Location Uncertainty in Mobile Networks: a theoretical framework *

Christopher Rose & Roy Yates

November 26, 1996

Department of Electrical and Computer Engineering
WINLAB
Rutgers University
PO Box 909
Piscataway NJ 08855

ABSTRACT

As users, services, databases and computers become increasingly mobile, so fades the era of the *fixed network*. Modern networks are becoming *mobile networks* which must accommodate a broad range services with differing mobility characteristics. Consequently, there is an impetus to understand mobility and its effect on communications systems. Of particular interest are the unique stresses imposed by mobile computing and especially mobile computer programs (agents).

As an aid to such greater understanding, we propose a theoretical framework for the study of mobility tracking based on user/service/host *location probability distributions.* We show how this methodology, using stochastic ordering and information theory, can enable quantitative comparison of various mobility management schemes as well as insight into the mobility tracking problem over a wide range of mobility characteristics. This general approach should aid both applications and future research.

*This work supported in part by NSF grants NCRI-9206148 and NCRI-9506505.

Contents

1 Introduction

Our highly mobile society and its increasing demand for immediate access to knowledge requires that future information networks gracefully accommodate mobility of both users and services. For example, a particular user might wish to gain network access through any number of different ports or connection media. Likewise, a mobile network service or agent [1–4] might reside on one of many possible processors and could possibly spawn related processes, all of which may need to communicate with other services/agents. Under such a scenario, where both users and network services change location, the distinction between the "fixed" and "mobile" network blurs; all networks are mobile networks.

The overall costs of maintaining accurate location records are at present only poorly understood. However, recent work indicates that simply for telephone traffic, the excess network signaling load expense would be much larger than that required for classical fixed traffic [5–10]. If migrant programs and databases are included, the aggregate signaling load can only be greater. In addition, for wireless systems, the relevant signaling events require use of radio channels and such use must be minimized owing to the scarcity of bandwidth. Thus, either from the standpoint of modifying existing fixed network signaling structures or designing wireless network paging/registration strategies, it is important to understand and quantify the impact of location uncertainty on signaling.

1.1 A Brief Synopsis of Prior Work

First consider the problem of routing a call to a mobile unit U in slightly more detail. We emphasize that by mobile unit it is meant any migrant user, database or program with which communication is desired. A request for connection to U is generated from some point P in the network. Since mobile units change locations, a physical address must be found. The means by which the network obtains information about unit locations is through some combination of *paging* (issuing polling signals in likely locations) and *registration* (the unit identifying it's location). A fair amount of previous work develops paging and registration procedures which minimize some cost function (usually signaling over the wireless and/or fixed network) for simple mobility models or specific systems [11–18].

Another body of work addresses the problem of how to optimally store and update location information throughout the network. Specifically, network resources at point P must have some knowledge of how to route a call to unit U. A number of possibilities exist. For example, should there be a central database which contains the exact location of every mobile unit, should every location have its own local copy of the database, or should locations have partial information and if so, how should it be ordered and updated? This basic problem and its variants are considered in [15, 19–32].

Thus, one body of literature optimizes location information gathering (paging and registration), whereas the other assumes unit locations are known and optimizes dissemination of location information over the network. Both, however, make assumptions about unit mobility either through the use of simple mobility models or by more general intuitive notions such as locality of movement.

1.2 A More General Approach

It is difficult to compare various mobility management schemes since as yet there exist no absolute bounds on optimum performance which can be applied to any procedure regardless of assumptions about mobility, network structure and the like. It would be useful to fill this gap by developing some general theory, which optimally, would precisely relate quantities of interest such as minimum average effort/delay in determining unit locations and minimum average location information dissemination rate over the network in fundamental terms.

We therefore note that in order to quantitatively evaluate the average cost of *any* mobility management scheme, time-varying probability distributions on mobile unit location are required. These distributions might be derived from specific motion models, approximated via empirical data [18, 33] or even provided by the mobile units in the form of a partial itinerary at the time of last contact. Regardless, for quantitative evaluation, the underlying distributions must be either assumed or measured. From a theoretical standpoint, it therefore seems sensible to work *directly* with location probability distributions.

Thus, we identify two distinct subtasks common to all mobility management problems:

- finding the mobile unit
- disseminating location information to the rest of the network

and then seek clearly defined lower bounds on the cost of information gathering (registration/paging) and location information dissemination based on assumed location probability distribution functions and calling frequency patterns. In this paper, we explore some of the ramifications of this approach, and hope that it will both prove useful for practitioners as well as stimulate further quantitative discussion in the research community.

2 Preliminaries

We enumerate possible unit locations by $j = 1, 2, \ldots$. Let us again emphasize that the term mobile unit can refer to a migrant program [1, 2, 4], database or network user. The locations might be cells in a wireless system, nodes/hosts in a communications network or particular wall outlets in a building. Each unit m moves between locations j according to some stochastic process.

Given that the unit was last in contact with the network at time $t = t_0$ and place $x(t_0)$, the time varying distribution on location is defined as $p_j(t, t_0, x(t_0))$. We assume that $p_j(t_0, t_0, x(t_0)) = \delta_{j,x(t_0)}$ where $\delta_{i,j}$ is unity for $i = j$ and zero otherwise. The implication is that successful paging, registration or calls initiated by the mobile unit inform the system of the unit location and essentially reset t_0.

We assume that an effective paging process is rapid compared to the rate of unit motion; i.e., the unit to be found does not change locations during the paging process. Thus, if a page request arrives at time $t \geq t_0$, the distribution is assumed fixed at $p_j(t, t_0, x(t_0)) = p_j$. For simplicity we also assume that a mobile unit can always "hear" and will answer a page request at its current location. These assumptions can be relaxed if desired however [34–36].

Although any model of unit motion could in principle be reduced to its corresponding time-varying probability distribution, as a convenient, useful and concrete example we consider an isotropic brownian motion process with drift [37]. Drift is defined as mean velocity in a given direction and can be used to model directed traffic such as vehicles along a highway. With zero mean velocity we model the possibly more aimless meanderings of a shopper at a flea market.

In the one-dimensional version of brownian motion, a particle moves to the right by one "space step" Δx with some probability p and to the left with probability q, and stays put with probability $1-p-q$ for each "time step" Δt. The simplest analogy is either pedestrian or vehicular traffic in a Manhattan street grid (confined to one avenue in the one-dimensional case). Given the particle started at time $\tau = 0$ from position $x = 0$, as the time and space steps are made infinitesimally small we obtain a Gaussian probability density function on particle location:

$$p_{X(\tau)}(x(\tau)) = \frac{1}{\sqrt{\pi \mathcal{D} \tau}} \mathrm{e}^{-(x - v\tau)^2/\mathcal{D}\tau} \tag{1}$$

where $v = (p-q)\Delta x/\Delta t$ is the drift velocity and $\mathcal{D} = 2((1-p)p + (1-q)q + 2pq)(\Delta x)^2/\Delta t$ is the diffusion constant, both functions of the relative values of time and space steps [38]. Note that equally likely steps left or right implies $p = q$ and therefore $v = 0$. The location probability p_j would be obtained by integrating the density function over the region associated with location j.

Similarly, for n-dimensional isotropic motion we have

$$p_{\mathbf{x}(\tau)}(\mathbf{x}(\tau)) = \prod_{i=1}^{n} \frac{1}{\sqrt{\pi \mathcal{D} \tau}^{1/2}} \mathrm{e}^{-(x_i - v_i \tau)^2/\mathcal{D}\tau} = \frac{1}{(\pi \mathcal{D} \tau)^{n/2}} \mathrm{e}^{-|\mathbf{x}-\mathbf{v}\tau|^2/\mathcal{D}\tau} \tag{2}$$

Of particular interest is the fact that the variance of the distributions $\mathcal{D}\tau/2$, a measure of location uncertainty, *does not depend on the mean drift velocity* $\mathbf{v}$. This is an important feature since intuitively, and in some of the literature, higher unit velocity implies higher mobility. There is no such relationship for the probability distribution-based method proposed here – we consider only location uncertainty.

As a specific example, consider a pedestrian in Manhattan who covers 80 meters (one city block) per minute and proceeds randomly in one of 4 directions at each corner; a two dimensional diffusion process. In each dimension of motion we have $p = q = 1/4$ so that $\mathcal{D} = (\Delta x)^2/2\Delta t$. Substituting $\Delta x = 80m$ and $\Delta t = 60s$ we have $\mathcal{D}_{\text{ped}} = 53.\overline{333}m^2/s$. If a taxi covers the same city block in 6 seconds and also makes random turns at each corner we use the same argument to obtain $\mathcal{D}_{\text{taxi}} = 5.33 \times 10^2 m^2/2$.

Similarly, consider highly directed and planned motion such as that of an aircraft or commuter train [39]. If these conveyances followed exact courses at constant speed, then the diffusivities of would be identically zero – the departure time and current time would completely determine location. In reality, the average arrival time is a random variable for a trip of any reasonable duration. We therefore try to infer the parameters of the equivalent diffusion model from the trip distance, mean duration and arrival time variation.

Consider a train traveling a distance D non-stop with an average trip duration of T for an average speed of $\bar{V} = D/T$. Let the standard deviation in arrival time be δt. Ignoring a number of details about the motion model we roughly approximate the position standard deviation at $t = T$ as $\bar{V}\delta t$. For motion in 1 dimension, the diffusion process variance at time T is $\mathcal{D}T/2$. Equating these two quantities and solving for $\mathcal{D}$ yields,

$$\mathcal{D} = 2(\bar{V}\delta t)^2/T$$

Using the distance between Boston and New York City (approximately $D = 320km$), average trip duration $T = 2.5$hr and an optimistic arrival standard deviation of $\delta t = 10$ minutes, we obtain $\mathcal{D}_{\text{train}} = 1.01 \times 10^5 m^2/s$. For the corresponding shuttle flight with $T = 1$hr we obtain $\mathcal{D}_{\text{air}} = 1.58 \times 10^6 m^2/s$. So, although mean velocity does not figure *directly* into location uncertainty via the diffusion model, we conclude empirically that the practical rapid motion might engender high diffusivity.

3 Hunting Users: signaling and delay costs of paging

3.1 The Basic Problems of Paging

Since paging amounts to issuing queries about unit location and these queries require signaling messages, three obvious questions arise:

> *How much information does the network need to resolve a mobile unit location at a given point in time?*
>
> *What is the relationship of this information to the amount of signaling required?*
>
> *What is the tradeoff between delay in finding a user and the signaling effort?*

The obvious simple answer to the first question is *the entropy of the location distribution.* Likewise, considering that locations vary with time, we see that the average information *rate* necessary to completely specify unit location for all time is the entropy rate of the motion process [40, 41]. However, the structure of allowable queries about location profoundly affects the relationship between information content and the necessary signaling in the paging problem.

For example, consider the game of "guess my number" where a number n is chosen between 1 and J according to some probability distribution p_j [42, 43]. Using the standard information theoretic formulation, the minimum average number of yes/no questions necessary to identify n is the entropy of the probability distribution. However, in a mobile communications system, each polling event at a given location requires signaling in the network, and possibly use of a radio channel as well in the case of a wireless system. Thus, from a signaling standpoint, the appropriate queries are of the form *is your number (location) k?*.

3.1.1 Stochastic Ordering and Paging

Methods based on the ordered location probability distribution [18] and its stochastic order [38, 44] have been developed previously. The *ordered* distribution is simply the original distribution with its elements arranged in decreasing order of probability. Stochastic order is defined in terms of the complementary CDF of the ordered distribution. This approach seems promising since it allows the calculation of bounds on minimum signaling and delay.

For example, using stochastic ordering arguments we have shown that the mean of the ordered distribution, $E[L]$, is *exactly* the minimum number of location queries necessary to find a single mobile unit. The basic principle is simple; more probable locations (or groups of locations) should be searched before less likely ones. The general concept is explained pictorially in FIGURES 1 through 3.

We have also shown that among distributions over a finite number of locations, the uniform location distribution requires the max-min amount of signaling and delay: intriguingly analogous to the information theoretic result for uniform distributions. For comparison, a uniform distribution requires on average a minimum of $(N+1)/2$ signaling queries to resolve unit location whereas the *information* conveyed on average is still $\log_2 N$ bits. The question of how these two quantities might be related prompts further inquiry.

3.1.2 Difficulties for Information Theory and Paging

With $E[L]$ the mean of the ordered distribution used as a shorthand notation for the minimum mean number of locations searched, we note that the uniform location distribution maximizes $E[L]$. This begs a comparison to information theory. Since information theory is the *lingua franca* of uncertainty, there exists a large body of results which might be applied if location distribution entropy, H, is closely related to $E[L]$.

We first note that increased stochastic order "broadens" the probability distribution. Thus, we might suspect that increased stochastic order implies increased entropy. This hypothesis is true and stated formally in the following theorem.[1]

Theorem 1 *If* $\mathbf{p}$ *and* $\mathbf{q}$ *are distributions on location random variables* X *and* Y *with* $X \geq_{\text{st}} Y$*, then* $H(X) \geq H(Y)$.

Unfortunately, an information theoretic approach to paging has a glaring deficiency in that the entropy and the mean number of locations polled can be vastly different:

Theorem 2 *Some distributions can have finite* $H(X)$ *but infinite* $E[L]$.

Thus, entropy is not tightly related to the mean number of locations searched although bounds have been derived for the two quantities in the context of sequential decoding (see Massey or Arikan, [42, 43]). Therefore, entropy is probably not the right tool for measuring signaling cost due to paging.

[1]Proof of all theorems are given in Appendix A for clarity.

3.2 Delay and Single-unit Paging

Since searching one location at a time maximizes delay, we have also considered arranging locations into groups and polling elements of a group simultaneously (and thence the groups sequentially). The same search principles apply in deciding membership of the groups. That is, the first group, whatever its size, should contain the most likely elements. The next group should contain the next most likely elements and so forth. The problem of finding optimal group sizes is a dynamic program [18], but a continuous approximation can also be employed [38] to obtain analytic results for a variety delay constraint types.

In FIGURE 4 we reproduce a figure from [38] which relates signaling cost $E[L]$ (denoted by L^* on the ordinate) achieved by optimal group sizes for a given fixed average number of groups polled D^* (polling delay). The worst case uniform distribution is assumed since for non-uniform distributions over the same number of locations, both signaling cost and delay must decrease: a consequence of the uniform distribution having maximum stochastic order [38]. The result shown is typical for a relatively broad range of problems involving delay and paging effort.

The most important point of FIGURE 4 is that near-optimum $E[L]$ can be obtained even under relatively severe constraints on D^*. Specifically, for the uniform distribution of 20 elements, the unconstrained minimum $E[L]$ is 10.5. However, even when a mean polling delay of $D^* = 2$ is required, $E[L] = 13$ can still be achieved.

Furthermore, we have also shown previously that these results *scale* so that the relative $E[L]$ remains virtually constant under fixed delay constraints *regardless of the actual number of locations*. For a uniform distribution with 200 elements we can expect through application of the scaling theorem [38] that $E[L] = 130$ with $D^* = 2$. Since the absolute minimum $E[L]$ for this distribution is 100.5, the relative $E[L]$ of the absolute minima for the 20 and 200 element cases are comparable.

3.3 $E[L]$ and $\mathcal{D}$

In [45] it was shown that for a diffusion process, the cost of paging is proportional to $(\mathcal{D}\tau)^{n/2}$ where τ is the time elapsed since last contact with the mobile unit and n is the dimensionality of the unit motion. Once again notice that this cost is *independent* of the mean unit velocity. Also notice that this cost is in units of length whereas we have stressed that signaling cost is the unitless $E[L]$.

The translation between physical motion and paging effort is accomplished through a *location density* – the number of pollable locations per unit length. For example, a normal cellular system might have a density of one polling location per kilometer whereas a LAN could have a density of one polling location per 10 meters. For processes whose mobility is defined in terms of the network connectivity such as mobile software agents, no translation is needed since physical space is subordinated to the abstraction of a pollable location.

This distinction, however, will be virtually ignored in what follows since the relevant cost units are obvious by context.

3.4 Delay and Multi-unit Paging

Suppose that many units must be found over J possible locations. Assume that each location can be polled independently of the others and that the polling rate at each location is fixed.[2] Only a single unit can be polled per location during a polling interval. A block diagram of the procedure is shown in Figure 5.

Our goal is to find polling sequence policies which minimize average delay in finding mobile units. Unfortunately, optimal solutions appear difficult to obtain in general, so we seek basic governing principles which will lead to good solutions. To this end we have shown that the basic idea of minimizing $E[L]$ for single units can be applied in a suboptimal way to obtain good multi-unit paging policies [46, 47].

Specifically, in response to an incoming call, polling requests can be issued sequentially in decreasing order of location probability for each unit until the unit is found. This sequential method assures that the absolute minimum average number of locations are searched per mobile unit. As such, this sequential discipline results in lower load on the paging system than the usual "blanket" paging technique in which polling requests are issued simultaneously to *all* locations in the system for each incoming call. Assuming fixed polling rate (paging bandwidth) at each location, sequential polling allows larger incoming call rates to be sustained than for blanket paging. This result is shown in FIGURE 6. Furthermore, if the unit location distributions are non-uniform, $E[L]$ is decreased which further reduces the paging load.

Of course as mentioned earlier, strictly sequential search maximizes delay. Thus, issuing polling requests in groups could be employed to reduce delay. This tension between delay reduction and paging load reduction leads to a tradeoff between polling effort $E[L]$ and delay. To explore this tradeoff, locations were partitioned into approximately fixed size groups and the delay calculated as a function of load for various group sizes. For low paging load, the optimal group size is the entire set of locations; blanket paging. For increasing load, the groups become smaller until groups consist of only single elements. It can thus be inferred that for any given load, there exists

[2] The latter point is a consequence of assumed fixed paging/signaling bandwidth.

an optimal group size which achieves minimum delay.

4 Cooperative Users: registration & paging

4.1 Classical Methods

Classically, registration strategies have used the concept of a *location area*. Location areas are groups of (usually) contiguous locations. These areas can be global in the sense that they are identical for all mobile units, or personal in that each unit has its own location area. The two concepts are illustrated in FIGURES 7 and 8.

Under the classical scenario, a service area is partitioned into groups of locations. Incoming calls result in page requests at all locations in the appropriate location area. When a mobile unit crosses a group boundary, a registration is mandated. Location area boundaries are chosen based on incoming call rates, aggregate mobility patterns and the relative cost of paging and registration.

Unfortunately the simple classical strategy suffers from spurious registrations since mobile units which dwell near boundaries can "ping pong" between location areas. Furthermore, since mobile units may have different mobility characteristics, designing location areas for the aggregate can be inefficient.

A key element of the registration problem is the rate at which calls arrive to the mobile unit. If calls never arrive, then the unit need never register since paging cost is always zero. Likewise, if calls arrive regularly then registration might not often be necessary because the network has reasonably accurate location information. Thus, the procedure used depends both upon the mobility process and the incoming call rate as well as the relative cost of registration.

Using these ideas, the concept of personal location areas has been proposed [48,49] where each mobile unit contact with the system results in the calculation of a new location area, centered around the current location. Since the boundaries are reset after each registration, this method does not suffer from spurious location updates, and since location areas can be designed for each user, the problem of designing for the aggregate does not exist.

However, as with the classical scheme, all locations in a location area are paged in response to an incoming call even if the call arrives shortly after a previous call has terminated. Furthermore, since the boundaries of these location areas are fixed, high velocity mobile units must register more frequently or require larger location areas, both of which tend to increase cost.

4.2 Paging/Registration Using Optimal Paging and the Mobility Index

As mentioned previously, minimum average paging cost $E[L]$ can be approximated by $(\mathcal{D}\tau)^{n/2}$ [45] where $\mathcal{D}$ is the diffusion constant, τ is the time since last sighting of the unit, and n is the dimensionality of motion. It is useful to further define a *mobility index* as $\rho = \mathcal{D}/\lambda_p$. The inverse of ρ is known as the *call to mobility ratio* [21, 31]. The paging cost can then be written as $(\rho t)^{n/2}$ where the new variable t, the time since last sighting, is measured in units of average call interarrival times $1/\lambda_p$. The utility of the mobility index lies in its explicit inclusion of the frequency with which a unit is contacted. For example, if the unit is often paged then its whereabouts are better known to the system, thereby decreasing the unit mobility index.

These basic notions were applied in [45, 50, 51] where registration strategies were derived based on time and/or place. It is assumed that the mobile unit knows the paging method employed by the network and thereby knows the cost (or expected cost) to be incurred at the current point in (time, place). Assuming Poisson arrivals of paging requests makes for a relatively straightforward optimization.

When only time information is used by the mobile unit, a deadline registration policy is appropriate. That is, the mobile unit should register at time τ^* after the time of last contact with the network where τ^* depends upon the mobility index and the relative costs of registration and paging. Intervening call initiations or received calls reset the timer. With both time and place information, however, optimality is an open question.

Application of the timer-only registration method leads to an improvement over simpler place-based methods [48] especially at higher mobile unit velocities since the optimal paging algorithm is affected only by location uncertainty and not directly by the unit velocity. Suboptimal registration methods based on time *and* place were explored in [51, 52], and similar to an optimal method based on distance from last sighting derived for the random walk with drift velocity zero [17], threshold rules for registration were obtained. Typical registration regions are illustrated in FIGURE 9. It was found that the (time, place)-based method performed only slightly better than the timer-alone method over a range of conditions. The primary improvement afforded by spatial information was a reduction in cost variance.

5 Getting the Word Out: location information dissemination

We have so far considered only the effort associated with finding a given mobile unit through paging and registration. Here we take a slightly different view and assume that the location of every unit is known somewhere. Now we ask how much signaling effort is necessary to efficiently disseminate the location information over the network.

Consider then the Universal Phone Number (UPN) where a single number is used to identify each mobile unit and route the call appropriately. At the heart of the UPN problem is a question similar to that which arose in the paging/registration context:

Who needs to know?

That is, the frequency with which a given UPN routing entry is used is the primary index of the importance of maintaining its accuracy.

Let us assume that the answer to this question is given by the rate, r_{ij}, at which unit i calls unit j. This concept of calling rate, coupled to a view of mobility as a set of time-varying location distributions, allows a simple lower bound on the amount of information which must be disseminated over the network.

Specifically, let a random variable τ_{ij} describe a renewal process for the time between calls from unit i to unit j. We assume that τ_{ij} has a density function $f_{\tau_{ij}}(t)$ with mean $1/r_{ij}$. At the initiation of a call at time t, from the perspective of unit i, the location uncertainty of unit j is given by the location probability distribution $p_j(t, t_0, x_j(t_0))$ where $x(t_0)$ was the position of unit j at time t_0.[3] In information theoretic terms, however, the entropy of this distribution represents the average amount of information required to exactly specify the location of unit j to unit i at time t.

Define $H_{ij}(t)$ with $t \geq t_0$ as the entropy in bits of the location distribution $p_j(t, t_0, x_j(t_0))$. The average number of bits needed by unit i to specify the location of unit j just before the next call is,

$$\bar{H}_{ij} = \int_0^{\infty} f_{\tau_{ij}}(t') H_{ij}(t') \, dt' \tag{3}$$

and by renewal theory [53], the *absolute minimum* average number of bits per second needed by unit i to determine the location of unit j is simply $\bar{H}_{ij}/E[\tau_{ij}] = r_{ij}\bar{H}_{ij}$

The minimum necessary aggregate network signaling load can then be determined by considering the average number of hops [54]. That is, location information from

[3] Paging and registration are done by the *system* as opposed to the caller. Thus, the only *a priori* information available to the caller is the location distribution $p_j(t, t_0, x_j(t_0))$ which is independent of the registration/paging method used to track user j.

unit j must be routed to unit i and must traverse some number of nodes. As a simple example, M mobile units. With uniformly distributed mobile unit locations we define γ as the mean number of hops from any node to the rest of the network. Therefore, the aggregate signaling rate associated with location information dissemination is

$$\mathcal{R} = \gamma \sum_{ij} r_{ij} \bar{H}_{ij} \tag{4}$$

The generality of this approach allows connection rates between *any* two entities to be defined; i.e., not only "mobile units" but possibly routing tables as well. Thus, we believe this basic method can be extended to derive lower bounds for signaling costs associated with any mobility management scheme on any given network.

Note that for a particular signaling system, the average number of bits per second of location information actually transmitted to unit i may be significantly greater than the calculated lower bound because the signaling system may support only inefficient signaling messages. Thus, the lower bound provides a basis upon which to judge the efficiency of a particular signaling system in conveying location information.

6 Mobility Management and the Future

When considering mobility in present-day communications systems, what usually comes to mind is a person, conveyance or mobile computer. However, in future networks, programs as well as physical objects might also be mobile. For example, suppose you are an investment banker and seek financial market information distributed over the network. Specifically, you might wish to exploit small price differences over numerous "cyber-markets" – analogous to but more extensive than current arbitrage practices.

If the processing and communication capacity of the machines and network were infinite or there were no competing users, it would be a relatively simple task to spawn search processes on all machines in the network and have them report back information in some fashion. However, capacity constraints allow only a small number of programs to be launched into the network and run on other machines. For efficient search or to execute trades, the programs should be able to communicate and modify their behavior as information is gathered. For effective search, the programs should be able to relocate themselves to appropriate databases or markets. Such migrant programs charged with ferreting out information are generically called mobile agents [1–4].

The need for communication between agents implies a need for registration and paging since the agents are effectively mobile units to which calls may be routed. In motion processes which involve movement of mass, there are inherent constraints on motion between physical locations. Mobile agents, however, suffer few such constraints since their motion processes are influenced primarily by the location of contacts relevant to the search. Thus, one instant an agent could be active at a host in Los Angeles, and when done relocate to a suggested host in Madras.

For a simple random walk in n dimensions, paging cost is proportional to $(\rho t)^{n/2}$ where once again, ρ is the mobility index, t is the time elapsed since last contact with the system (measured in units of intercall arrival). The intrinsic lack of constraints on both mobility index (how fast and far) and motion dimensionality (how many choices) for mobile agents suggests that groups of intercommunicating agents moving rapidly over a network of effective dimension $n \geq 3$ could severely stress signaling resources. Under this not-too-futuristic scenario, efficient mobility management could easily become the principal issue in network design.

7 Discussion & Conclusions

The concept of mobility management based on time-varying mobile unit location probability distributions was introduced. Mobile units could be cellular telephone users, mobile computers or even mobile computer programs such as mobile agents. The approach uses stochastic ordering of the distributions to rank the relative *minimum* difficulty in finding units and thereby provides a common framework for comparison across mobility management schemes.

Since location uncertainty is at the heart of the mobility management problem and information theory is the lingua franca of uncertainty, it was tempting to apply information theory to the paging problem. We found that information theory, and in particular entropy arguments, can be highly misleading about the amount of effort which must be expended to find a mobile unit. It therefore seems that stochastic ordering and the mean of the ordered distribution, $E[L]$, provide the proper theoretical context for the paging problem.

We outlined various simple paging/registration procedures based on this methodology, compared them to more classical methods. A byproduct of this study was the definition of a *mobility index* which is a useful reification of mobile unit location uncertainty and its growth as a function of time since last contact. We then showed how paging/registration cost varies as a function of mobility index.

We also showed how this cost varies with the dimensionality of the motion process which led to a consideration of non-classical mobility: groups of mobile programs ranging over a network in a coordinated way seeking out information. One could imagine agents which roam the network for information [4] or to buy and sell goods [2]. The results suggest that severe stress could be placed on a communication network by widespread use of such programs.

In the networks context, we then suggested ways in which time-varying location probability distributions for mobile units might be used to underbound the amount of signaling traffic necessary for distributing location information. In this case, an information theoretic approach *is* helpful.

Finally we reiterate the basic contention that mobility management problems should be couched directly in terms of unit location probability distribution and how they vary with time. Given a stochastic motion process, in principle the associated time-varying probability distribution function should be derived, and then the minimum amount of signaling effort necessary to track the mobile unit should be calculated. Unfortunately, closed form solutions for minimum signaling effort can be difficult to derive for arbitrary motion processes. Thus, rather than provide a fragmentary treatment of specific mobility models, we have illustrated the general approach using a diffusion process with its intuitively appealing natural formulation.

It must be stressed, however, that we do not propose diffusion as THE fundamental mobility model in the sense that the Gaussian distribution is fundamental in communications theory. Rather, we suggest that the diffusion's hueristic appeal – location uncertainty increase with time and spatial dimension – and its analytic tractability make it a useful surrogate for other more accurate models of unit mobility.

8 Acknowledgements

The authors would like to thank the two anonymous reviewers for their comments which helped us more carefully hone the presentation. In addition we thank David J. Goodman here at Rutgers/WINLAB for first bringing the mobility management problem to our attention.

A APPENDIX: Proofs

Proof: Theorem 1 The Kullback-Liebler distance measure [40] between two distributions is defined as

$$D(\mathbf{y}||\mathbf{x}) = \sum_{i=1}^{\infty} q_i \log \frac{q_i}{p_i} \geq 0 \tag{5}$$

which can be rewritten as

$$\sum_{i=1}^{\infty} q_i \log \frac{1}{p_i} - H(Y) \geq 0 \tag{6}$$

or

$$E_{\mathbf{y}}[\log \frac{1}{p_i}] - H(Y) \geq 0 \tag{7}$$

Remembering that p_i is a location probability distribution, we have $\log(1/p_i)$ is an increasing function of i. Since $X \geq_{\text{st}} Y$, we must then have

$$H(X) = E_{\mathbf{x}}[\log \frac{1}{p_i}] \geq E_{\mathbf{y}}[\log \frac{1}{p_i}] \geq H(Y) \tag{8}$$

thus proving the theorem.

♠

Proof: Theorem 2 Let $p_i = (1 - p_1)/N$, $i = 2, 3, ..., N + 1$ and $p_1 \geq p_2$ (see FIGURE 10 for example).

$$H(X) = p_1 \log \frac{1}{p_1} + (1 - p_1) \log \frac{N}{1 - p_1} = H_b(p_1) + (1 - p_1) \log N \tag{9}$$

With no loss of generality assume $\log_2$ so that $0 \leq H_b() \leq 1$. We can then write

$$\frac{H(X) - 1}{\log N} \leq 1 - p_1 \leq \frac{H(X)}{\log N} \tag{10}$$

Fix $H(X) - 1 = \kappa > 0$. Calculating the mean of this ordered distribution, $E[L]$, we obtain:

$$E[L] = \sum_{i=1}^{\infty} i p_i = p_1 + (1 - p_1)\frac{N+3}{2} \geq p_1 + (H(X) - 1)\frac{N+3}{2 \log N} \tag{11}$$

so that

$$\lim_{N \to \infty} E[L] = \infty \tag{12}$$

♠

References

[1] P. Maes. Agents that Reduce Work and Information Overload. *Communications of the ACM*, 37(7):31–40, 1994.

[2] A. Chavez and P. Maes. Kasbah: An Agent Marketplace for Buying and Selling Goods. In *First Interantional Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*, London, UK, 1996.

[3] P. Maes. Intelligent Programs. *Scientific American*, 273(3):84–88, 1995. September.

[4] H. Lieberman. Letizia: An Agent That Assists Web Browsing. In *1995 International Joint Conference on Aritifical Intelligence*, Montreal, CA, 1995.

[5] K. Meier-Hellstern, E. Alonso, and D. O'Neill. The use of SS7 and GSM to support high density personal communications. In *Wireless Communications: Future Directions*. Kluwer Academic, 1993. J.M. Holtzman and D.J. Goodman eds.

[6] M.J. Beller, E.H. Lipper, and M.P. Rumsewicz. Switching system impacts of PCS traffic. *2nd International Conference on Universal Personal Communications*, Oct. 1993. Ottawa, Canada.

[7] G. Columbo, L. DeMartino, C. Eynard, and L. Gabrielli. Mobility load control in future personal communications networks. *2nd International Conference on Universal Personal Communications*, Oct. 1993. Ottawa, Canada.

[8] C.N. Lo, S. Mohan, and R.S. Wolff. An estimate of network database transaction volumes to support voice and data personal communications services. *Proc. of the 8th ITC Specialist Seminar on Universal Personal Communications*, Oct. 1992. Santa Margherita, Italy.

[9] E.H. Lipper and M.P. Rumsewicz. Teletraffic considerations for widespread deployment of PCS. *IEEE Networks, Special Issue on Nomadic Personal Communications*, Sept./Oct. 1994.

[10] E.H. Lipper. Switching system performance problems for universal personal communications. *Computer Networks and ISDN Systems*, 1995. (P. Enslow, Editor-in-Chief).

[11] H. Xie, S. Tabbane, and D. J. Goodman. Dynamic location area management and performance analysis. In *Proc. of IEEE Vehicular Technology Conference*, 1993.

[12] R. Thomas, H. Gilbert, and G. Mazziotto. Influence of the movement of the mobile station on the performance of a radio cellular network. In *Proc. 3rd Nordic Seminar*, September 1988. Paper 9.4.

[13] S. Okasaka, S. Onoe, S. Yasuda, and A. Maebara. A New Location Updating Method for Digital Cellular Systems. In *Proc. IEEE Vehicular Technology Conference, VTC'91*, pages 345–350, May 1991. Denver, CO.

[14] A. Bar-Noy and I. Kessler. Tracking mobile users in wireless networks. In *INFOCOM'93*, pages 1232–1239, March 1993. San Francisco, CA.

[15] A. Bar-Noy, I. Kessler, and M. Sidi. To Update or Not to Update? *Wireless Networks*, 1(1), 1995. Preprint.

[16] I. Kessler A. Bar-Noy and M. Sidi. Tracking strategies in wireless networks. In *Proceedings, IEEE Int. Symp on IT*, page 413, June 1993. Trondheim, Norway.

[17] U. Madhow, M.L. Honig, and K. Steiglitz. Optimization of Wireless Resources for Personal Communications Mobility Tracking. *IEEE Transactions on Networking*, 3(6):698–707, December 1995.

[18] S. Madhavapeddy, K. Basu, and A. Roberts. Adaptive paging algorithms for cellular systems. *Fifth WINLAB Workshop on Third Generation Wireless Information Networks*, April 1995. New Brunswick, NJ.

[19] B. Awerbuch and D. Peleg. Concurrent Online Tracking of Mobile Users. In *Proc. ACM SIGCOMM Symposium on Communication, Architecture and Protocols*, October 1991.

[20] J.Z. Wang. A Fully Distributed Location Registration Strategy for Universal Peronsal Communication Systems. *IEEE J. SAC*, 11(6):850–860, August 1993.

[21] R. Jain, Y.-B. Lin, C. Lo, and S. Mohan. A caching strategy to reduce network impacts of PCS. *IEEE J. on Sel. Areas of Comm.*, 12(8):1434–1444, Oct. 1994.

[22] Y.-B. Lin. Determining the user locations for personal communications services networks. *IEEE Transactions on Vehicular Technology*, 43(3):466–473, Aug. 1994.

[23] T. Blackwell *et al.* Secure short-cut routing for mobile IP. *USENIX Summer 1994 Technical Conference*, June. 1994.

[24] J. Ioannidis and G.Q. Maguire Jr. The design and implementation of a mobile internetworking architecture. In *USENIX Winter 1993 Technical Conference*, pages 489–500, Jan. 1993.

[25] W.A. Simpson. IP mobility support. *Mobile IP Working Group, IETF Internet Draft*, May 1994.

[26] C. Perkins and Y. Rekhter. Optimal routing for mobile hosts using IPs loose source route option. *Internet Draft*, Oct. 1992.

[27] F. Teraoka and M. Tokoro. Host migration transparency in IP networks: The VIP approach. *ACM Computer Communications Rev.*, pages 45–65, Jan. 1993.

[28] T. Imielinski and B. Badrinath. Querying in highly mobile environments. In *18th Proceedings VLDB*, Aug 1992.

[29] S. Rajagopalan. An adaptive location management scheme for mobile internetworking. Master's thesis, Rutgers University, January 1995.

[30] S. Rajagopalan. An adaptive location management scheme for mobile internetworking. Winlab-TR 93, WINLAB, Rutgers University, December 1994.

[31] R. Yates, C. Rose, S. Rajagopalan, and B. Badrinath. Analysis of a mobile-assisted adaptive location management strategy. *ACM Wireless Networks*, 1995. (in press).

[32] G. Pollini and S. Tabbane. The Intelligent Network Signaling and Switching Cost of an Alternative Location Strategy Using Memory. In *Proc. IEEE Vehicular Technology Conference, VTC'93*, May 1993. Seacaucus, NJ.

[33] R. Yates and C. Rose. Preliminary study of mobile cellular data: Tucson. WINLAB-TR 59, WINLAB, Rutgers University, October 1993.

[34] C. Yener, A. & Rose. Paging Strategies for Highly Mobile Users. In *IEEE VTC'96*, April 1996. Atlanta, GA.

[35] M. Verkama. Optimal paging – a search theory approach. In *ICUPC'96*, November 1996.

[36] D.O. Awduche, A. Ganz, and A. Gaylord. An optimal search strategy for mobile stations in wireless networks. In *ICUPC'96*, November 1996.

[37] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, third edition, 1991.

[38] C. Rose and R. Yates. Minimizing the average cost of paging under delay constraints. *ACM Wireless Networks*, 1(2):211–219, 1995.

[39] S.S. Rappaport. Blocking, Hand-Off and Traffic Performance for Cellular Systems with Mixed Platforms. *IEE Proceedings I*, 140(5):389, October 1993.

[40] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 1991.

[41] R.E. Blahut. *Data Principles and Practice of Information Theory.* Addison-Wesley, 1988.

[42] J.L. Massey. Guessing and Entropy. In *IEEE International Symposium on Information Theory*, page 204, Trondheim, Norway, 1994.

[43] E Arikan. An Inequality on Guessing and its Application to Sequential Decoding. *IEEE Transactions on Information Theory*, 42(1):99–105, January 1996.

[44] C. Rose and R. Yates. Paging Cost Minimization Under Delay Constraints. In *IEEE Infocom'95*, pages 490–495, April 1995. Boston, MA.

[45] C. Rose. Minimizing the average cost of paging and registration: A timer-based method. *ACM Wireless Networks*, 2(2):109–116, June 1996.

[46] David Goodman, P. Krishnan, and Binay Sugla. Minimizing Queueing Delays and Number of Messages in Mobile Phone Location. *ACM-NOMAD*, 1(1), 1996. (in press).

[47] C. Rose and R. Yates. Ensemble polling strategies for increased paging capacity in mobile communications networks. *ACM Wireless Networks*, 1996. (in press).

[48] H. Xie, S. Tabbane, and D.J. Goodman. Dynamic location area management and performance analysis. In *Proc. IEEE Vehicular Technology Conference, VTC'93*, May 1993. Seacaucus, NJ.

[49] S. Tabbane. An Alternative Strategy for Location Tracking. *IEEE Journal on Selected Areas in Communications*, 13(5):880–892, June 1995.

[50] C. Rose. Minimization of paging and registration costs through registration deadlines. *IEEE/ICC'95*, pages 735–739, 1995. Seattle.

[51] C. Rose. A Greedy Method of State-Based Registration. *IEEE/ICC'96*, June 1996. Dallas, TX.

[52] C. Rose. State-based paging/registration: A greedy technique. Winlab-TR 92, Rutgers University, December 1994. (submitted, IEEE Transactions on Vehicular Technology, 12/94).

[53] S.M. Ross. *Stochastic Processes.* John Wiley & Sons, New York, 1983.

[54] C. Rose. Mean Internodal Distance in Multihop Store & Forward Networks. *IEEE Transactions on Communications*, 40(8):1310–1318, 1992.

B FIGURES

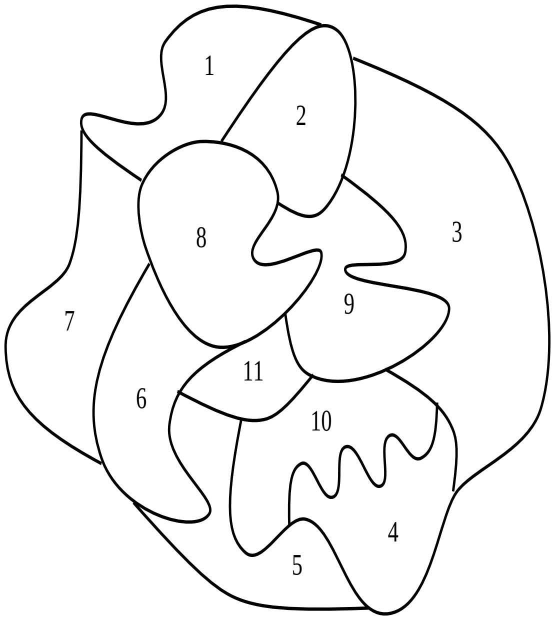

Figure 1: Possible locations associated with a mobile unit.

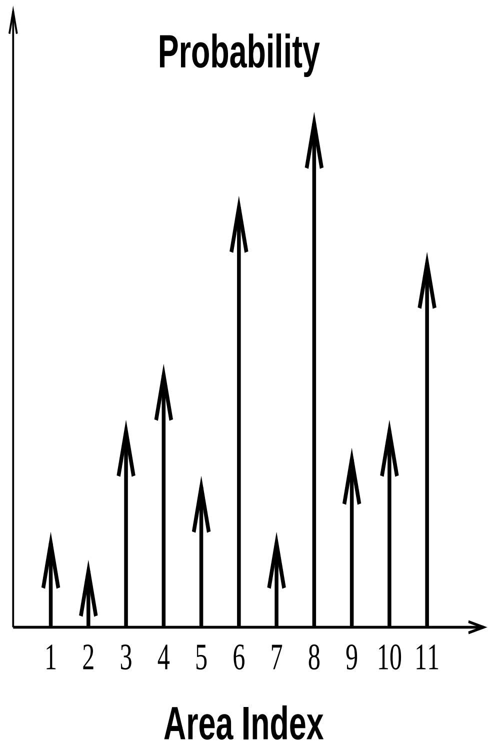

Figure 2: Raw probability distribution on unit location.

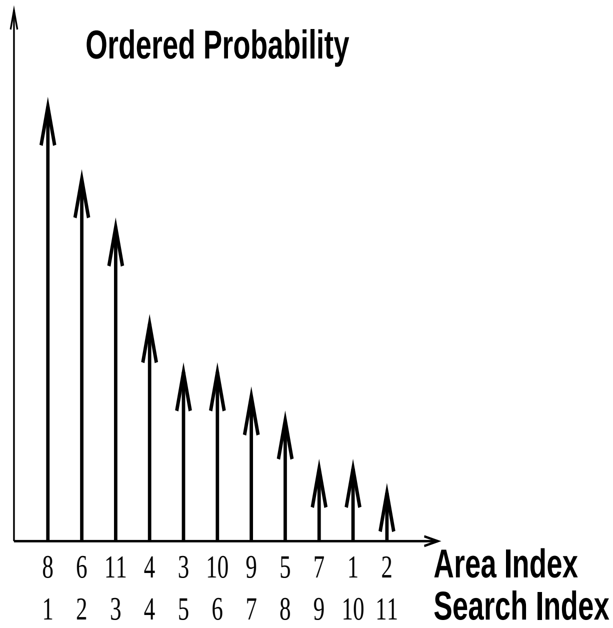

Figure 3: Ordered probability distribution on unit location. The mean of the ordered distribution is the minimum number of locations which must be searched ($E[L]$) to find the mobile unit on average.

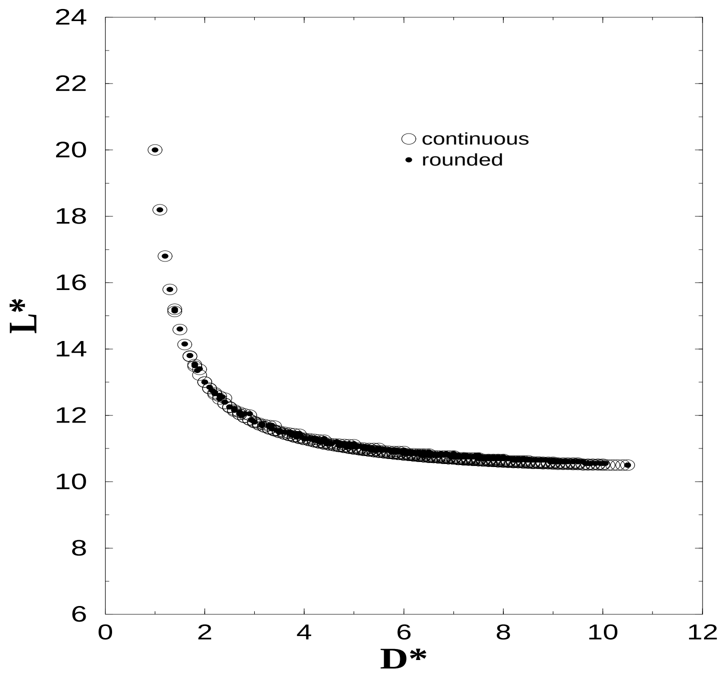

Figure 4: Montage of minimum paging cost $E[L]$ versus fixed mean polling delay D^* for a uniform distribution with 20 elements.

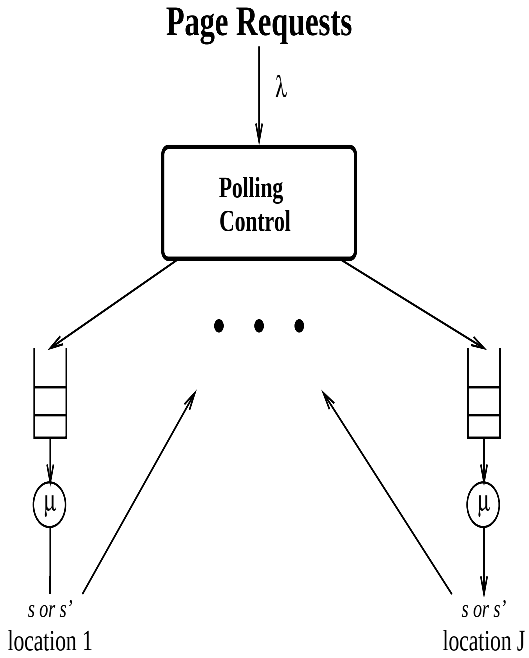

Figure 5: Block diagram of polling system. Page requests arrive according to a Poisson process with rate λ_p, each having an associated probability distribution p_j. Polling Control issues polling requests to appropriate locations. These requests are queued at the locations and serviced FCFS at average rate μ. The result (success: s, failure: s') of completed polls are fed back to the controller for appropriate further action.

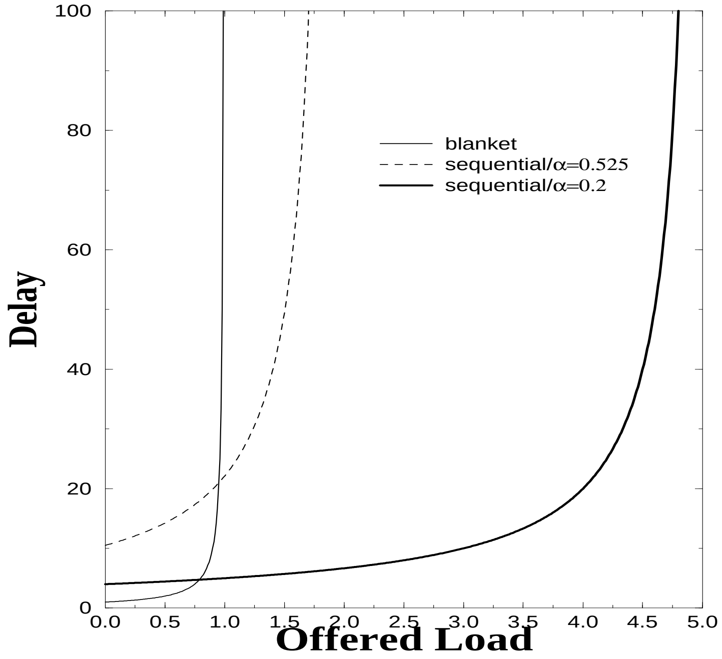

Figure 6: Delay as a function of offered paging load λ_p/μ where μ is the polling rate at each location for sequential and blanket polling. Twenty locations ($J = 20$), uniform ($E[L] = 10.5$, $\alpha = EL/J = 0.525$), and nonuniform ($\alpha = 0.2$) location distributions.

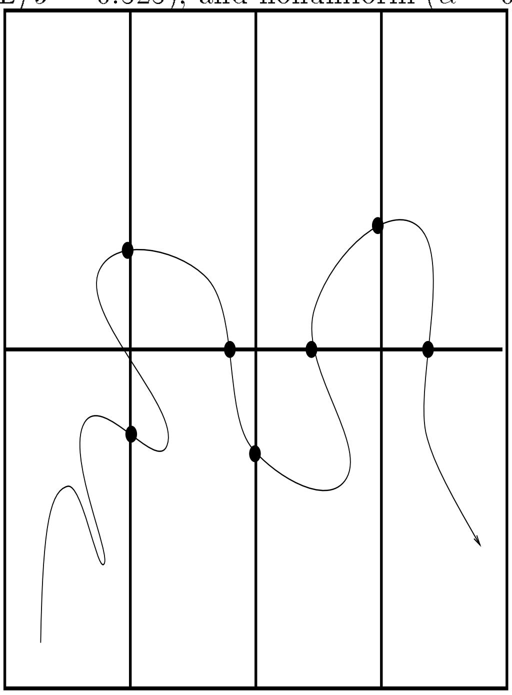

Figure 7: Illustration of classical registration strategy with geographically fixed location areas. Registration occurs at location area boundary crossings (denoted by solid dots). The arrival of an incoming call triggers polling requests over all locations contained in the current location area.

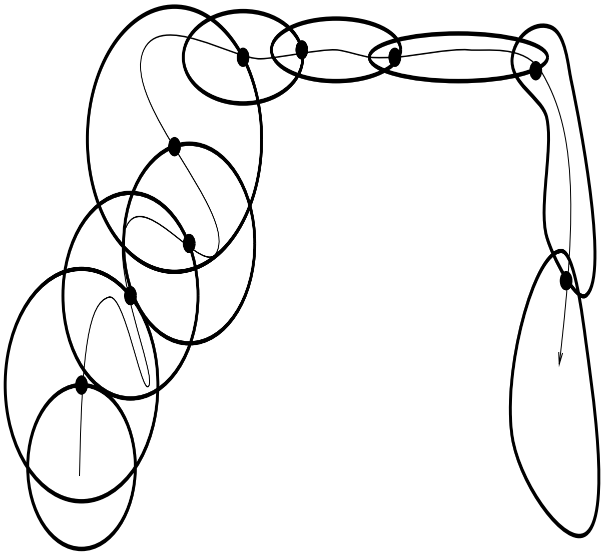

Figure 8: Illustration of personal location areas, recalculated after every contact with system. The location area shape may change with time depending upon the mobility characteristics of the user. For example, the mobile unit in this example might first be driving around a city and then enter a high-speed thoroughfare. The location areas change shape to accommodate the changed mobility. Registration points denoted by solid dots.

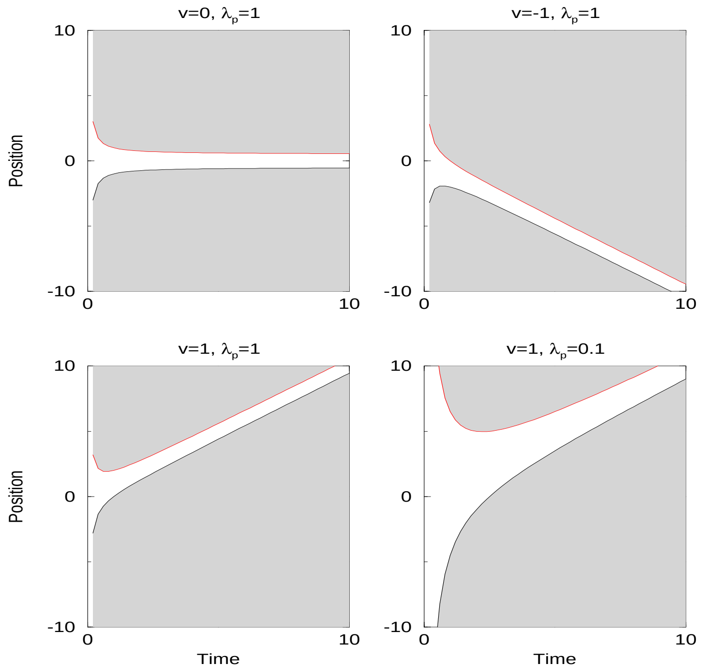

Figure 9: Representative registration regions using the greedy algorithm for various drift velocities v and call arrival rates λ_p. If the mobile position and time correspond to a point in the gray region, registration takes place and the process timer is reset. Notice how the registration region tracks the expected unit position with time. Also notice how decreased incoming call rate widens the range of (time,place) values which may be visited without registration since paging events are less likely.

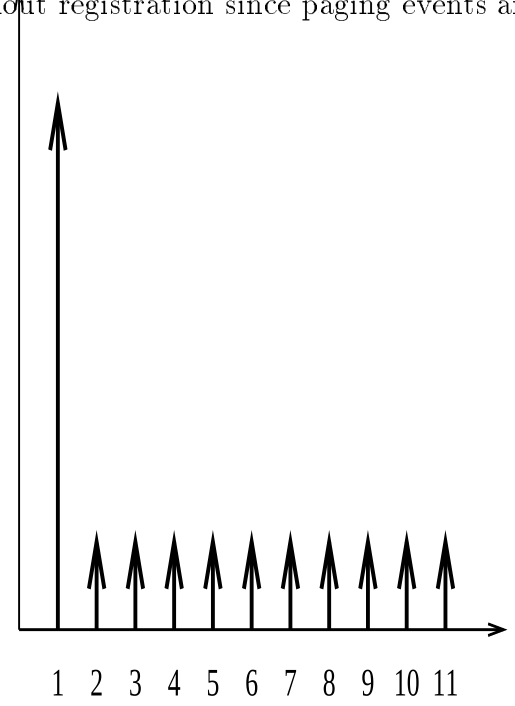

Figure 10: Example distribution for Theorem 2 with $N = 10$.