Modeling and Estimating the Spatial Distribution of Healthcare Workers

ABSTRACT

This paper describes a spatial model for healthcare workers’ location in a large hospital facility. Such models have many applications in healthcare, such as supporting time-and-motion efficiency studies to improve healthcare delivery, or modeling the spread of hospital-acquired infections. We use our model to estimate spatial distributions for healthcare workers in The University of Iowa Hospitals and Clinics (UIHC), a 700-bed comprehensive academic medical center spanning a total of 3.2 million square feet and employing about 8,000 healthcare workers. We model the UIHC as a metric space induced by walking distance between pairs of rooms, and with each room having a level of attractiveness representing the activity level in that room. We combine this with a model in which each healthcare worker has a center of activity and a probability density function that decays polynomially as we move away from the center. Using 12 million Electronic Medical Record (EMR) logins collected over 22 months, we solve for the model parameters for each room and each healthcare worker using heuristic techniques to make the problem computationally tractable. We then validate the model parameters obtained by comparing real-world expectations of healthcare worker behavior for several job categories to our model predictions (e.g., we verify that Unit Clerks are much more stationary than Respiratory Therapists). Finally we present solutions to two important applications. First, using healthcare worker spatial distributions, we find a near-optimal placement of hospital resources (e.g., time clocks) which minimizes the average distance a healthcare worker has to travel to access that resource. Second, we use the healthcare worker spatial distributions to generate random walks representing their movement through the hospital. We use these random walks to simulate healthcare worker contact networks in order to study the spread of hospital-acquired infections.

1. INTRODUCTION

The University of Iowa Hospitals and Clinics (UIHC) is a 3.2 million square foot facility that employs approximately 8,000 healthcare workers, of which about 4,000 are in the hospital at any given time. The UIHC has, at any given time, over 650 in-patients with an average 3-6 days length of stay. In 2007, 740,000 out-patients were treated at the UIHC. Given this scale of operations, there are numerous opportunities to implement new policies and procedures that could lead to improvements in patient outcomes while simultaneously reducing healthcare costs. Attempting to understand and exploit these opportunities can sometimes lead hospital administrators to very specific questions. Two examples that motivate our work are the following. (1) We want to locate 80 time clocks in the UIHC facility for staff to punch in their hours. Where should we locate these in order to minimize overall access time? (2) We have a limited supply of vaccine. Which healthcare workers or which categories of healthcare workers should we target for vaccination? There are numerous other examples: time-and-motion efficiency studies for improving healthcare delivery, optimal placement of incoming patients with certain diagnoses, etc. In this paper, we present a computational approach for solving these and other similar problems.

Different classes of healthcare workers inhabit different parts of the hospital, and exhibit very different levels of mobility and temporal patterns. Administrators might have an office in which they spend most of their time, while In-Patient Nurses will regularly move between a small cluster of patient bedrooms. Residents and Therapists see patients all over the hospital and are highly mobile. Taking cues from work on estimating the spatial distribution of web search queries [BKKN08] and older work from spatial biostatistics [SLR97], our approach utilizes the “electronic footprints” that healthcare workers leave behind in the hospital’s electronic medical record system to estimate their spatial probability distributions within the UIHC facility.
1.1 The Hospital Metric Space and Login Data

We start by imposing a metric space on the UIHC facility that corresponds to the “walking distance” between pairs of hospital rooms. To construct this metric space, we painstakingly created a graph model by manually examining detailed printed architectural drawings of the hospital and matching these drawings with a master spreadsheet of hospital rooms and their utilization, provided by the UIHC. We represented each room in the hospital by a node in the so called hospital graph, and for every pair of rooms in the hospital between which it is possible to move directly (e.g., through a doorway), we created an edge between the corresponding nodes. In order to have a consistent concept of distance, corridors and large rooms (e.g., cafeterias, atriums) are divided into smaller “room-sized” chunks. Traveling along an edge of the hospital graph is roughly equivalent to walking 5-6 meters. For greater flexibility, we view each edge \( \{u, v\} \) as two directed edges \((u, v)\) and \((v, u)\), allowing for distinct weights to be assigned to edges in opposite direction. This helps us in modeling certain natural preferences of people who move about in the hospital, such as using corridors rather then walking through other units (even if that meant a shorter walk) or using an elevator to go up the floors rather than a staircase. The weights we assign to edges are all between 0.8 and 1.6. The hospital graph that results from this construction has 18,961 nodes and 46,884 directed edges and we work with the metric space\(^1\) induced by the shortest path distances on this graph. It is worth pointing out that distances in this metric space maybe quite different from Euclidean distances between rooms (viewing rooms as points in 3-D space). See Figures 1 and 2.

Figure 1: A small portion of the hospital graph corresponding to the second floor of the UIHC. Each room or corridor segment is represented by a vertex, connected by edges to adjacent rooms or corridor segments. This particular image was produced by superimposing the graph onto a CAD drawing of the floor plan.

The data that form the basis for our models and applications are healthcare worker login records to the UIHC electronic medical record (EMR) system. Each login record to the UIHC EMR system contains information on the healthcare worker initiating the login (uniquely anonymized), the ID of the machine being used, the location of the machine, and the start time-stamp and the end time-stamp of the login. The login records also contain the job code and department of each healthcare worker at any given time. After filtering out roughly four million records with missing location information, and four million additional records with ambiguous or unusable location information, we were left with approximately 11.7 million records from a 22 month period beginning September 2006 and ending July 2008. There are 14,596 healthcare workers who login to the EMR at least once during this period using 17,522 different machines distributed over 4,379 locations in the UIHC facility. The machine locations are well spread out in the hospital with the average distance between a room in the UIHC facility and a room with a machine being 2.816 (± 2.258).

1.2 Model and Parameter Estimation

We start with the assumption that every healthcare worker has a geographical center for her activity and that her spatial distribution polynomially decays as we move away from her center, where the rate of decay is determined by two additional parameters. While our model is similar to the model of Backstrom et al. [BKKN08] in these aspects, there are important differences due to the fact that queries and people (specifically, healthcare workers) are after all fundamentally different entities. The same query, for example for the “New York Yankees,” can be issued from multiple locations simultaneously whereas a healthcare worker can be in at most one location in the hospital at any time. Our primary goal here is to estimate a static spatial distribution for each healthcare worker and to do this we suppose that each room \( v \) has an associated parameter \( \alpha_v \) that denotes, in a loose sense, the “attractiveness” of room \( v \) to healthcare workers. Certain UIHC rooms, e.g., the nurse’s station in a unit, may see a lot more traffic than a nearby patient room and such rooms will have a correspondingly higher \( \alpha_v \) value.

Thus for a setting with \( m \) healthcare workers and \( n \) rooms, we need to estimate \( 3m + n \) model parameters. Solving for the maximum likelihood estimators of the model parameters could potentially require solving \( \Omega(mn) \) continuous optimization problems. On the face of it, this is computationally infeasible (since \( m \) is roughly 15,000 and \( n \) is roughly 19,000). In Section 3 we describe heuristics that take advantage of properties of the underlying metric space to prune

\(^1\)Technically, this is not a metric space since distance \( d(u, v) \) may be distinct from distance \( d(v, u) \). However, since the weights are in the range 0.8 to 1.6, symmetry is satisfied approximately and triangle inequality is satisfied exactly.
away from the search space many of the UIHC rooms for each healthcare worker. These heuristics make the model parameter estimation computationally feasible.

1.3 Validation and Applications

We validate our model in several ways. For example, we show that the centers of Pediatricians cluster in 3 areas in the hospital that exactly correspond to where one would expect to see Pediatricians: in the Pediatrics unit, in the Cystic Fibrosis unit, and in the Neonatal Intensive Care Unit. By separately considering login data from night shifts and day shifts, we obtain very different sets of centers of activity that clearly show which clinics and units shut down at night. From the estimated spatial distributions of healthcare workers, we can calculate how dispersed in the UIHC facility each healthcare worker is. These calculations match our expectations quite nicely and correctly predict that job categories such as Unit Clerks are relatively stationary and job categories such as Physical Therapists are highly mobile.

We also present two applications of our model. The first application involves the construction of contact networks for healthcare workers. We model the movement of healthcare workers by random walks in the hospital metric space whose spatial distributions are the static spatial distributions of the healthcare workers. These random walks allow us to simulate healthcare worker contact networks: analysis of these networks shows that they have many of the expected classic “small world” properties [WS98; New03]. Healthcare contact networks are critical for understanding and controlling the spread of disease hospital-acquired infections such as MRSA or VRE. In the second application, we use the healthcare worker spatial distributions to find near optimal placement for k time clocks, where the objective is to minimize the average extra distance a healthcare worker has to travel in order to punch in their hours. This application is a canonical example of the problem of locating resources (e.g., medical equipment, crash carts, pharma stores) in a large hospital might be approached.

2. MODEL

Our goal is to estimate the spatial distribution of each of the 14,596 healthcare workers (HCWs, from now on) that appear at least once in our EMR login database. We associate with each HCW i three parameters: a center $c_i$ (a room in the UIHC facility), a dispersion $\gamma_i$, and a constant $\beta_i$. Fix a time window $T$ (e.g., $T$ might be a week). We assume that for any room $v$ in the UIHC facility the probability that a person chosen randomly from room $v$ at a random time instant in $T$ is the HCW $i$, is proportional to

$$\beta_i \cdot (d(c_i, v))^{-\gamma_i}.$$ 

Here, $d(c_i, v)$ is the distance in the hospital graph between nodes $v$ and $c_i$. For a room $v$, the HCWs who are most likely to be found in $v$ are those whose $\beta$ parameter values are high, whose centers are near $v$, and whose dispersion parameter values (i.e., $\gamma$) are small (implying a spatial probability density function that has not fallen off too much before reaching $v$). While this model tells us for each room $v$ which HCW is more likely to be found at $v$, it does not directly provide spatial distributions for the HCWs.

Here is a simple example (see Figure 3) that clarifies this point. The figure shows a “toy” hospital graph that is populated by two HCWs, labeled 1 and 2, whose centers are nodes $B$ and $D$ respectively. Let $s_i$ be the normalizing factor for HCW $i$, i.e., $s_i$ is a value such that

$$s_i \cdot \beta_i \cdot \sum_v \alpha_v \cdot (d(c_i, v))^{-\gamma_i} = 1.$$ 

Given these spatial distributions for the HCWs, the expected number of HCWs in a room $v$ at any time instant is

$$\alpha_v \cdot \sum_i \beta_i \cdot (d(c_i, v))^{-\gamma_i},$$

where the sum is over all HCWs $i$.

3. ALGORITHM

For a given HCW $i$, we use a maximum likelihood approach to find a center $c_i$, dispersion $\gamma_i$, and constant $\beta_i$. To do this, we first partition the set of EMR logins into two subsets: $L$ consisting of only the logins associated with HCW $i$ and $\overline{L}$ denoting the complement, i.e., the set of logins with which HCW $i$ is not associated. For each login $\ell$, let $\ell$ denote the location in the UIHC facility at which the login occurs. Then the log of the likelihood that the EMR login records are obtained by sampling from HCW $i$’s assumed distribution is:

$$LL(c, \beta, \gamma) := \sum_{\ell \in L} \log (\beta \cdot \text{dist}(c, \ell)^{-\gamma}) + \sum_{\ell \in \overline{L}} \log (1 - \beta \cdot \text{dist}(c, \ell)^{-\gamma}).$$

This quantity is a function of $c$, $\beta$, and $\gamma$ and we denote it by $LL(c, \beta, \gamma)$. Our problem is to find values of $c$, $\beta$, and $\gamma$ that maximize this function. These maximizing values are the maximum likelihood estimators of the model parameters for HCW $i$.

Backstrom et al. [BKKK08] show that for a fixed center $c$, the function $LL(c, \beta, \gamma)$ has exactly one local maximum over its 2-dimensional parameter space $0 \leq \beta \leq 1$ and
γ ≥ 0. Thus, for a given room c, optimal values of β and γ can be found by any one of various local continuous maximization techniques; we have implemented the well-known Nelder-Mead local optimization algorithm [NM65]. To maximize LL(c, β, γ) we can simply call the Nelder-Mead algorithm for each candidate center c. However, this approach is computationally problematic because there are roughly 19,000 candidate centers (i.e., rooms in the hospital) and LL(c, β, γ) has to be maximized separately for about 15,000 HCWs. This implies more than quarter billion calls to the Nelder-Mead subroutine. Furthermore, we want to estimate spatial distributions for HCWs from various subsets of login records (e.g., separately for each month, separately for night and day, etc.) and this requires having to repeat the entire computation 15-20 times.

We employ two “tricks” to significantly reduce the running time of our computation. We first preprocess the login records and bucket them by (HCW, location). For example, we would compress two EMR logsins, one for 2 seconds and one for 3 seconds, from Room 4200 by the HCW with userID 22 to be (5, 4200, 22). The first entry, 5, refers to the total amount of login time of HCW 22 into Room 4200. After fixing a candidate center c, we do a second bucketing of the logins by their distance to c, just before the call the Nelder-Mead subroutine. These two bucketings of the logins significantly reduce (i.e., by at least an order of magnitude) the number of logarithm and exponentiation operations in practice.

Our second idea, a search pruning technique, leads to a significant reduction in the number of candidate centers we need to consider. As a result, we examine only 3% of the rooms in the hospital on average per HCW. This idea is suggested by a discussion in [BKKN08], but the way it is used there sacrifice optimality. In our work, we take advantage of the fact that the distances in the hospital graph form a metric to prune the search space of candidate centers without sacrificing optimality.

Our pruning algorithm starts with a single cluster consisting of the entire hospital graph. An arbitrary node in the graph is taken to be the cluster center and the radius of the cluster is bounded above by the diameter of the hospital graph. In a typical iteration, the algorithm considers a collection of clusters all of which have radii bounded above by r1. The goal in this iteration is to partition each cluster (if necessary) into clusters of radius at most r1/2. For any cluster C that has radius r such that r1/2 < r ≤ r1, we repeatedly pick an arbitrary node v in C and make v a cluster center of a cluster C′ and assign to the cluster C′ all nodes in C that are at a distance at most r1/2 from v. We then delete C′ from C and continue until C becomes empty. This hierarchy of clusters can be naturally viewed as a tree with the cluster consisting of the entire graph at the root and with the leaves being singleton clusters of radius 0. Note that this tree has height O(log(diam)), where diam is the diameter of the hospital graph.

Now, a simple observation leads to significant pruning, provided distances satisfy the triangle inequality. Given an upper-bound r on the distance between the center c of a cluster and any other node in the cluster, it possible to obtain an upper-bound on the maximum value of LL(v, β, γ) for any room v in the cluster. Consider the following 4-parameter function:

\[
MLL(c, β, γ, r) := \sum_{i ∈ L} \log (β \cdot (d(c, loc(ℓ)) − r)^{−γ}) + \sum_{i ∈ \mathcal{V}} \log (\frac{1}{1 - β \cdot (d(c, loc(ℓ)) + r)^{−γ}})
\]

Given a cluster C with cluster center c and radius r, we can use the Nelder-Mead algorithm to solve the continuous optimization problem to find values of β and γ that maximize MLL(c, β, γ, r). Our claim, expressed precisely in the following theorem, is that the maximum value of MLL(c, β, γ, r) (for fixed c and r) is an upper bound on the maximum value of LL(v, β, γ) for any node v in cluster C.

**Theorem 1.** Fix a node v and r > 0. For any node v, if d(c, v) ≤ r then \(MLL(c, β_v, γ_v, r) ≥ LL(v, β_v, γ_v)\) where \(β_v\) and \(γ_v\) are values that maximize \(MLL(c, β, γ, r)\) and \(β_v\) and \(γ_v\) are values that maximize \(LL(v, β, γ)\).

We defer the proof of this theorem to the full version of the paper. Theorem 1 allows us to process the hierarchy of clusters while pruning away clusters that are guaranteed not to contain a “good enough” center. In a typical step, we process a cluster C somewhere in the cluster tree. Suppose that C has cluster center c and radius r. We compute the maximum value of \(MLL(c, β, γ, r)\) over values of β and γ by making a call to the Nelder-Mead algorithm. If this value is no greater than the current best log-likelihood seen by the algorithm, we can stop searching this cluster immediately. Otherwise, we compute the maximum value of \(LL(c, β, γ)\) and continue.

Once the \(c, β,\) and \(γ\) values have been found for all HCWs, the attractiveness parameter \(α_v\) can be calculated for each node v in the hospital. To calculate the attractiveness value we first calculate the total number of seconds of login activity in each room that has a login terminal. Let M be all the rooms with a login machine in the hospital and for each u ∈ M let \(T_u\) be the total number of seconds any HCW is logged into room u. For each room v in the hospital we then calculate a smoothed activity level \(T'_v\) which reduces the effect of outliers and assigns activity values to rooms with no login terminals. For each room v in the hospital we set

\[
T'_v = s_v \sum_{u ∈ M} T_u \cdot d(u, v)^{−δ}
\]

where \(s_v\) is a scaling factor chosen so that \(s_v \sum_{u} d(u, v)^{−δ} = 1\). δ is a constant that sets the level of decay on the smoothing. For our experiments we found \(δ = 6\) gave us reasonable results.

The attractiveness value for each room v can then be found by solving for \(α_v\) in the following equation

\[
α_v \cdot \sum_i β_i \cdot d(c_i, v)^{−γ_i} = T_v
\]

where the summation is over all HCWs i.

**4. VALIDATION**

In order to verify that spatial dispersions estimated by our model match up with real-world HCW behavior, we consulted with a diverse team from the UIHC consisting of members from industrial engineering, infection control, general medicine and information technologies. This behavioral consulting team is well equipped to describe behavior
in the hospital because knowing when and where HCWs are within the facility is part of their job. We asked the consulting team to tell us about expected HCW behavior along three dimensions (1) where in the UIHC facility would we find different types of HCWs? (2) who are the most mobile and the most static departments and job categories? (3) are there particular times of the day or week during which activity would change significantly? In this section we validate our results against what happens in a real hospital setting according to these “behavior consultants.”

4.1 Centers

The Pediatrics department at UIHC consists of specialists that only deal with kids and thus the majority of HCWs working in Pediatrics spend almost all of their time in the Pediatrics unit. Figure 4 shows the centers locations for HCWs from Pediatrics on the 2nd floor for March, 2007. The cluster of dark circles in the middle of the hospital shows that, based on our model, a large number of Pediatric HCWs have their centers located there and this is where the Pediatrics unit is located in the hospital. Additionally, the behavioral consultants noted that some members of Pediatric department work in the Neonatal Intensive Care Unit and the Cystic Fibrosis Unit. Our results confirm that there are a small number (not shown here) of Pediatric HCWs that have their center in the Neonatal Intensive Care Unit or Cystic Fibrosis Unit.

4.2 Dispersions

To validate the dispersions of HCWs in our models our behavior consultants identified the most and least mobile types of HCWs. Among the most mobile job types are House Staff (Residents), Respiratory Therapists, and Information Technology Support, and Dietitians. CT Service Techs, Unit Clerks, and Secretaries were identified as being some of the least mobile job types in the hospital. To capture the notion of mobility, we introduce the concept of a t-radius. The t-radius of a healthcare worker \( i \) is the minimum \( r \) such that

\[
\text{Prob}[i \text{ is in } \text{Disk}(c_i, r)] \geq t.
\]

Here \( c_i \) is center of HCW \( i \) and \( \text{Disk}(c_i, r) \) is the set of nodes in the hospital graph within \( r \) hops of \( c_i \). For \( t = 0.8 \), Figure 5 shows the average \( t \)-radii for each job category in the hospital. Notice that CT Service Techs have a very low 0.8-radius indicating they spend 80% of their time in only a handful of rooms near their center. On the other hand, House Staff have a relatively high 0.8-radius indicating they tend to spend time in many rooms further from their center. The job categories not shown in Figure 5 and not mentioned by our behavior consultants as being very mobile or very static, lie in the middle of these extremes. A small number of job types, such as Information Technology Support, are missing in our model because they don’t login to the EMR system.

Our behavior consultants noted that even within the same job type, seniority and mobility may be (negatively) correlated. It was observed that senior members tend to move around less and assign tasks requiring going to other parts of the hospital to their junior counterparts. We looked at the difference in \( t \)-radii for House Staff. Figure 6 shows that while all house staff HCWs are fairly mobile, there is a notable difference in how rank plays a role in the mobility.

As additional validation, we use spatial dispersions to cal-

<table>
<thead>
<tr>
<th>Job category</th>
<th>Average t-radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT Service Tech</td>
<td>1.50</td>
</tr>
<tr>
<td>Secretary</td>
<td>5.06</td>
</tr>
<tr>
<td>Unit Clerk</td>
<td>7.20</td>
</tr>
<tr>
<td>Nurse Manager</td>
<td>11.0</td>
</tr>
<tr>
<td>Sonographer</td>
<td>13.6</td>
</tr>
<tr>
<td>Pharmacy Tech</td>
<td>14.0</td>
</tr>
<tr>
<td>Clinical Lab Scientist</td>
<td>16.5</td>
</tr>
<tr>
<td>Professor</td>
<td>20.1</td>
</tr>
<tr>
<td>Social Worker</td>
<td>21.2</td>
</tr>
<tr>
<td>Dietician</td>
<td>21.4</td>
</tr>
<tr>
<td>Imaging Tech</td>
<td>25.6</td>
</tr>
<tr>
<td>Respiratory Therapist</td>
<td>25.8</td>
</tr>
<tr>
<td>House Staff</td>
<td>30.3</td>
</tr>
</tbody>
</table>

Figure 5: Average \( t \)-radii with \( t = 0.8 \) for selected job categories. The higher the average \( t \)-radii, the more mobile the job category tends to be.
<table>
<thead>
<tr>
<th>Job category</th>
<th>Average t-radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>House Staff I</td>
<td>35.6</td>
</tr>
<tr>
<td>House Staff II</td>
<td>31.3</td>
</tr>
<tr>
<td>House Staff III</td>
<td>31.8</td>
</tr>
<tr>
<td>House Staff IV</td>
<td>25.0</td>
</tr>
<tr>
<td>House Staff V</td>
<td>29.6</td>
</tr>
</tbody>
</table>

Figure 6: Average t-radius for House Staff. Senior House Staff (IV and V) tend to move around less than the younger House Staff (I and II).

culate the expected distance on the hospital graph between two HCWs at any given time. Within almost all job types, the expected distance between two HCWs is about 40 with a couple notable exceptions. Staff Nurse Anesthesiologists have a significantly lower average expected distance of 20 hinting that these job types having a single focus point for activity. Within departments, the expected distance between two HCWs is lower, being in the range of 25-35 weighted hops. The fact that departments cluster together more tightly than job categories is to be expected since departments tend to have one or two units in which their employees will spend the majority of their time. A single job type will normally have HCWs spread across many departments all over the hospital. Notably, Internal Medicine is the most distributed department with a mean expected distance of 43.0, and Nursing department has an expected distance of 37.3. These two departments are large in size and include a majority of House Staff and Staff Nurses, two of the most mobile job categories.

4.3 Activity

Overnight many areas in the UIHC are shut down. For example, the in-patient and out-patient units, both on the fifth floor, are open during the day. At night the out-patient unit closes but the in-patient unit remains busy. Figure 7 shows the expected number of people in each room of the hospital’s 5th floor from 2am to 3am. For these pictures we calculated values for our centers model considering only the people who login overnight. You’ll notice that there are entire parts of the hospital where no activity is expected during these hours.

During the night there are also significantly fewer HCWs active in the hospital. The HCWs that are active tend to have to cover more territory and we would expect that the dispersion values for these healthcare workers would see a significant change. HCWs that login from 2am to 3am have a t-radii of 16.4 compared to a t-radii of 4.8 for those HCWs that login from 10am to 11am, indicating HCWs logged in at night do move around more.

4.4 Internal Consistency

We expect that centers and dispersions (γ-values) should be stable over time. Figure 8 shows the median change in center location and γ-values for each pair of consecutive months in 2007. Indeed, centers move very little from one month to the next, and γ-values are stable as well. Our behavior consultants told us that members of House Staff I are reassigned every month and this group of HCWs exhibited the largest change in center location from month to month.

5. APPLICATIONS

Knowing the spatial distributions of HCWs is useful for a variety of applications. Here we are present two: (1) inferring HCW contact graphs and (2) locating time clocks in the hospital.

5.1 Contact Graphs

Contact network epidemiology [Mey07, New02] is a relatively new area of research that investigates the spread of disease through a population based on intrinsic features of the pathogen and structural properties of a contact network (graph) that explicitly models physical interactions between pairs of individuals. The success of contact network epidemiology depends partly on the quality of the contact networks. Constructing reliable contact networks is a challenge since these networks model physical interactions as opposed to online interactions. Knowing the spatial distributions of HCWs provides a natural way for generating HCW movement within the hospital and also HCW contact networks. Here we describe our approach to generating these networks and provide a brief analysis. In the future, we aim to use these networks to evaluate control strategies such as vaccination, quarantining, cohorting, etc. for the mitigation of hospital-acquired infections.
Our basic idea is to generate, for each HCl i, a random walk in the hospital metric space whose stationary distribution is the static spatial distribution of the healthcare worker. If a HCl is in room v at time step t, then in time step t+1 she either stays in room v or moves to an adjacent room according to the transition probabilities of the random walk. Our goal therefore is to solve for these transition probabilities, given the target stationary distribution. Fix a person i and let p_v be the probability of person i being in room v. Recall that according to our model,

\[ p_v = \alpha_v \cdot \beta_i \cdot d(c_i, v)^{-\gamma_i}, \]

where c_i is i’s center, \( \gamma \) is i’s dispersion, \( \beta_i \) is the constant of proportionality, and \( \alpha_v \) is room v’s relative “attractiveness”. Note that \( \sum_v p_v = 1 \).

Assuming there are \( n \) nodes in the hospital graph labeled \( 1, 2, \ldots, n \) let \( p \) be the spatial distribution vector \( (p_1, p_2, \ldots, p_n) \) for i. Let \( T = (t_{uv})_{u,v\in\gamma} \) be HCl i’s transition matrix with \( t_{uv} \) being the probability of i moving from u to v. Note that \( t_{uv} \) and \( t_{vu} \) may be distinct and if there is no edge between u and v, \( t_{uv} = t_{vu} = 0 \). We are interested in finding a \( T \) such that \( p \cdot T = p \) and the following additional constraints are satisfied. Here \( N[u] \) denote the closed neighborhood of u, i.e., the neighborhood including u.

1. \( \sum_{v\in N[u]} t_{uv} = 1 \) for all nodes u in the hospital graph.
2. \( t_{uv} \geq 0 \) for all nodes u and v in the hospital graph.

The following claim (see for e.g., Chib and Greenberg [CG95] for a simple proof) leads to a simple algorithm for constructing \( T \).

**Lemma 1.** Suppose that for all \( u \neq v \), \( p_u \cdot t_{uv} = p_v \cdot t_{vu} \) and \( \sum_{v\in N[u]} t_{uv} = 1 \), then \( p \cdot T = p \).

This claim suggests that one way to generate entries in \( T \) is to first generate for each corresponding pair of directed edges \((u, v)\) and \((v, u)\) (\( u \neq v \)) the values \( t_{uv} \) and \( t_{vu} \) satisfying the constraint that \( p_u \cdot t_{uv} = p_v \cdot t_{vu} \). Once these values are all generated, it is possible that the sum \( \sum_{v\in N[u]} t_{uv} \) exceeds 1. If so, we scale down all the \( t_{uv} \) values so that for all u, the sum \( \sum_{v\in N(u)} t_{uv} \) is at most 1. Note that this scaling does not cause any violation of the \( p_u \cdot t_{uv} = p_v \cdot t_{vu} \) constraint. Finally, for each node u, the residual probability \( 1 - \sum_{v\in N(u)} t_{uv} \) is assigned to \( t_{uu} \). Clearly, this algorithm ensures that \( T \) is nonnegative and that \( \sum_{v\in N[u]} t_{uv} = 1 \) is satisfied for all u. Furthermore, since the \( p_u \cdot t_{uv} = p_v \cdot t_{vu} \) constraint is satisfied, from the above lemma we obtain that \( p \cdot T = p \). The algorithm is described below and we obtain the following lemma via the above argument.

**Algorithm GenTransit**

1. For each edge \((u, v)\) in the hospital graph, where \( u < v \), pick an arbitrary value for \( t_{uv} \) and assign \( t_{uv} := p_u \cdot t_{uv}/p_v \).
2. Let \( S = \max_u \sum_{v\in N[u]} t_{uv} \). If \( S > 1 \) then scale all \( t_{uv} \) by \( S \) i.e., assign \( t_{uv} := t_{uv}/S \).
3. For each u set \( t_{uu} = 1 - \sum_{v\in N[u]} t_{uv} \).

**Lemma 2.** Algorithm GenTransit constructs an \( n \times n \) matrix \( T \) such that \( t_{uv} \geq 0 \) for all nodes u and v. \( \sum_{v\in N[u]} t_{uv} = 1 \) for all u, and \( p \cdot T = p \).

### 5.2 Graph Generation

After generating a transition matrix \( T \) for each HCl i, we generate a HCW contact graph via a simple discrete simulation of the movement of the HCWs in the hospital. We start with a graph whose nodes are healthcare workers and whose edge set is empty. We add edges to this graph using the following process:

1. For each HCl i pick an initial location for i according to their static spatial distribution.
2. Add an edge of weight of one to the HCl contact graph between any two HCWs who are in the same room. If an edge already exists, increase the weight of that edge by one.
3. Move each HCl i to another room based on their current location and their transition matrix \( T \). Note that such a move may keep HCl i in the same room or may move her to an adjacent room.
4. Repeat step 2 and 3 until the contact graph has the desired density.

The fact that each HCl was introduced into the hospital according to her spatial distribution \( p \) and the fact that \( p \) is a stationary distribution for the random walk \( T \), implies that even as she is walking around the hospital her spatial distribution continues to be \( p \).

The graphs generated by this process, despite being sparse (average degree = 1% of graph size) exhibit properties such as having a high clustering coefficient, small diameter, and a heavy tailed degree distribution. See Table 1. These are properties that have been observed in other “small world” social networks [WS98, New03]. For comparison we also report corresponding features for the Erdős-Rényi random graph model \( G(n, p) \) with \( n = 6229 \) and \( p = 01 \) to give approximately the same expected degree. More specifically, we see that our graphs have a clustering coefficient of 0.5847, which is orders of magnitude larger than the clustering coefficient of the Erdős-Rényi graph of similar size and edge-density.

Our contact graphs have one giant component, consisting of more than 97% of the nodes, along with many tiny connected components. This makes it structurally very different.

<table>
<thead>
<tr>
<th>Months</th>
<th>Center movement</th>
<th>( \gamma ) movement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Feb</td>
<td>4.0</td>
<td>0.797</td>
</tr>
<tr>
<td>Feb-Mar</td>
<td>1.0</td>
<td>0.527</td>
</tr>
<tr>
<td>Mar-Apr</td>
<td>1.0</td>
<td>0.528</td>
</tr>
<tr>
<td>Apr-May</td>
<td>1.0</td>
<td>0.551</td>
</tr>
<tr>
<td>May-Jun</td>
<td>1.0</td>
<td>0.533</td>
</tr>
<tr>
<td>Jun-Jul</td>
<td>1.0</td>
<td>0.585</td>
</tr>
<tr>
<td>Jul-Aug</td>
<td>1.0</td>
<td>0.543</td>
</tr>
<tr>
<td>Aug-Sep</td>
<td>1.0</td>
<td>0.542</td>
</tr>
<tr>
<td>Sep-Oct</td>
<td>1.0</td>
<td>0.508</td>
</tr>
<tr>
<td>Oct-Nov</td>
<td>2.0</td>
<td>0.690</td>
</tr>
<tr>
<td>Nov-Dec</td>
<td>1.0</td>
<td>0.528</td>
</tr>
</tbody>
</table>

Figure 8: Median difference in centers and \( \gamma \)s in consecutive months in 2007. While a few HCWs have centers in very different parts of the hospital from one month to the next, the vast majority move less than a few weighted hops. \( \gamma \) values also change very little for most people.
from the comparable Erdős-Renyi graph, that has a single connected component. The giant component in our contact graphs have a very small diameter (14) and average path length (4.429) relative to their size.

This attempt at generating HCW contact networks extends two of our earlier, more primitive attempts [CKP+09, CHP+10, PTPS10]. In [PTPS10] we actually “shadowed” HCWs in various job categories to acquire data on their contacts. The limitation of this attempt is of course that our sample size is necessarily tiny because of how labor intensive this approach is. In [CKP+09, CHP+10] we used the EMR login data directly to obtain HCW contacts. In addition to being a way of generating contact networks, the HCW random walks we generated can be used, more generally, in discrete-event simulations of hospitals.

5.3 Time clock placement

In order to cut operating costs, the UIHC is interested in the problem of efficiently placing time clocks. There are two kinds of clocks: a basic model whose functionality is limited to clocking in, and a premium model that is capable of displaying information such as the cafeteria menu for the day and use of sick days. UIHC is interested in placing the clocks near where people start the work day so they can clock in and out immediately before and after work. The clocks with more advanced functionality are targeted toward the nursing staff.

Roughly speaking, the UIHC is interested in placing clocks so as to minimize average distance from a HCW to a time clock. If HCWs had fixed locations or if we had full knowledge of their start and finish locations, then this problem can be modeled as the well known k-mediants clustering problem. But it turns out that the k-medians problem is a useful model even without precise location information, because we have estimated HCW spatial distributions. Any algorithm for the k-median problem can work with spatial distributions rather than exact locations and minimize the average expected distance between HCWs and clocks.

We solve the problem in two separate steps. The first step is figuring out where to put the time clocks, and the second step is deciding which clocks should be the basic model, and which should be the premium model.

5.3.1 k-means and k-medians

We first implemented a simple 2-approximation algorithm for the k-means problem [HS86]. We find that a typical run yields a solution with at least one HCW needing to travel 30 units to reach a clock, and on average a room has a clock 10 units away.

We then implemented a 5-approximation algorithm for the k-median problem. This algorithm starts with a feasible solution and then tries to swap a clock from a room with one to one without one until no swaps with improvement above a threshold value are found. Arya et al. [AGK+01] show the approximation ratio of the algorithm to be bounded by $OPT (3 + 2/numSwaps)$, but with nearly 20,000 candidate locations and 80 clocks it was not computationally feasible to use anything other than $numSwaps = 1$. By solving the k-medians problem we were typically able to get the average distance from the expected location of an HCW to a clock to be 7.1 units. Figure 9 shows a fifth floor slice of a sample solution, with the usual cool tones indicating infrequent activity, and the warm tones indicating frequent activity.

### Table 1: Mean statistics for 20 generated graphs.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$ (num. vertices)</td>
<td>6,229</td>
</tr>
<tr>
<td>$m$ (num. edges)</td>
<td>193,985</td>
</tr>
<tr>
<td>$\langle k \rangle$ (mean degree)</td>
<td>62.29</td>
</tr>
<tr>
<td>$k_{\text{max}}$ (max. degree)</td>
<td>318.95</td>
</tr>
<tr>
<td>$\sigma$ (std. dev. degree dist.)</td>
<td>58.68</td>
</tr>
<tr>
<td>$\sigma_{\text{rand}}$ (std. dev. degree dist. $G(n,p)$)</td>
<td>7.84</td>
</tr>
<tr>
<td>$cc$ (clust. coeff.)</td>
<td>0.5847</td>
</tr>
<tr>
<td>$cc_{\text{rand}}$ (clust. coeff. $G(n,p)$)</td>
<td>0.0100</td>
</tr>
<tr>
<td>$c$ (num. components)</td>
<td>150.11</td>
</tr>
<tr>
<td>$c_{\text{rand}}$ (num. components $G(n,p)$)</td>
<td>1</td>
</tr>
<tr>
<td>$n_{\text{giant}}$ (num. vertices giant comp.)</td>
<td>6,052.37 (97.16%)</td>
</tr>
<tr>
<td>$m_{\text{giant}}$ (num. edges giant comp.)</td>
<td>193,945.63 (99.98%)</td>
</tr>
<tr>
<td>$diam$ (diam. giant comp.)</td>
<td>14.47</td>
</tr>
<tr>
<td>$\langle l \rangle$ (ave. path len. giant comp.)</td>
<td>4.464</td>
</tr>
</tbody>
</table>

5.3.2 UIHC: Was k appropriate?

For any clustering problem we need to be considerate of $k$. UIHC was also interested in determining if 80 was a sufficient number of clocks to purchase. Assuming that queuing is
not an issue, Figure 10 suggests that 80 clocks provides adequate coverage and adding more clocks would provide little benefit.

![Distance to nearest time clock](image)

**Figure 10:** Shows the maximum and average distance in hops a HCW needs to travel to reach a clock versus the number of clocks $k$.

### 5.3.3 Allocating the premium clocks

The process of allocating the premium clocks is straightforward. For each clock, count the expected number of nurses assigned to that clock. The clocks with the highest number of expected nurses are allocated the premium model, and the rest of the clocks are the basic model.

### 6. FUTURE WORK

The EMR login data, despite being rather noisy, seems to have enough “signal” in it to be able to provide robust estimates of spatial distributions of HCWs in a hospital environment. The estimated spatial distributions match our expectations quite nicely. As far as we know, this is the first attempt at modeling the spatial distributions of HCWs and there are a number of modeling aspects and applications that can be further pursued.

We intend to extend our work in several directions in the near future. Below, we briefly outline some of these.

1. HCWs in certain job categories seem to have two or more natural centers of activity, e.g., physicians might exhibit some activity in or near their office and some activity near the unit in which they see patients. We are exploring the use of several alternate models that permit individual users to have multiple centers of activity. Preliminary work indicates that these models with multiple centers are better at capturing HCW activity. Presently we are doing a preliminary evaluation of these models and also looking for ways to make the maximum likelihood estimation of model parameters more efficient.

2. We are working on more procedures for validating our model. Additional internal validation will be performed for single center models by using 10-fold cross validation. 10% of each user’s logins will be withheld during model fitting, and then we will compare the spatial distributions obtained from the model with the corresponding distributions of the withheld logins. We also intend to examine hospital epidemiology literature for other sources of comparison for HCW movement within a hospital. A recent paper on how the architecture of a unit affects the movement of nurses [HCBH09] and patient care in general, seems to be a good starting point. We have access to a data set obtained from shadowing HCWs at the UIHC and recording when interactions took place and how long they lasted [PTPS10]. We intend to compare contact graphs generated from the models described in this paper with contact graphs generated from the shadow data. Lastly, when time clocks are eventually placed in UIHC, we can replay the login data and see how our suggested placement holds up against the real-world placement in simulations.

3. Other probability decay functions that are not polynomial in the distance (e.g., exponential) might better capture the behavior of HCWs of certain type. We have begun exploring how these multiple decay functions might be incorporated into our model.

It should also be noted that this research has implications outside of healthcare. Indeed, the work was inspired by a framework developed to find the centers of search engine queries. It could answer questions as seemingly unrelated as “Where should the Iowa City Police Department increase its presence to combat the rise in downtown violence?”.

### Acknowledgments

We would like to thank Stephanie Holley RN BSN CIC, Tom Persoon, Hendrick Shultz M.D., and Chris Smith for their input as “behavioral consultants.” We’d also like to acknowledge Jason Fries, Matt Stults, Matt Hootman, Gaurav Kanade, Shobha Kazinka, Justin Hoffman, James Paton, and Brandyn Kusenba for their help manually building the hospital graph from UIHC CAD printouts.

### 7. REFERENCES


[HS86] D Hochbaum and D Shmoys, *A unified


APPENDIX

A. PREFERRED REVIEW APPROACH

A.1 Focus

1. Computing
2. Information Science

A.2 Topics

1. Innovative applications in electronic health records
2. Technologies for capturing and documenting clinical encounter information in electronic systems