The distribution of communication cost for a mobile service scenario

June 13, 2009

Jesper Møller, Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9220 Aalborg Ø, Denmark.
Email: jm@math.aau.dk.

Man Lung Yiu, Department of Computer Science, Aalborg University, Selma Lagerlöfs Vej 300, DK-9220 Aalborg, Denmark.
Email: mly@cs.aau.dk.

Abstract: We consider the following mathematical model for a mobile service scenario. Consider a planar stationary Poisson process, with its points radially ordered with respect to the origin (the anchor); these points may correspond to locations of e.g. restaurants. A user, with a location different from the origin, asks for the location of the first Poisson point and keeps asking for the location of the next Poisson point until the first time that he can be completely certain that he knows which Poisson point is his nearest neighbour. The distribution of this waiting time, called the communication cost, is analysed in detail. In particular the expected communication cost and the asymptotic behaviour as the distance between the user and the anchor increases are of interest.

Keywords: nearest-neighbour search; Poisson process; radial simulation algorithm; waiting time.

1 Introduction

The mobile Internet offers services that e.g. receive the location of the nearest point of interest such as a store, restaurant, or tourist attraction, see Yiu et al. (2008, 2009) and the references therein. Tools from applied probability and in particular stochastic geometry should be useful for an analysis of the performance of such services.

We illustrate this by considering a user located at a given point $q \in \mathbb{R}^2$ and data points $p_1, p_2, \ldots \in \mathbb{R}^2$ corresponding to the locations of e.g. stores. In order to preserve some privacy, the user queries a server for nearby points but he reports not his correct location q but another location $q' \in \mathbb{R}^2$ referred to as the anchor. An incremental query processing on the server is used so that the data points are ordered in increasing distance to the anchor. The user then stops to queries the server as soon as possible, i.e., when the nearest data point with respect to q can be determined. The waiting time for this to happen is called the communication cost and is denoted M; further details are given in Section 2.

The purpose of this note is to analyse the distribution of M assuming that the data points follow a stationary Poisson process $\Phi = \{p_1, p_2, \ldots\}$ with intensity $\rho > 0$ (see, e.g., Kingman, 1993, or Chapter 3 in Møller and Waagepetersen, 2004). In particular the expected communication cost and the asymptotic behaviour as the distance between q and q' increases are of interest. Section 3 discusses our results.

2 The communication cost

We use the following notation. By stationarity and isotropy of Φ, we can without loss of generality take $q' = (0,0)$ and $q = (l, 0)$, where $l > 0$ is the distance between the anchor and the user location. Let $(R_i, \theta_i) \in [0, \infty) \times [0, 2\pi)$ be the polar coordinates of p_i with respect to the anchor (the origin), and order the data points such that the R_i are increasing, $R_0 := 0 \leq R_1 \leq R_2 \leq \ldots$. Let $p = p_i$ if p_i be the first NN (nearest-neighbour) to q among $p_1, p_2, \ldots$. For $i = 1, 2, \ldots$, let q_i denote the first NN to q among the i first data points $p_1, \ldots, p_i$, and let $\tilde{R}_i = \|q - q_i\|$ denote the distance from q to q_i. Moreover, let $B(x, r) = \{y \in \mathbb{R}^2 : \|y - x\| \leq r\}$ denote the closed disc with centre $x \in \mathbb{R}^2$ and radius $r \geq 0$.

Then the communication cost M is defined as the first time m such that the user can be completely ensured that $p = q_m$ when he has only received $p_1, \ldots, p_m$ from the server, i.e., no matter where the points $p_{m+1}, p_{m+2}, \ldots$ potentially could be located in $\mathbb{R}^2 \setminus B(q', R_m)$. In other words, if for $m \in \mathbb{N}$, we define the demand space $\mathcal{D}_m$ and the supply space $\mathcal{S}_m$ by

$$\mathcal{D}_m = \mathrm{disc}(q, \tilde{R}_m), \quad \mathcal{S}_m = \mathrm{disc}(q', R_m),$$

then

$$M = \inf\{m \in \mathbb{N} : \mathcal{D}_m \subseteq \mathcal{S}_m\} \tag{1}$$

(setting $\inf \emptyset = \infty$). Furthermore,

$$p = q_M \tag{2}$$

is returned as the NN to q, provided of course $M < \infty$. See Figure 1.

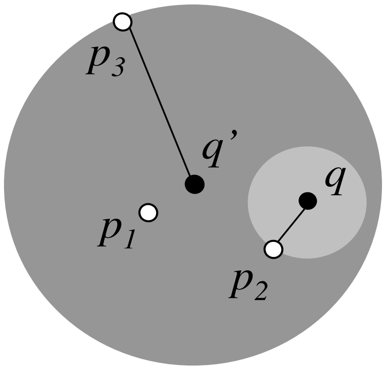

Figure 1: An example with $M = 3$, showing the corresponding demand space $\mathcal{D}_3$ and supply space $\mathcal{S}_3$.

3 Results

Since $M = 1$ implies that p_1 lies on the halfline H with endpoint q and direction $q - q'$, and this happens with probability zero, we have almost surely that $M \geq 2$ and $R_M \geq l$. Similarly, with probability one, for $m \in \{2, 3, \ldots\}$, $M = m$ implies that p_m cannot be the NN to q among $p_1, \ldots, p_m$, because this would imply that $p_m \in H$. Moreover, with probability one, the sequence $R_1, R_2, \ldots$ is strictly increasing to infinity, and so the sequence of supply spaces $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots$ tends to $\mathbb{R}^2$. On the other hand, the sequence of demand spaces decreases. Consequently, by (1) and (2), with probability one, $2 \leq M < \infty$, $\mathcal{D}_1 \supseteq \mathcal{D}_2 \supseteq \ldots \supseteq \mathcal{D}_M$, and

$$p = q_M = q_{M-1}, \quad \tilde{R}_M = \tilde{R}_{M-1}, \quad R_M \geq l. \tag{3}$$

The distribution of M may be estimated by Monte Carlo methods using a radial simulation algorithm due to Quine and Watson (1984). The radial simulation algorithm simply utilizes the fact that the squared radial coordinates $R_1^2, R_2^2, \ldots$ form a homogeneous Poisson process with intensity $\pi\rho$ on the positive halfline, so $R_i^2 - R_{i-1}^2$, $l = 1, 2, \ldots$, are independent and exponentially distributed with mean $1/(\pi\rho)$, and they are independent of the angular coordinates $\theta_1, \theta_2, \ldots$, which in turn are independent and uniformly distributed on $[0, 2\pi)$. The following Lemma 1 follows straightforwardly from these properties. In Lemma 1, when we consider the first m data points and ignore their radial ordering, we write $\{x_1, \ldots, x_m\}$ for the corresponding point configuration. Moreover, $\mathrm{Unif}([0, 2\pi))$ denotes the uniform distribution on $[0, 2\pi)$, and $\Gamma(\alpha, \beta)$ the gamma distribution with shape parameter α and inverse scale parameter β.

Lemma 1 For any $m \in \mathbb{N}$, we have the following.

(i) $R_m^2 \sim \Gamma(m, \pi\rho)$ is independent of $\theta_m \sim \mathrm{Unif}([0, 2\pi))$.

(ii) Conditional on $R_{m+1} = r$, the configuration of the first m points $\{x_1, \ldots, x_m\}$ is a binomial point process on $B(q', r)$ (i.e., the x_i are independent and uniformly distributed on $B(q', r)$).

(iii) Conditional on $R_{m+1} = r$, we have that $\theta_m \sim \mathrm{Unif}([0, 2\pi))$ is independent of R_m, where $R_m/r \sim B(1, m-1)$. If $m \geq 2$ and we condition on both $R_{m+1} = r$ and (R_m, θ_m) with $R_m = t$, then $\{x_1, \ldots, x_{m-1}\}$ is a binomial point process on $B(q', t)$.

We now turn to exact results for the distribution of the communication cost M, where its distribution function turns out to be more tractable than its probability density function, and we can obtain an expression for the expectation $\mathrm{E}M$. We need the following notation. Let

$$f_{m+1}(r) = 2\frac{(\pi\rho)^{m+1}}{m!} r^{2m+1} \exp\left(-\pi\rho r^2\right), \quad r > 0, \tag{4}$$

denote the density function of R_{m+1}, cf. (i) in Lemma 1. For $r \geq l$ and $0 \leq \varphi < 2\pi$, let

$$v(\varphi, r) = \left(r^2 - l^2 \sin^2\varphi\right)^{1/2} - l\cos\varphi. \tag{5}$$

Finally, for $0 \leq r - l \leq s \leq v(\varphi, r)$, define

$$g(r, s) = \frac{|B(q', r) \setminus B(q, s)|}{|B(q', r)|} = 1 - h(r, s)/\left(\pi r^2\right) \tag{6}$$

where

$$\begin{aligned} h(r, s) =& |B(q', r) \cap B(q, s)| \qquad\qquad (7) \\ =& r^2 \arccos\left(\frac{l^2 + r^2 - s^2}{2lr}\right) - \frac{l^2 + r^2 - s^2}{4l^2}\left[4l^2r^2 - \left(l^2 + r^2 - s^2\right)^2\right]^{1/2} \\ &+ s^2 \arccos\left(\frac{l^2 + s^2 - r^2}{2ls}\right) - \frac{l^2 + s^2 - r^2}{4l^2}\left[4l^2s^2 - \left(l^2 + s^2 - r^2\right)^2\right]^{1/2}. \end{aligned}$$

Here $|\cdot|$ denotes area (Lebesgue measure), and we suppress in the notation that the functions in (5)-(7) depend on $l > 0$.

Proposition 1 For $m = 2, 3, \ldots$, we have that

$$\begin{aligned} \mathrm{P}(M > m) = & \int_0^l f_{m+1}(r)\,\mathrm{d}r + \\ & m \int_l^\infty \int_0^{2\pi} \int_{r-l}^{v(\varphi,r)} \frac{s}{\pi r^2} g(r,s)^{m-1} f_{m+1}(r)\,\mathrm{d}s\,\mathrm{d}\varphi\,\mathrm{d}r \end{aligned} \tag{8}$$

and

$$\begin{aligned} \mathrm{E}M = & \, 2 + \pi\rho l^2 - \int_0^l \left[1 + \pi\rho r^2\right] \exp\left(-\pi\rho r^2\right)\,\mathrm{d}r + \\ & \int_l^\infty \int_0^{2\pi} \int_{r-l}^{v(\varphi,r)} 2\pi\rho^2 sr \left[\exp(-\rho h(r,s)) - \exp\left(-\pi\rho r^2\right)\right]\,\mathrm{d}s\,\mathrm{d}\varphi\,\mathrm{d}r \end{aligned} \tag{9}$$

is finite.

Proof Using (1) and (3), considering each of the cases $R_{m+1} < l$ and $R_{m+1} \geq l$, we see that the probability that $M > m$ is equal to the sum of two terms, viz the term

$$\mathrm{P}(R_m \leq l) = \int_0^l f_{m+1}(r)\,\mathrm{d}r$$

and the term

$$\begin{aligned} \int_l^\infty \sum_{i=1}^m \mathrm{P}(& x_i \text{ is the NN to } q \text{ among } x_1, \ldots, x_m \text{ and} \\ & B(q, \|x_i - q\|) \not\subset B(q', r) \mid R_{m+1} = r) f_{m+1}(r)\,\mathrm{d}r. \end{aligned}$$

By (ii) in Lemma 1,

$$\begin{aligned} & \mathrm{P}(x_i \text{ is the NN to } q \text{ among } x_1, \ldots, x_m \text{ and} \\ & \quad B(q, \|x_i - q\|) \not\subset B(q', r) \mid R_{m+1} = r) \end{aligned} \tag{10}$$

$$= \int_{B(q',r)\setminus B(q,r-l)} \left(\frac{|B(q',r) \setminus B(q, \|x-q\|)|}{|B(q',r)|}\right)^{m-1} \frac{1}{|B(q',r)|}\,\mathrm{d}x. \tag{11}$$

Hence, by making a shift of coordinates from

$$x = q + s(\cos\varphi, \sin\varphi)$$

to the polar coordinates (s, φ), we obtain (8) after a straightforward calculation, noticing the following facts: in (11) we can set $q = (l, 0)$ and $q' = (0, 0)$; for fixed $r > l$ and $\varphi \in [0, 2\pi)$, since $x \in B(q', r) \setminus B(q, r-l)$, we obtain that s ranges from $r - l$ to $v(\varphi, r)$; here the latter bound follows from the equation $\|(l, 0) + s(\cos\varphi, \sin\varphi)\| = r$, which has only one solution because $s > 0$. See also Figure 2.

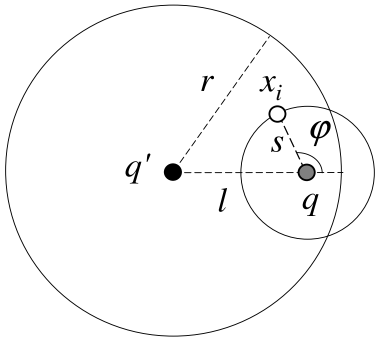

Figure 2: Illustration of the notation used in the proof of Proposition 1. The function $v(\varphi, r)$ is the upper bound on s considering a case with i data points such that $x_i = q_i$ is the nearest-neighbour to q, $B(q', r)$ is the corresponding supply space, and $B(q, s)$ is the corresponding demand space.

Since $M - 2$ is a non-negative discrete random variable, we have that $\mathrm{E}M = 2 + \sum_{m=2}^{\infty} \mathrm{P}(M > m)$. Combining this with (8), we obtain

$$\begin{aligned}\mathrm{E}M = 2 &+ \int_0^l \sum_{m=2}^{\infty} f_{m+1}(r)\,\mathrm{d}r \\ &+ \int_l^{\infty} \int_0^{2\pi} \int_{r-l}^{k(\varphi,r)} \frac{s}{\pi r^2} \sum_{m=2}^{\infty} m g(r,s)^{m-1} f_{m+1}(r)\,\mathrm{d}s\,\mathrm{d}\varphi\,\mathrm{d}r.\end{aligned}$$

From this and (4), we easily obtain (9).

Since $k(\varphi, r) \le r + l$, it follows from (9) that $\mathrm{E}M$ is finite if for some number $r_0 > l$,

$$I := \int_{r_0}^{\infty} \int_{r-l}^{r+l} sr \left[\exp(-\rho h(r,s)) - \exp\left(-\pi\rho r^2\right)\right] \mathrm{d}s\,\mathrm{d}r$$

is finite. This is indeed the case, since for any $\epsilon > 0$, if $r_0 > l + \epsilon$ is sufficiently large, then $h(r,s) \ge \pi(r - l - \epsilon)^2$ whenever $r \ge r_0$ and $r - l \le s \le r + l$, and so

$$0 \le I \le 2l \int_{r_0}^{\infty} r^2 \exp\left(-\pi\rho(r-l-\epsilon)^2\right)\,\mathrm{d}r - 2l \int_{r_0}^{\infty} r^2 \exp\left(-\pi\rho r^2\right)\,\mathrm{d}r$$

where both integrals are finite.

The integrals in (8) and (9) can be evaluated by numerical methods. In (9), as l increases towards infinity, $\mathrm{E}M$ is dominated by the term $\pi\rho l^2$, meaning that the dominating part of the expected communication cost is proportional to ρ and l^2. It is worth noticing that $\pi\rho l^2$ is the expected number of points in the disc $b(q', l)$.

Note that the distribution of M depends only on (ρ, l) through $\sqrt{\rho}l$. Figure 3 visualizes the difference between $\mathrm{E}M$ and its dominant term $\pi\rho l^2$ (shown as 'O'), where for different values of l and with ρ fixed to 1, $\mathrm{E}M$ has been estimated both by using (9) and a numerical method (shown as '$\times$') and by the Monte Carlo method based on 1000 independent simulations obtained by the radial simulation algorithm (shown as '$\bigtriangledown$'). The axes in Figure 3 are shown in the log-scale, the estimates of $\mathrm{E}M$ obtained by the two methods are rather close, and the dominating term of $\mathrm{E}M$ is seen to converge to $\mathrm{E}M$ as $l \to \infty$. The figure also shows the Monte Carlo estimates of the 5% and 95% quantiles (shown as '+').

Figures 4 and 5 show further Monte Carlo estimates with $\rho = 1$ and based on 1000 independent simulations obtained by the radial simulation algorithm for each tested value of l. Figure 4 suggests an approximate log-linear relation between $\sqrt{\mathrm{Var}M}/\mathrm{E}M$ and l. Considering values of $l \geq 2$, the regression line in Figure 4 was estimated by the method of least squares to $y = 0.5524x^{-0.9147}$, where (x, y) corresponds to $(\log_{10} l, \log_{10} \sqrt{\mathrm{Var}M}/\mathrm{E}M)$ ($\log_{10}$ denotes the logarithm function with base 10). Figure 5 shows four histograms of the distribution of M corresponding to the simulations with $l = 0.1, 1, 10, 100$. For the larger values of l, the distribution of M seems to be well approximated by a Gaussian distribution.

Acknowledgements:

Supported by the Danish Natural Science Research Council, grant 272-06-0442, "Point process modelling and statistical inference".

References

[1] J. F. C. Kingman. *Poisson Processes*. Clarendon Press, Oxford, 1993.

[2] J. Møller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton, 2004.

[3] M. P. Quine and D. F. Watson. Radial simulation of n-dimensional Poisson processes. *Journal of Applied Probability*, 21:548–557, 1984.

[4] M. L. Yiu, C. S. Jensen, X. Huang, and H. Lu. SpaceTwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. In M. Castellanos, A. P. Buchmann, and K. Ramamritham, editors, *Proceedings of the 24th IEEE International Conference on Data Engineering*, pages 366–375. IEEE Computer Society, Los Alamitos, CA, 2008.

[5] M. L. Yiu, C. S. Jensen, J. Møller, and H. Lu. Design and analysis of an incremental approach to location privacy for location-based services. Submitted for publication, 2009.

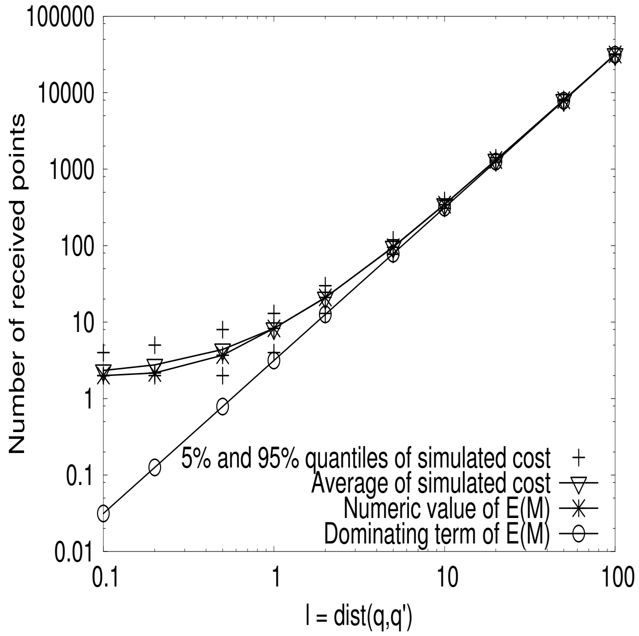

Figure 3: Estimates of EM vs. l when $\rho = 1$ and EM is estimated by the result in Proposition 1 using a numerical method (shown as '$\times$') and by the Monte Carlo method (shown as '$\bigtriangledown$'), together with the dominant term $\pi\rho l^2$ (shown as 'O') and Monte Carlo estimates of the 5% and 95% quantiles (shown as '+').

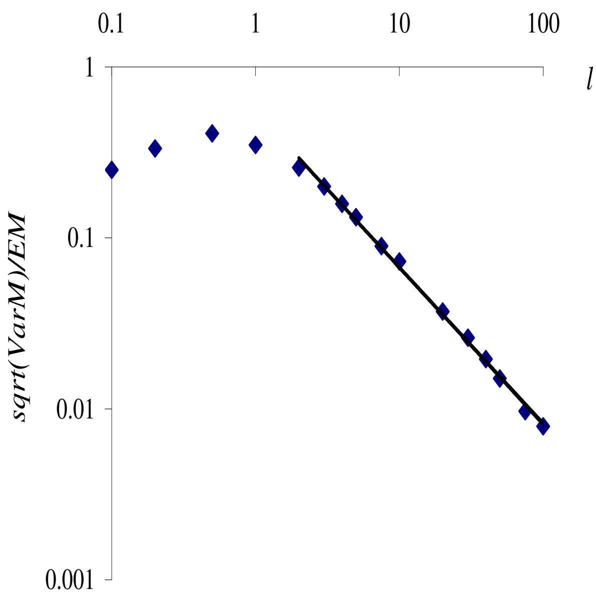

Figure 4: Monte Carlo estimates of $\sqrt{\mathrm{Var} M}/\mathrm{E} M$ vs. l when $\rho = 1$.

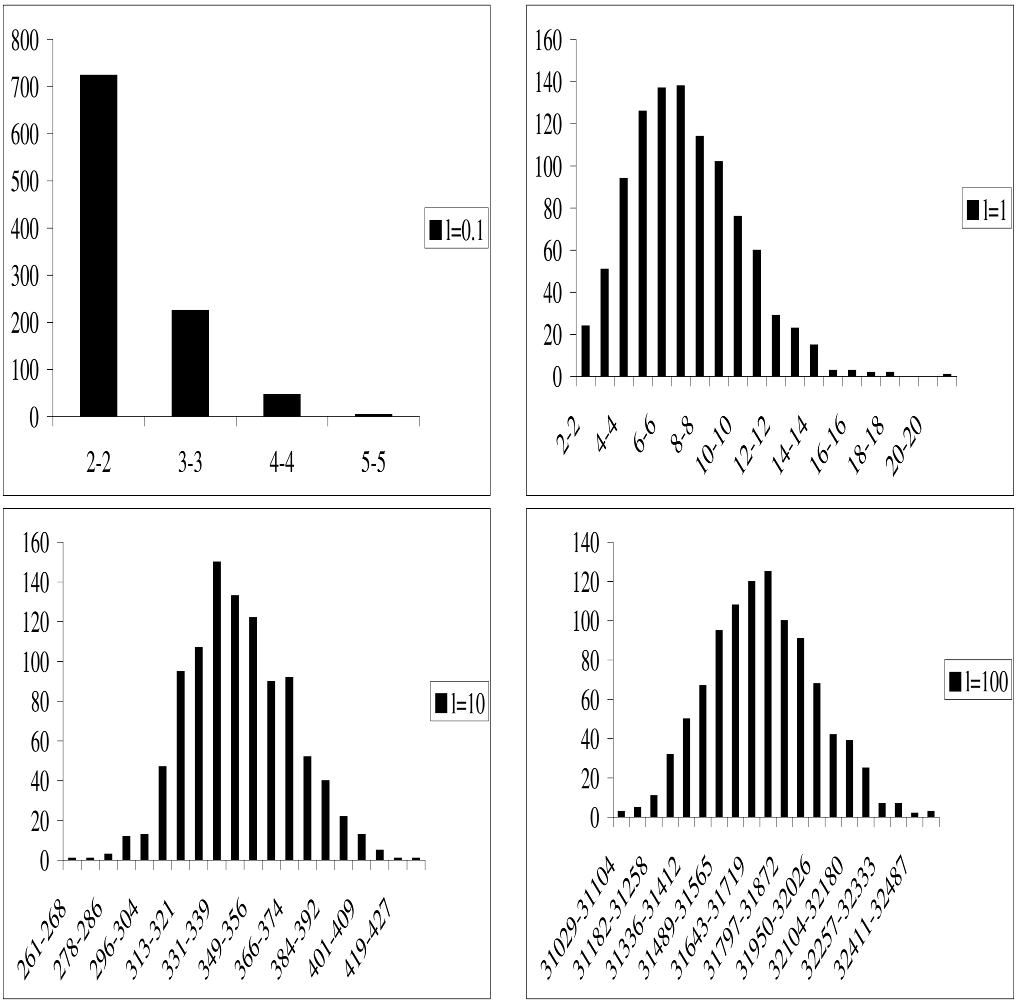

Figure 5: Histograms of the distribution of M corresponding to 1000 independent simulations with $\rho = 1$ and $l = 0.1, 1, 10, 100$.