i(X_j)$. Then we can write

$$\mathbf{e}_i^* \leq \mathbf{e} + \sum_{j \in N(X_i)} D_i(X_j),$$

i.e.

$$\mathbf{e}_i^* - \mathbf{e} \leq \sum_{j \in N(X_i)} D_i(X_j).$$

Combining this with the previous estimate we obtain

$$(\mathbf{e}_i^* - \mathbf{e})^2 \leq D^2(X_i) + \left(\sum_{j \in N(X_i)} D_i(X_j) \right)^2 .$$

By the Cauchy – Schwarz inequality the rightmost term can be bounded above by

$$D^2(X_i) + |N(X_i)| \sum_{j \in N(X_i)} D_i^2(X_j),$$

where $|N(X_i)|$ denotes the cardinality of $N(X_i)$. It is easy to see [3] that $|N(X_i)| \leq K_d$, where K_d is a constant depending only on the dimension d. Thus taking expectations, we infer that

$$\begin{aligned} \operatorname{var}(\mathbf{e}_n^*) &\leq \sum_{i=1}^{n} \mathbf{E} \left(D^2(X_i) + K_d \sum_{j \in N(X_i)} D_i^2(X_j) \right) \\ &= n\mathbf{E} \left(D^2(X_1) + K_d \sum_{j \in N(X_1)} D_1^2(X_j) \right), \end{aligned}$$

since the random variables in question are identically distributed. Thus the proof will be complete once we show that

$$\mathbf{E}D^2(X_1) \leq C,$$

and

$$\mathbf{E} \sum_{j \in N(X_1)} D_1^2(X_j) \leq C.$$

As we will see below, the first estimate follows from the proof of the second so we concentrate on the second bound. To this end, let I_j be the indicator of the event that the nearest neighbor of X_j is X_1. Further, let $\mathcal{G}$ be the σ-algebra generated by $X_2, \ldots, X_n$. Then we have

$$\mathbf{E} \sum_{j \in N(X_1)} D_1^2(X_j) = \sum_{j=2}^{n} \mathbf{E}\left(I_j D_1^2(X_j)\right) = \sum_{j=2}^{n} \mathbf{E}\left(D_1^2(X_j)\mathbf{E}(I_j|\mathcal{G})\right).$$

Denote by p_j the conditional expectation $\mathbf{E}(I_j|\mathcal{G})$. Using Hölder's inequality, we see that the sum above can be bounded above by

$$\sum_{j=2}^{n} (\mathbf{E}D_1^3(X_j))^{2/3}(\mathbf{E}p_j^3)^{1/3}.$$

Thus, in order to complete this part of the proof, it suffices to show that $\mathbf{E}p_j^3 \leq C/n^3$ and $\mathbf{E}D_1^3(X_j) = \mathbf{E}D_1^3(X_2) \leq C$.

In order to justify the second statement (which we will do for $D(X_1)$ rather than $D_1(X_2)$ in order to simplify notation), let $I_{i,j}$ be the event that there is an edge between sites X_i and X_j, and for simplicity we will denote its indicator by the same symbol.

Then we have

$$\begin{aligned} \mathbf{E}D^3(X_1) = \mathbf{E}\left(\sum_{j=2}^{n} I_{1,j}\right)^3 &= \mathbf{E} \sum_{2 \leq i,j,k \leq n} I_{1,i} I_{1,j} I_{1,k} \\ &= (n-1)\mathbf{P}(I_{1,2}) + 3(n-1)(n-2)\mathbf{P}(I_{1,2} \cap I_{1,3}) \\ &\qquad + (n-1)(n-2)(n-3)\mathbf{P}(I_{1,2} \cap I_{1,3} \cap I_{1,4}). \end{aligned}$$

It was shown in [5] (or in [3]) that $\mathbf{P}(I_{1,2}) \approx c_d/n$ for a constant c_d depending only on d. So, it remains to show that $\mathbf{P}(I_{1,2} \cap I_{1,3}) \leq \dfrac{c_d}{n^2}$ and $\mathbf{P}(I_{1,2} \cap I_{1,3} \cap I_{1,4}) \leq \dfrac{c_d}{n^3}$. Since both computations are essentially the same, we will present only the second one. We also note that the first inequality implies that $\mathbf{E}D^2(X_1) \leq C$.

Let $\mathbf{P}_{t_2,t_3,t_4}(\,\cdot\,)$ be the conditional probability given that $|X_1 - X_2| = t_2$, $|X_1 - X_3| = t_3$, and $|X_1 - X_4| = t_4$. Then, by symmetry

$$\mathbf{P}(I_{1,2} \cap I_{1,3} \cap I_{1,4}) = 3! \int_{0<t_2<t_3<t_4} \mathbf{P}_{t_2,t_3,t_4}(I_{1,2} \cap I_{1,3} \cap I_{1,4}) f_{2,3,4}(t_2,t_3,t_4)\, dt_2 dt_3 dt_4,$$

where $f_{2,3,4}$ is the joint density of the distances between X_1 and X_2, X_1 and X_3, and X_1 and X_4.

In the computation below, the statement "$B_r(X_j)$ is empty" means that no other site is contained in a ball with radius r centered at X_j. If the distance between X_1 and X_4 is t_4, and there is an edge between those two sites, then one of the balls $B_{t_4/2}(X_1)$ or $B_{t_4/2}(X_4)$ must be empty. It follows that

$$\begin{aligned} &\mathbf{P}_{t_2,t_3,t_4}(I_{1,2} \cap I_{1,3} \cap I_{1,4}) \leq \mathbf{P}_{t_2,t_3,t_4}(I_{1,4}) \\ &\quad \leq \mathbf{P}_{t_2,t_3,t_4}\left(\left\{B_{t_4/2}(X_1) = \emptyset\right\} \cup \left\{B_{t_4/2}(X_4) = \emptyset\right\}\right). \end{aligned}$$

Now, $B_{t_4/2}(X_1)$ or $B_{t_4/2}(X_4)$ is empty means that one of these balls contains none of the remaining $n-4$ points. It follows that this last probability is no more than (assuming as we may that $t_4 < \sqrt{d}$)

$$2\left(1 - c_d\left(\frac{t_4}{2}\right)^d\right)^{n-4},$$

where $c_d > 0$. If $t_4 > t_3 > t_2$, then $t_4^d \geq (1/3)(t_4^d + t_3^d + t_2^d)$, and using the inequality $1 - x \leq e^{-x}$ we see that the quantity above does not exceed

$$2\exp\Big\{-\frac{(n-4)c_d}{2^d}t_4^d\Big\} \leq 2\exp\Big\{-\frac{(n-4)c_d}{3\cdot 2^d}\big(t_4^d + t_3^d + t_2^d\big)\Big\}.$$

Substituting this quantity in the integral above, and noting that (by conditioning on X_1) $f_{2,3,4}(t_2,t_3,t_4) \leq c_d t_2^{d-1}t_3^{d-1}t_4^{d-1}$, we find that

$$\begin{aligned}
&\mathbf{P}(I_{1,2}\cap I_{1,3}\cap I_{1,4})\\
&\quad\leq c_d\int_{0<t_2<t_3<t_4}\exp\Big\{-\frac{(n-4)c_d}{3\cdot 2^d}\big(t_2^d+t_3^d+t_4^d\big)\Big\}t_2^{d-1}t_3^{d-1}t_4^{d-1}dt_2dt_3dt_4\\
&\quad\leq c_d\left(\int_0^\infty \exp\Big\{-\frac{(n-4)c_d}{3\cdot 2^d}t^d\Big\}t^{d-1}dt\right)^3\\
&\quad= \frac{c_d}{(n-4)^3},
\end{aligned}$$

as desired.

Finally, to obtain an upper bound on $\mathbf{E}p_j^3$, we note that since p_j is equal to the volume of the intersection of the unit cube and the sphere of influence about X_j (in $\mathbf{e}_1^*$), and denoting this sphere by $SIG_1(X_j)$ and its radius by $r_{SIG_1(X_j)}$, we see that

$$\mathbf{E}p_j^3 \leq \mathbf{E}\mathrm{vol}^3(SIG_1(X_j)) = v_d^3\mathbf{E}r_{SIG_1(X_j)}^{3d} = 3dv_d^3\int_0^{\sqrt{d}} r^{3d-1}\mathbf{P}(r_{SIG_1(X_j)} \geq r)dr,$$

and since (for $r \leq \sqrt{d}$)

$$\mathbf{P}(r_{SIG_1(X_j)} \geq r) \leq C(1 - c_d r^d)^{n-2} \leq \exp\{-c_d r^d(n-2)\},$$

it follows that $\mathbf{E}p_j^3 \leq C/n^3$.

3. Lower bound for the variance.

To obtain the lower bound we will use a technique developed by Avram and Bertsimas [1], see also Steele [13], Section 5.8. It is convenient to change our notation; let S be a subset of $[0,1]^d$, and denote by $\mathbf{e}(S)$ the sum of the degrees of vertices contained in S. (Note that, if both endpoints of an edge are contained in S, this edge is counted twice.) Therefore, we have $\mathbf{e}([0,1]^d) = 2\mathbf{e}$, and if $S_1, \ldots S_k$ are pairwise disjoint subsets of $[0,1]^d$ then

$$\mathbf{e}(\bigcup_{j=1}^{k} S_j) = \sum_{j=1}^{k} \mathbf{e}(S_j).$$

Let $\ell = \lceil n^{1/d} \rceil$ and subdivide $[0,1]^d$ into $n_1 = \ell^d$ congruent subcubes $Q_1, \ldots, Q_{n_1}$ with edge length $1/\ell$. Let $m_d = 4(1 + \lceil \sqrt{d} \rceil) + 1$ and subdivide each subcube into congruent subcubes with edge length $\varepsilon = 1/m_d \ell$ (so that Q_i is subdivided into m_d^d cubes). Let C_i be the cube in the center of Q_i and for each $1 \le i \le n_1$ let A_i be the event that:

(i) C_i contains exactly 4 of the sites $X_1, \ldots, X_n$,

(ii) each of the $m_d^d - (m_d - 2)^d$ subcubes of Q_i sharing a face with the boundary of Q_i contains exactly one site, and

(iii) the remainder of the cube Q_i contains no more sites.

Since there are n sites and the volume of each of the subcubes of Q_i is proportional to $1/n$, $\mathbf{P}(A_i) \ge \alpha_d$, where α_d is a positive number not depending on n. In particular, $\mathbf{E} \sum_{i=1}^{n_1} I_{A_i} \ge \alpha_d n$. Moreover, m_d is chosen large enough so that on the event A_i the four sites inside C_i have edges only among themselves. Let $\mathcal{G}$ be the σ-algebra generated by everything except the location of the four sites within C_i, for those C_i for which A_i occurs. That is, if $J(\omega) = \{j : \omega \in A_j\}$, then

$$\mathcal{G} = \sigma \left\{ J, \{X_1, \ldots, X_n\} \bigcap \left(\bigcup_{j \in J} C_j \right)^c \right\}.$$

Since for any random variable Y and any σ - algebra $\mathcal{F}$ we have

$$\mathrm{var}(Y) = \mathrm{var}(\mathbf{E}_{\mathcal{F}} Y) + \mathbf{E}\mathrm{var}_{\mathcal{F}}(Y),$$

where $\mathrm{var}_{\mathcal{F}}(Y) = \mathbf{E}_{\mathcal{F}}(Y - E_{\mathcal{F}} Y)^2$, we have

$$\begin{aligned} 4\mathrm{var}(\mathbf{e}) &= 4\mathrm{var}(\mathbf{E}_{\mathcal{G}}(\mathbf{e})) + 4\mathbf{E}\mathrm{var}_{\mathcal{G}}(\mathbf{e}) \\ &\ge 4\mathbf{E}\mathrm{var}_{\mathcal{G}}(\mathbf{e}) = \mathbf{E}\mathrm{var}_{\mathcal{G}} \left(\sum_{i=1}^{n} \mathbf{e}(C_i) \right) \\ &= \mathbf{E}\mathrm{var}_{\mathcal{G}} \left(\sum_{i \in J} \mathbf{e}(C_i) + \sum_{i \notin J} \mathbf{e}(C_i) \right) \\ &= \mathbf{E}\mathrm{var}_{\mathcal{G}} \left(\sum_{i \in J} \mathbf{e}(C_i) \right), \end{aligned}$$

where the last equality follows from the fact that $\sum_{i\notin J} \mathbf{e}(C_i)$ is $\mathcal{G}$ - measurable and that the variance is translation invariant. Further, given $\mathcal{G}$, the random variables $\mathbf{e}(C_i)$, $i \in J$, are independent. Therefore,

$$\mathbf{E}\mathrm{var}_{\mathcal{G}}\left(\sum_{i\in J} \mathbf{e}(C_i)\right) = \mathbf{E}\sum_{i\in J} \mathrm{var}_{\mathcal{G}}\mathbf{e}(C_i).$$

Conditionally on $\mathcal{G}$, for $i \in J$, the sphere of influence graph built on the four sites that are contained in C_i can have 2, 3, 4, 5, or 6 edges, each with positive probability. Moreover, the number of edges depends only on the location of those four sites within C_i, and thus is independent of n. Consequently, $\mathrm{var}_{\mathcal{G}}\mathbf{e}(C_i) \geq c$, for some constant c not depending on n, and therefore,

$$\mathbf{E}\sum_{i\in J} \mathrm{var}_{\mathcal{G}}\mathbf{e}(C_i) \geq c\mathbf{E}\sum_{i=1}^{n} I_{A_i} \geq c\alpha_d n.$$

Putting all of these estimates together gives

$$\mathrm{var}(\mathbf{e}) \geq cn,$$

for some absolute constant c.

4. Remarks.

This section contains some consequences and results related to Theorem 1. Perhaps the most significant one is a central limit theorem for the size of the random sphere of influence graph built on a Poisson number of points. It seems reasonable to conjecture that the CLT holds for a nonrandom number of sites, but we have not established that. We begin with the CLT and we will discuss higher moments and other proximity graphs later in this section.

(i) Central limit theorem for the size of the sphere of influence graph. In what follows we use the variance bounds of Theorem 1 and the local behavior of the sphere of influence graph to establish the following result (throughout $N(n)$ denotes a Poisson random variable with parameter n, independent of all other random variables under consideration, and in order to emphasize the dependence on n we will write in this section $\mathbf{e}_n = \mathbf{e}(X_1, ..., X_n)$.)

Theorem 2. *(CLT for the number of edges in the sphere of influence graph) We have*

$$\frac{\mathbf{e}_{N(n)} - \mathbf{E}\mathbf{e}_{N(n)}}{\mathrm{var}^{1/2}(\mathbf{e}_{N(n)})} \Longrightarrow N(0,1), \qquad \textit{as} \qquad n \to \infty.$$

The local behavior is formalized through the notion of dependency graphs, an idea used by Avram and Bertsimas [1] to establish the asymptotic normality of the length of the k-nearest neighbor graph on a random sample as well as the length of the Delaunay and

Voronoi tessellations on a random sample. This approach is also discussed in Steele [13, Sections 5.7, 5.8] and Yukich [16, Section 8.3]. The underlying idea is as follows: subdivide $[0,1]^d$ into $Cn/\log n$ subcubes of edge length $C(\log n/n)^{1/d}$. Consider the high probability event that all subcubes contain at least one point and at most $C\log n$ points. Then, conditionally on this event, the sphere of influence graph around a point is determined by only a finite number of the neighboring subcubes. Consequently, the number of edges in the sphere of influence graph satisfies "m-dependence" and thus a central limit theorem, by the theory of dependency graphs (see Baldi and Rinott [2], Janson [9], Petrovskaya and Leontovich [11]). To formalize this discussion, let us recall that if V is a collection of vertices (but not necessarily points in $\mathbf{R}^d$) and $\{Y_i\}$ random variables, then the graph $G = (V, E)$ is a *dependency graph* for the family $\{Y_i\}$ of random variables if the following two conditions are satisfied:

(i) $\{Y_i\}$ are indexed by the vertex set V,

(ii) if V_1 and V_2 are two disjoint sets of vertices of G such that no edge E of G has one endpoint in V_1 and the other in V_2, then the sets of random variables $\{Y_i\}_{i\in V_1}$ and $\{Y_i\}_{i\in V_2}$ are independent.

We then have

Theorem 3. *(Baldi and Rinott, [2]) Let $\{Y_{ni},\ i \in V_n\}$ be random variables with a dependency graph $G_n = (V_n, E_n)$. Let $S_n = \sum_{i\in V_n} Y_{ni}$, $\sigma_n^2 := \mathrm{var} S_n < \infty$. Let D_n denote the maximum degree of G_n and suppose that $|Y_{ni}| \le B_n$ a.s. for all $i \in V_n$. Then for all $x \in \mathbf{R}$ we have*

$$\left|\mathbf{P}\left(\frac{S_n - ES_n}{\sigma_n} \le x\right) - \mathbf{P}(N(0,1) \le x)\right| \le 32(1+6^{1/2})\left(\frac{\mathrm{card}(V_n)D_n^2B_n^3}{\sigma_n^3}\right)^{1/2}.$$

To apply Theorem 3 to the random variables $\mathbf{e}_{N(n)}$ we subdivide $[0,1]^d$ into m subcubes $Q_1, ..., Q_m$ of edge length $s \approx K(\log n/n)^{1/d}$, where K is a large constant to be chosen later. For all $1 \le i \le m$ let $N_i = \frac{1}{2}\mathbf{e}(Q_i)$. This evidently yields the decomposition

$$\mathbf{e}_{N(n)} = \sum_{i=1}^m N_i.$$

Let n_i, $1 \le i \le m$, denote the number of points falling in subcube Q_i. Since each n_i, $1 \le i \le m$, is a Poisson random variable with parameter $\approx K^d \log n$, it follows that if

$$A_n := \bigcap_{i=1}^m \{1 \le n_i \le K' \log n\},$$

then A_n is a high probability event, namely for every $\alpha > 0$ there are $K = K(\alpha)$ and $K' = K'(\alpha)$ such that

$$\mathbf{P}(A_n) \ge 1 - n^{-\alpha}.$$

We now fix $K := K(\alpha)$ and $K' := K'(\alpha)$ with $\alpha = 5$, say.

We next define a distance function on sets of subcubes as follows: if $\mathcal{A} = \bigcup_{i \in A} Q_i$ and $\mathcal{B} = \bigcup_{i \in B} Q_i$ for subsets A and B of $\{1, 2, ..., m\}$, then

$$d(\mathcal{A}, \mathcal{B}) = d_H(\mathcal{A}, \mathcal{B})/s,$$

where d_H denotes the usual Hausdorff distance. (This distance essentially measures the width of the "moat" separating two collection of subcubes where the common unit of measurement is the edge length of a subcube.) We now make the key observation that conditionally on A_n, the sphere of influence around any point has radius at most $\sqrt{d+3}s$; hence neighbors in the sphere of influence graph have distance at most $2\sqrt{d+3}s$, and the degree of a site depends only on the location of the sites with a distance less than $3\sqrt{d+3}s$. Thus, if $\mathcal{A} = \bigcup_{i \in A} Q_i$ and $\mathcal{B} = \bigcup_{i \in B} Q_i$ satisfy $d(\mathcal{A}, \mathcal{B}) > 6\sqrt{d+3}$, then conditionally on A_n the sets of random variables

$$\{N_i, \ i \in A\} \ \text{ and } \ \{N_i, \ i \in B\}$$

are independent.

Define the random variable Y_i to be N_i, conditioned on the event A_n. We now define a dependency graph $G_n := (V_n, E_n)$ as follows. Let V_n consist of the m subcubes $Q_1, ..., Q_m$. Say that the edge $E = (V_i, V_j)$, $1 \leq i, j \leq m$, belongs to E_n if and only if $d(Q_i, Q_j) \leq 6\sqrt{d+3}$. By the argument above, G_n is a dependency graph for $\{Y_i\}$ (although not for the unconditioned variables N_i).

The maximal degree D_n of G_n satisfies $D_n \leq K_d$. Moreover, given the event A_n we have

$$N_i \leq C \log n, \quad 1 \leq i \leq m,$$

and thus $Y_n \leq B_n := C \log n$. Moreover, Theorem 1 and the identity

$$\mathrm{var}(\mathbf{e}_{N(n)}) = \mathrm{var}(\mathbf{E}_{\sigma(N(n))}(\mathbf{e}_{N(n)})) + \mathbf{E}\mathrm{var}_{\sigma(N(n))}(\mathbf{e}_{N(n)})$$

imply that

$$\sigma_n^2 = \mathrm{var}(\mathbf{e}_{N(n)}) \geq cn,$$

where $\sigma(N(n))$ denotes the σ-algebra generated by $N(n)$. Since $\mathbf{P}(A_n) \geq 1 - n^{-5}$, it follows easily that the variance of $\mathbf{e}_{N(n)}$ conditioned on A_n equals $\mathrm{var}(\mathbf{e}_{N(n)}) + o(1) \geq cn$. Thus, by Theorem 3 we have established that conditionally on the event A_n, $Z_n := (\mathbf{e}_{N(n)} n - \mathbf{E}\mathbf{e}_{N(n)})/(\mathrm{var}(\mathbf{e}_{N(n)}))^{1/2}$ converges to a normal $N(0, 1)$ random variable. Finally, it follows that Z_n converges unconditionally to a $N(0, 1)$ random variable by taking limits as $n \to \infty$ in the expression

$$P\{Z_n \leq x\} = P\{Z_n \leq x | A_n\} P\{A_n\} + P\{Z_n \leq x | A_n^c\} P\{A_n^c\}$$

and using the fact that $P\{A_n\} \to 1$ as $n \to \infty$. This proves Theorem 2.

(ii) Higher moments of degrees. Although it is not needed in this paper, it is perhaps worth recording that our method can be used to get bounds for higher moments of the degrees of the vertices in the sphere of influence graph. Indeed, we have

Proposition 4. *For some constants C_d, c_d we have for all $k \geq 1$,*
(a) $\|D(X_1)\|_k \leq C_d k$,
(b) $\mathbf{E}e^{c_d D(X_1)} \leq 2$ *and thus* $\mathbf{P}(D(X_1) \geq k) \leq 2e^{-c_d k}$.

Proof: (a) We have

$$\mathbf{E}D^k(X_1) = \mathbf{E}\left\{ \sum_{2 \leq i_1, \ldots, i_k \leq n} I_{1,i_1} \cdot \ldots \cdot I_{1,i_k} \right\} = \sum_{j=1}^{k} K_{n-1,j} \mathbf{P}(I_{1,2} \cap \ldots \cap I_{1,j+1}),$$

where $K_{n-1,j}$ is the number of terms $I_{1,i_1} \cdot \ldots \cdot I_{1,i_k}$ such that exactly j of the i_m's are distinct. It can, of course, be computed exactly, but for our purposes it is enough to know that $K_{n-1,j} \leq \binom{n-1}{j} j^k$. (This is, in fact, the correct order.) So we get

$$\mathbf{E}D^k(X_1) \leq \sum_{j=1}^{k} \binom{n-1}{j} j^k \mathbf{P}(I_{1,2} \cap \ldots \cap I_{1,j+1}).$$

With analogous notation as before we have, for $n \geq 2j+2$ and $G = \{0 < t_2 < \ldots < t_{j+1}\}$,

$$\begin{aligned}
&\mathbf{P}(I_{1,2} \cap \ldots \cap I_{1,j+1}) \\
&\leq j! \int_G \mathbf{P}_{t_2,\ldots,t_{j+1}} \left(\left\{ B_{t_{j+1}/2}(X_1) = \emptyset \right\} \cup \left\{ B_{t_{j+1}/2}(X_{j+1}) = \emptyset \right\} \right) f_{t_2,\ldots,t_{j+1}} dt_2 \ldots dt_{j+1} \\
&\leq 2j! \int_G (1 - c_d (t_{j+1}/2)^d)^{n-(j+1)} f_{t_2,\ldots,t_{j+1}} dt_2 \ldots dt_{j+1} \\
&\leq j! c_d^j \int_G \exp\left\{ - \frac{(n-j-1)c_d}{j \cdot 2^d} \left(t_2^d + \ldots + t_{j+1}^d \right) \right\} t_2^{d-1} \ldots t_{j+1}^{d-1} dt_2 \ldots dt_{j+1} \\
&= \left(\frac{c_d j}{n-j-1} \right)^j \leq \left(\frac{c_d j}{n} \right)^j.
\end{aligned}$$

For $n \leq 2j+1$ we have the same estimate, since trivially

$$\mathbf{P}(I_{1,2} \cap \ldots \cap I_{1,j+1}) \leq 1 \leq \left(\frac{3j}{n} \right)^j.$$

These give, since $j! > j^j e^{-j}$ by Stirling's formula,

$$\mathbf{E}D^k(X_1) \leq \sum_{j=1}^{k} \binom{n-1}{j} j^k \frac{(cj)^j}{n^j} \leq \sum_{j=1}^{k} k^k \frac{c^j j^j}{j!} \leq c^k k^k,$$

which completes the proof of part (a).
(b) By (a),

$$\mathbf{E}e^{tD(X_1)} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbf{E}D^k(X_1) \leq \sum_{k=0}^{\infty} \frac{t^k C_d^k k^k}{k!} \leq \sum_{k=0}^{\infty} (tC_d e)^k \leq 2,$$

provided $t \le 1/2eC_d$. The second inequality follows by Chebyshev's inequality.

(iii) Proximity graphs. The methods used for the sphere of influence graph can be applied with minor changes to other constructions that are of interest in computational geometry. We will discuss briefly what have been labeled proximity graphs by Devroye [4]. Given n points, $x_1, \ldots, x_n$, one draws an edge between x_i and x_j if and only if a certain region $S(x_i, x_j)$ contains x_i, x_j and no other sites. Typical examples are the relative neighbor graph (RNG) and the Gabriel graph (GG). In the first one, $S(x_i, x_j)$ is a lens which is the intersection of two balls centered and x_i and x_j, respectively, having radius equal to $|x_i - x_j|$, while in GG $S(x_i, x_j)$ is a ball centered at $(x_i + x_j)/2$ with radius $|x_i - x_j|/2$. Our terminology follows Preparata and Shamos [12], which provides more information on these constructions as well as on their applications. We will mention here only that since $S_{GG}(x_i, x_j) \subset S_{RNG}(x_i, x_j)$, RNG is a subgraph of GG, and thus $\mathbf{e}_{RNG} \le \mathbf{e}_{GG}$ (and both are of order n). Here, and throughout the rest of this section, $\mathbf{e}$ with a corresponding subscript denotes the number of edges in a graph whose name is indicated in the subscript. Occasionally, when the argument applies in general, we will omit the subscript. It is known that RNG is a supergraph of the MST and GG is a subgraph of the DT, where MST and DT represent the minimum spanning tree and the Delaunay triangulation, respectively, of n points (although neither MST nor DT is a proximity graph in the sense described above). Let us recall that in DT, there is an edge between x_i and x_j if and only if there is a ball which has both of these two points on the surface and which contains no other points.

In order to bound the variance via the Efron – Stein inequality note that $\mathbf{e} - \mathbf{e}_i^* \le \deg(X_i)$, and $\mathbf{e}_i^* - \mathbf{e} \le J_i$, where

$$J_i = \mathrm{card}\{(j,k) : \; X_i \in S_{j,k}, \; X_m \notin S_{j,k}; \; m \ne i, j, k\},$$

where for simplicity of notation we write $S_{i,j} = S(X_i, X_j)$. Hence,

$$(\mathbf{e} - \mathbf{e}_i^*)^2 \le \deg^2(X_i) + J_i^2$$

and we need to show

$$\mathbf{E} \deg^2(X_1) \le C \qquad \text{and} \qquad \mathbf{E} J_1^2 \le C.$$

Since $\deg_{RNG}(X_1) \le \deg_{GG}(X_1)$, we will show that $\mathbf{E} \deg_{GG}^2(X_1) \le C$. (Note, that $\deg_{RNG} \le C$, but this is not necessarily true for GG). By the same argument as for SIG, we need to bound $\mathbf{P}_{s,t}(I_{1,2} \cap I_{1,3})$, for $t > s > 0$. But, if $|X_1 - X_2| = s$, $|X_1 - X_3| = t$, and $s < t$ then

$$\mathbf{P}_{s,t}(I_{1,2} \cap I_{1,3}) \le \big(1 - \mathrm{vol}(B_{t/2}((X_1 + X_3)/2) \cap [0,1]^d)\big)^{n-3} \le \big(1 - c_d (t/2)^d\big)^{n-3},$$

and the bound on $\mathbf{E} D^2(X_1)$ follows by the same integration. To bound $\mathbf{E} J_1^2$ let $J_{j,k}$ denote the event $\{X_1 \in S_{j,k}, \; X_m \notin S_{j,k}; \; m \ne 1, j, k\}$ as well as its indicator. Then

$$\mathbf{E} J_1^2 = \sum_{(j,k),(\ell,m)} \mathbf{P}(J_{j,k} \cap J_{\ell,m}) = \binom{n-1}{2} \mathbf{P}(J_{2,3}) + 2\binom{n-1}{2}(n-3)\mathbf{P}(J_{2,3} \cap J_{2,4})$$
$$+ \binom{n-1}{2}\binom{n-3}{2}\mathbf{P}(J_{2,3} \cap J_{4,5}),$$

and we need to show that each of those terms is bounded. If $\mathbf{P}_s(\ \cdot\ )$ is the conditional probability given that $|X_2 - X_3| = s$, then

$$\mathbf{P}_s(J_{2,3}) = \mathrm{vol}(S_{2,3})(1 - \mathrm{vol}(S_{2,3} \cap [0,1]^d))^{n-3} \le K s^d (1 - K s^d)^{n-3},$$

and integrating against the density of the distance between X_2 and X_3 we see that the first term is bounded. A proof of the second bound is similar; with obvious notation we have

$$\begin{aligned}\mathbf{P}_{s,t}(J_{2,3} \cap J_{2,4}) &= \mathrm{vol}(S_{2,3} \cap S_{2,4}) \left(1 - \mathrm{vol}(S_{2,4} \cap [0,1]^d)\right)^{n-4} \\ &\le K s^d \left(1 - K(t^d + s^d)/2\right)^{n-4},\end{aligned}$$

if $t > s$. The computation for the third probability is essentially the same, except that $X_1 \in S_{2,3} \cap S_{4,5}$ entails that $|X_2 - X_4| \le \mathrm{diam}(S_{2,3}) + \mathrm{diam}(S_{4,5})$, and if $|X_2 - X_3| = s$ and $|X_4 - X_5| = t$, then $\mathrm{diam}(S_{2,3}) + \mathrm{diam}(S_{4,5}) \le K(s+t)$. Thus

$$\begin{aligned}\mathbf{P}_{s,t}(J_{2,3} \cap J_{4,5}) &= \mathbf{P}_{s,t}\left(J_{2,3} \cap J_{4,5}\Big| |X_2 - X_4| \le K(s+t)\right) \mathbf{P}_{s,t}(|X_2 - X_4| \le K(s+t)) \\ &\le K s^d (s^d + t^d) \left(1 - K t^d\right)^{n-5},\end{aligned}$$

andthe last estimate follows.

As for the lower bound, the basic principle is the same as for the sphere of influence graph: we subdivide the unit cube into small cubes, and consider those cubes for which there is a boundary behavior that isolates the behavior inside the cube from the developments outside that cube and for which there is a fixed number of points near the center of the cube. A variability in the placements of those central points will cause variability in the number of edges within that cube, and since the average number of cubes for which that happens is of order n, and different cubes behave independently, superlinearity of the variance follows. For example, for GG, assuming $d = 2$ for simplicity, after subdividing the unit square into Q_i's as before, consider the event A_i that the square Q_i contains exactly 5 points located as follows. Assuming $Q_i = [a-h, a+h] \times [b-h, b+h]$, let x^0 be its center (a, b) and let $y_1^0, \ldots, y_4^0$ be the four points $(a \pm \frac{1}{2}h, b)$, $(a, b \pm \frac{1}{2}h)$; A_i then is the event that Q_i contains exactly five points $x, y_1, \ldots, y_4$ with $|x - x^0| < \varepsilon h$ and $|y_j - y_j^0| < \delta h$, where ε and δ are some fixed small numbers.

It is easily seen that if first ε and then δ are chosen small enough, then A_i implies the following, denoting the points outside Q_i by $\{z_j\}$:

(i) If a Gabriel sphere $S(z_j, z_k)$ or $S(z_j, y_k)$ contains x, then it also contains some y_l in its interior.

(ii) A Gabriel sphere $S(z_j, x)$ always contains some y_k.

(iii) A Gabriel sphere $S(y_j, x)$ never contains any other point.

(iv) A Gabriel sphere $S(y_j, y_k)$ contains no other point except possibly x; moreover, if y_j and y_k do not lie opposite each other, both $S(y_j, y_k)$ and its complement intersect $\{x : |x - x^0| < \varepsilon h\}$.

It follows that conditioned on A_i, moving around the central point x may only affect the edges between the y_j's, and the number of such edges will vary and has a non-zero variance. Consequently, conditioning as before on everything except the location of the central point for those Q_i's for which A_i occurs, we obtain a lower bound of order n on the variance of N_{GG}. For RNG the construction is similar.

It should be noted that this approach for a lower bound does not work for DT. In fact, the result is not true for DT. For example, if $d = 2$ and X_i's are uniformly distributed on the unit square then the variance of the *number of edges* in DT is of order $\log n$ rather than n. This is perhaps a bit surprising, especially if one takes into account the fact that the *total length* of the edges in DT has variance of order n (see [1]). To see that variance of the number of edges in DT (or any other triangulation, for that matter) is of order $\log n$ one just notes that $\mathbf{e}_{DT}$ is linked to the number of extreme points of the convex hull of X_i's by the formula $\mathbf{e}_{DT} = 3(n-1) - \mathrm{card}\{\mathrm{ext}(\mathrm{conv}(\{X_i\}))\}$, so that $\mathrm{var}(\mathbf{e}_{DT})$ is equal to the variance of the number of extreme points in the sample $X_1, \ldots, X_n$. But the latter variance is of order $\log n$ if the X_i's are i.i.d. uniform on the unit square [8]. This means that the behavior of DT is not really "local" which is perhaps a bit counterintuitive.

Acknowledgment.

The first author would like to thank Rex Dwyer from the Department of Computer Science at North Carolina State University for several useful conversations.

References.

1. Avram, F. and Bertsimas, D. (1993) On central limit theorems in geometric probability, *Ann. Appl. Probab.* **3**, 1033 – 1046.
2. Baldi, P. and Rinott, Y. (1989) On normal approximations of distributions in terms of dependency graphs, *Ann. Probab.* **17**, 1646-1650.
3. Chalker, T. K., Godbole, A. P., Hitczenko, P., Radcliff, J. and Ruehr, O. G. (1997) On the size of a random sphere of influence graph, preprint.
4. Devroye, L. (1988) The expected size of some graphs in computational geometry, *Computer Math. Appl.* **15**, 53 – 64.
5. Dwyer, R. A. (1995) The expected size of the sphere of influence graph, *Comp. Geometry* **5**, 155 – 164.
6. Efron, B. and Stein, C. (1981) The jackknife estimate of variance, *Ann. Statist.* **9**, 586 – 596.
7. Füredi, Z. (1995) The expected size of a random sphere of influence graph, *Intuitive Geometry*, Bolyai Mathematical Society, **6**, 319 – 326.
8. Groeneboom, P. (1988) Limit theorems for convex hulls, *Probab. Theory Rel. Fields* **79**, 327-368.
9. Janson, S. (1988) Normal convergence by higher semiinvariants with applications to sums of dependent random variables and random graphs, *Ann. Probab.* **16**, 305-312.
10. Michael, T. S. and Quint, T. (1994) Sphere of influence graphs: a survey, *Congressus Numerantium* **105**, 153–160.
11. Petrovskaya, M. and Leontovitch, A. (1982) The central limit theorem for a sequence of random variables with a slowly growing number of dependencies, *Theory Probab. Appl.* **27**, 815-825.
12. Preparata, F. P. and Shamos, M. I. (1985) *Computational Geometry. An Introduction.* Springer.

13. Steele, J. M. (1997) *Probability Theory and Combinatorial Optimization*, NSF-CBMS Regional Research Conference Lecture Notes Series Vol. 69, Society for Industrial and Applied Mathematics, Philadelphia.
14. Steele, J. M. (1995) Variations on the monotone subsequence theme of Erdös and Szekeres, *Discrete Probability and Algorithms*, volume 72, D. Aldous, P. Diaconis, J. Spencer and J. M. Steele, eds. The IMA Volumes in Mathematics and Its Applications, Springer.
15. Toussaint, G. T. (1980) Pattern recognition and geometric complexity. *Proceedings of the 5th International Conference on Pattern Recognition*, Miami Beach, FL, 1324–1347.
16. Yukich, J. E. (1998) *Probability Theory of Classical Euclidean Optimization Problems*, Lecture Notes Math., **1675**, Springer.

P. Hitczenko
Department of Mathematics
North Carolina State University
Raleigh, NC 27695–8205, USA
e-mail: pawel@math.ncsu.edu

S. Janson
Department of Mathematics
Uppsala University
S – 751 06 Uppsala, Sweden
e-mail: svante@math.uu.se

J.E. Yukich
Department of Mathematics
Lehigh University
Bethlehem, PA 18015
e-mail: jey0@lehigh.edu