2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust

Inferring realistic intra-hospital contact networks using link prediction and computer logins

Ted Herman*, Mauricio Monsalve*, Sriram Pemmaraju*, Phillip Polgreen†,
Alberto Maria Segre*, Deepti Sharma‡ and Geb Thomas‡
* Department of Computer Science, University of Iowa, CompEpi group
† *Department of Internal Medicine, University of Iowa, CompEpi group*
‡ *Department of Industrial Engineering, University of Iowa, CompEpi group*

***Abstract*—Disease spread in hospital settings is a common and important problem in health care. Knowing the network of contacts between health care workers and patients can be very helpful in mitigating disease spread. In this work, we address the problem of inferring the contact network of health care workers at the University of Iowa Hospital and Clinics facilities by integrating two sources of data: hospital-wide computer login data and proximity data obtained from direct measurement in the Medical Intensive Care Unit using a wireless sensor network. We treat this problem as a variant of the *network completion* problem, where one small portion of the network is well known while the rest is sparingly sampled, and we want to complete the network. In this case, we want to transform the login network, where an edge connects two people who logged into computers within some time and distance, of the hospital into a contact network. We solve this problem by borrowing techniques from *link prediction*. We train and evaluate these techniques on synthetic login networks and contact networks obtained from the sensor data. Our results are promising in that we can predict contact networks from login networks with accuracies mostly between 70% and 90%.**

***Keywords*-contact networks, link prediction, network epidemiology**

I. Introduction

In-hospital acquired infections represent an important threat to health care; each year, approximately two million people are affected by hospital-acquired infections, resulting in thousands of deaths [1]. To study how diseases spread in hospital settings and to design adequate containment strategies, researchers have turned to social network analysis, in particular to the study of *contact networks* [2], [3]. A contact network is a network in which vertices represent people and edges represent contacts, which can be defined in terms of physical proximity, physical contact, room sharing, etc, depending on the transmission mechanisms of the disease being studied. However, the main difficulty of such research is the cost of acquiring extensive, detailed and near-complete hospital-wide contact network data. Privacy concerns and an understandable reluctance on the part of health care workers to submit to monitoring are also obstacles to this type of research. Therefore, researchers have turned to a number of strategies to infer contacts networks. Efforts have included the inference of contacts networks from: the relation between job structure and architecture [2], health records (which relate workers and patients) [4], ego networks obtained by shadowing health care workers [5], hospital-wide computer login data (to estimate the proximity between workers) [6], [7], and the combination of a worker mobility model with computer login data and architecture [8]. However, experience in simulation has shown that using slightly different inputs, such as different assumptions with regard to contacts, may lead to different results [10]. Smieszek makes a case for the need for details such as duration and intensity in contacts in epidemic models [11]. Because we want to obtain realistic results, we seek to make use of more complete and fine-grained data to reduce the need for modeling assumptions.

In this work, we describe the design and fine tuning of a method to infer hospital-wide contact networks that works by combining information from two data sets. We have a fine-grained data set which contains electronically measured distances between health care workers in the Medical Intensive Care Unit (MICU) of the University of Iowa Hospitals and Clinics (UIHC) during a period of ten days, all day long, with time resolution of seconds (an earlier version of the experiment is described in [9]). We propose using this data to *extend* our second data set, the hospital-wide UIHC computer login network introduced in [6]–[8] to generate realistic hospital-wide contact networks of proximity. In this second data set, workers are related if they logged into close-by computers within a fixed time interval, as an estimation of likely encounters between workers who are nearby. Since the login network data set is necessarily incomplete and could be missing critical contacts, the problem of enhancing the login network can be interpreted as a network completion problem, where we are to discover the missing edges using the most complete part of the network as reference. We utilize techniques from *link prediction* to perform such a transformation [12].

Our findings are promising: we are able to predict contact networks with accuracies mostly between 70% and 90%. The implication is that we can use link prediction that is trained by contacts obtained from the small-scale sensor network data to enhance the hospital-wide computer login

978-0-7695-4848-7/12 $26.00 © 2012 IEEE
DOI 10.1109/SocialCom-PASSAT.2012.113

network in order to make it epidemiologically useful. We also found that there are pairs of login and contact networks that closely resemble each other, implying that, in some cases, login networks are epidemiologically useful without further enhancements.

II. Link prediction

Link prediction consists of estimating the likelihood an edge exists between two vertices in a social network [12]. It is an important problem in social network research because it enables the completion of partial network data sets, which may be very expensive to capture, and permits the study of network dynamics [13].

A. Similarity scores

Several techniques are used in link prediction. Due to their simplicity, popular are those that measure the similarity of the neighborhoods of two vertices to determine whether they should be connected by an edge. Most of those popular *similarity scores* were introduced by Liben-Nowell and Kleinberg [14], while some were added by Fire et al. [17], Zhou et a.l [22], and others. Those scores, specially the simpler ones, have proven to have more predictive power than much more elaborated predictive methods [18]. In fact, the very simple scores of *common neighbors*, *Jaccard's index* and *Adamic-Adar's score* seem to be the most predictive [22]. Given graph $G=(V,E)$, let $\Gamma(v)$ be the neighborhood of vertex $v \in V$. The *common neighbors* score counts the common neighbors between two vertices $u,v \in V$:

$$CN(u,v) = |\Gamma(u) \cap \Gamma(v)|. \tag{1}$$

Jaccard's index, often used to measure the similarity between two sentences, can be redefined to apply to neighborhoods:

$$JI(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}. \tag{2}$$

The *Adamic-Adar's* score, originally defined to measure the similarity of web pages normalizing for word frequency, is:

$$AA(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(z)|}. \tag{3}$$

Many more scores have been proposed, some being similar to those previously mentioned. Some are based on random network models such as *preferential attachment* and the *configuration model*, and others are based on specific domains, such as ecology and protein networks [22]. Slightly different scores based on paths have also been proposed [14], [21], [22].

The previous measures are used as features for classification methods. An edge is reduced to a score, and then that score is classified to *yes* or *no*, predicting whether the edge exists.

For more methods of link prediction, refer to the survey of Lü and Zhou [12].

B. Weighted networks

Link prediction in weighted networks has meant a special opportunity for prediction: to use weights as predictors. Murata and Moriyasu introduced weighted modifications of the popular link prediction scores [19], which later were slightly modified in other works [16], [20], [21]. Consider an edge weighted graph $G=(V,E)$. Let $w(u,v)$ be the weight of edge $\{u,v\} \in E$ and $s(u) = \sum_{\{u,v\} \in E} w(u,v)$ be the *strength* of vertex $u \in V$. For example, *weighted common neighbors* [20] is defined as:

$$WCN(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} w(u,z) + w(v,z). \tag{4}$$

Observe that, if $\forall \{u,v\} \in E$, $w(u,v) = 1$, then eq. 4 is exactly twice as big as eq. 1. *Weighted Jaccard's index* and *weighted Adamic-Adar* are derived from their unweighted counterparts (Eqs. 2, 3) using the same rationale.

Using weights might not improve prediction. Unweighted predictors may outperform weighted predictors [20], which can be considered a *weak ties* phenomenon: even the weakest ties count. This also seems to happen with the Adamic-Adar score in unweighted graphs, where discarding repetitions increases the predictive power of the score [18]. To account for this, Lu and Zhou introduced the *weak ties* parameter α [20], which is used as follows:

$$CN_\alpha(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} w(u,z)^\alpha + w(v,z)^\alpha. \tag{5}$$

If $\alpha = 0$, then CN_α is equivalent to its unweighted version, Eq. 1 (just twice as big). If $\alpha = 1$, then CN_α becomes Eq. 4. Lu and Zhou found in their networks that using $\alpha < 0$ increased the predictive power of the scores.

C. Contact networks

Some researchers have used link prediction scores to predict contacts in human contact networks [23]–[25]. Wang et al. studied human mobility in a mobile phone data set by using several of the popular scores as well as some defined specifically for their problem [23]. Jahanbakhsh et al. performed link prediction on contact networks by modifying the standard link prediction scores to integrate multiple edge types, an evaluated them on a variety of sensor-measured networks [24], [25].

III. Materials and methods

A. The MICU health care worker mobility data

The MICU data set contains proximity and mobility data of health care workers in the MICU at the UIHC during 10 days (June 1 to June 10, 2011), for morning (7 am to 7 pm) and night (7 pm to 7 am) shifts. The data was obtained using a wireless sensor network that was partly mobile and partly stationary.

The experimental setup was as follows. Fixed sensors (*beacons*) were placed around the unit, in corridors, on

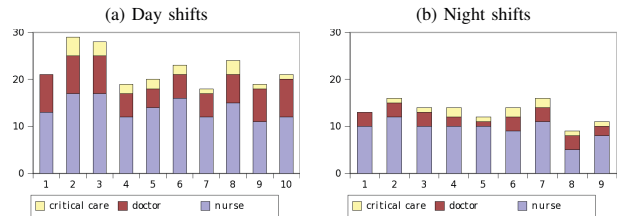

Figure 1. Number and distribution of job types per shift, during (a) 10 day shifts and (b) 9 night shifts.

beds and on doors. Then, a researcher carried some sensors and walked around the unit to obtain *ground truth* signal strengths. Finally, wearable sensors (*badges*) were given to health care workers to carry during their shifts. To protect the identity of health care workers, badges were given to them in a random order at the start of each shift. However, the same set of badges were used for a particular job type in every shift. There was a set of badges for nurses, another one for doctors, and the last one for critical care staff. The number of badges handed during each shift is depicted in Fig. 1. On average, 20.1 badges were handed during day shifts and 13.2 during night shifts.

All sensors emitted signals periodically (every 6 to 10 seconds). Badges received these signals and recorded the identifier of the sender, the signal strength (RSSI), and the moment the signal was received (time-stamp, in seconds). Signal strength between badges account for proximity while signal strength from beacons account for location.

B. The EMR login network data

The EMR login network data set induces a set of network files (graphs) built from login data of health care workers in the EMR system of the UIHC. The original login data set contained 19.8 million of logins of 15,945 users of 80 departments and 404 different job titles, and corresponds to a period of 22 months (89 weeks). Each login entry consisted of the time a worker logged into a computer, his/her hash id, his/her job type, and the identity of the computer used. The identity of the health care worker was removed due to privacy concerns, and replaced by an anonymous hash id.

From this data set, login networks were constructed by setting three parameters: a 4-week period T, a time t, a distance d, and a minimum weight $w_L^{\min}$. For a 4-week period indexed by T, two health care workers were in contact if they logged into computers at most d rooms away within t minutes. Room distance d, which accounts for rooms, hallways and stairs, ranges from 1 to 5. Time t can be 0, 5, 10, 15 or 30 minutes. Minimum weight $w_L^{\min}$ set the minimum number of *login-relations* that an edge must have. In the resulting networks, vertices have the following attributes: worker id, job title, number of logins, and total time logged into the system. Edges have the following attributes: number of contacts and total contact duration.

Ideally, we would have used the MICU contact networks to complete the EMR login contact networks. However, as we explain in the following section this is difficult to do because the networks span different time periods and consist of possibly distinct health care workers. To get around this problem we propose a method that involves generating synthetic login contact networks from the MICU mote data. Our method ensures an easy identification of the vertices in the login contact networks and vertices in the MICU mote contact networks. We provide more details in the following section.

C. Experiment design

We want to transform the EMR login networks into contact networks (of proximity). Specifically, let $G_L = (V, E_L)$ denote a login network, and a mapping $w_L : E_L \to \mathbb{N}$ that associates an edge to an integer number which represents the number of contacts (see Sec. III-B). We would like to estimate the hospital-wide contact network $G_C = (V, E_C)$ by using G_L and w_L. But we do not have a hospital-wide contact network to design or evaluate a method to perform such transformation. Instead, we have the MICU mote data, from which we can generate contact networks for a small number of workers who work in the unit. This suggests the following approach. Suppose we identify the subgraph of G_L corresponding to the MICU. Call this subgraph G_L^{MICU}. Suppose we *learn* a function $f : G_L^{MICU} \to G_C^{MICU}$ that transforms the login network restricted to the MICU into a contact network G_C^{MICU}. We could attempt to extend f to the entire hospital-wide login network G_L. The bottleneck for this approach is that the EMR login data and the MICU mote data come from different time periods and there is no way of identifying which health care workers appear in both. To get around this problem, we can use the MICU data to generate login-like networks. Contacts between workers are measured by the signal strength between their badges. Logins can be simulated by detecting when a person enters a room and stays inside for some time. Thus, we can generate pairs of contact and login networks with which we can prepare and evaluate methods for inferring contacts from logins. (And, in the future, apply them to infer hospital-wide contact networks.) In what follows, G_L and G_C denote login and contact networks for the MICU only.

Because of the way the experiment was conducted, the MICU data does not permit us to combine data from different shifts. Different health care workers were on duty on different shifts, and when they repeated, they did not necessarily wear the same badges. So, to generate graphs with consistent sets of vertices, and thus edges, we have to restrict the generation of each pair of graphs to span one shift only.

Since we generate pairs of graphs within a shift, they share the exactly same vertex sets, i.e. we generate pairs of graphs $G_L = (V, E_L)$ and $G_C = (V, E_C)$. Therefore, we basically need

to create functions f such that $f(E_L) \approx E_C$, which we can easily evaluate through *accuracy*, i.e. the number of times condition $\{u,v\} \in f(E_L) \Leftrightarrow \{u,v\} \in E_C$ holds.

In addition, we fit prediction functions f on day and night shifts separately. We expect contact patterns to be different between day and night.

Taking into account all of the above, testing the predictive power of f becomes straightforward through *cross-validation*. K-fold cross-validation is a test that evaluates the performance of a predictive function through out-of-sample testing. The data set is split into K bins or parts, and the predictive function is evaluated on each bin after it is trained on the rest. In practice, K is normally chosen to be 10. In our case, since we have 10 day and 9 night shifts, we evaluate the predictive power of f using 10-fold cross-validation for day shifts and 9-fold cross-validation for night shifts.

We do not generate just one pair of login and contact networks per shift. Instead, we generate several login and contact networks per shift by using different definitions of *login-relations* and contacts as we explain in the next section. Then, we generate predictive functions f for all the different combinations of login and contact networks. By doing so, we can determine which predictive function and login network definition best predicts each contact network.

D. Network generation

We generated a login network from shift j in the MICU data as follows. We first define the set of vertices V as the set of *id*s of the badges used in shift j. To define the set of edges E_L, we first need to detect when a badge entered a room; if the badge received a signal from a beacon with a signal strength above a threshold (obtained experimentally), then we say that the badge is inside the beacon's room. If the worker stays in the room for 4 or more minutes, we consider she/he logged into the room's computer. Then, if another health care worker logged into another computer at most d rooms away in less than t minutes, we consider it a *login-relation*. Define $w_L(u,v)$ as the number of contacts between u and v, for all $u,v \in V$. Then, we define the set of edges $E_L = \{\{u,v\} : u,v \in V \wedge w_L(u,v) > 0\}$. The resulting login network is then $G_L = (V,E_L)$ and its weight function is w_L. In this work, we limit ourselves to generating login networks with parameters $d \in \{0,2,4\}$ rooms of distance and $t \in \{10,15,30\}$ minutes. Note that we do not threshold the minimum number of login-relations $w_L^{\min}$; instead, we consider the weights explicitly in the experiment.

For a shift j, its contact network G_C is defined as follows. We define the set of vertices V as the set of *id*s of the badges used in shift j. Badges $u,v \in V$ are in contact if their mutual signal strength readings exceed a threshold chosen to correspond to a distance of 5 feet approximately. Define $w_C(u,v)$ as the fraction of time badges u and v were in contact. Then, we define the set of edges $E_C = \{\{u,v\} : u,v \in V \wedge w_C(u,v) > 0\}$. The resulting contact network is

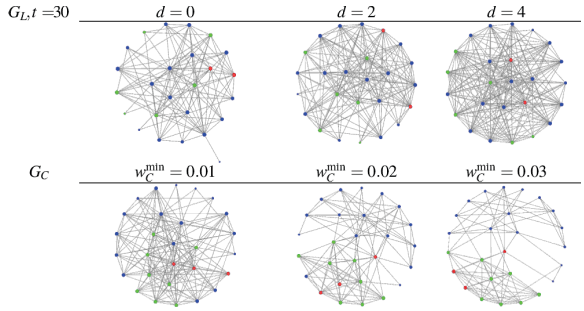

Figure 3. Some login and contact networks generated using the data of June 3, day shift.

$G_C = (V,E_P)$, and its weight function is w_C. Also, we remove relatively *insignificant* weights if $w_C < w_C^{\min}$. In this work, we limit ourselves to generating contact networks within 5 feet, and removing edges with $w_C^{\min} \in \{0.01, 0.02, 0.03\}$ (less than 1%, 2% and 3% of close encounters within the shift). We chose such small minimum weights because the distribution of weights is heavy tailed (see Fig. 2).

We generated 9 login and 3 contact networks per shift, resulting in $12 \times 19 = 228$ graphs. Figure 3 illustrates the graphs generated for the day shift of June 3, 2011. Table I shows some properties of the generated graphs, namely the averages of their: number of edges, average degree, density, clustering coefficient (the probability that two neighbors of a vertex are connected by an edge), and transitivity (three times the ratio between the number of triangles and the number of triples in a graph).

E. Training and prediction

The eight selected similarity scores are shown in Table II. The first six scores are adaptations of the traditional link prediction scores *common neighbors*, *Adamic/Adar* and *Jaccard's index*. The functions include the weak ties parameter α [20]. The last two scores, however, are meant for a direct translation of the login networks into contact networks. The *login hypothesis* score (LH) ignores the weights defined in the login networks, while the *login hypothesis improved* score ($LH^\star$) takes weights into account.

Consider login network $G_L = (V,E_L)$ and some similarity score f. To predict contact network $G_C = (V,E_C)$, function f is evaluated on all pairs of vertices $u,v \in V$. In case of *homophily*, if $f(u,v) \geq \theta$, then $\{u,v\} \in E_C$ is predicted, for some threshold θ. In case of *heterophily*, $\{u,v\} \in E_C$ is predicted if $f(u,v) \leq \theta$. For each similarity score, its threshold θ, usage style (homophily or heterophily), and weak ties parameter α are determined by maximizing the accuracy of the estimations in the training set. Also, training and prediction is performed separately for each type of edge (*c.care-c.care*, *c.care-doctor*, *c.care-nurse*, *doctor-doctor*, *doctor-nurse*, *nurse-nurse*). And as mentioned, the predictive accuracy of the scores is evaluated through N-

(a) w_C distribution in day shifts (b) w_L distribution in day shifts (c) w_C distribution in night shifts (d) w_L distribution in night shifts

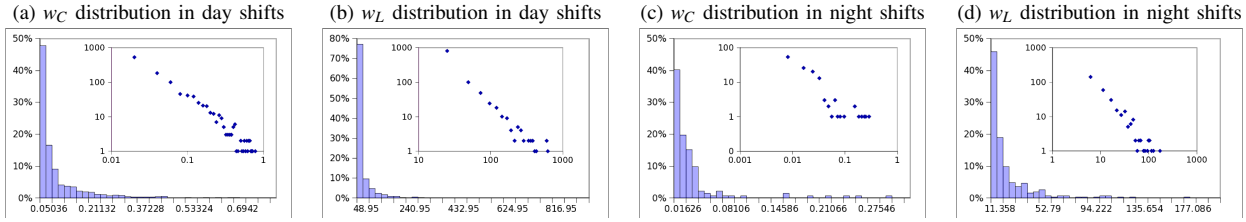

Figure 2. Histograms of the weight distribution in login and contact networks, for (a) contact networks G_C, $w_C^{\min} = 0.01$, of day shifts, (b) login networks G_L, $t = 15$ and $d = 2$, of day shifts, (c) contact networks G_C, $w_C^{\min} = 0.01$, of night shifts, and (d) login networks G_L, $t = 15$ and $d = 2$, of night shifts.. The insets are log-log scaled, and illustrate that weights are long tailed distributed.

(a) Day shifts

G_L		Edges	Avg. degree	Density	Clust.	Trans.
t	d					
10	$d=0$	52.6	4.75	0.23	0.38	0.43
	$d=2$	96.2	8.60	0.41	0.64	0.66
	$d=4$	128.7	11.48	0.55	0.73	0.77
15	$d=0$	66.1	5.95	0.29	0.48	0.51
	$d=2$	104.4	9.34	0.45	0.67	0.69
	$d=4$	134.7	12.00	0.58	0.75	0.79
30	$d=0$	81.5	7.33	0.35	0.58	0.58
	$d=2$	115.0	10.25	0.49	0.70	0.73
	$d=4$	144.1	12.82	0.61	0.78	0.83
G_C	$w_C^{\min}$					
	0.01	109.7	9.78	0.47	0.64	0.61
	0.02	74.2	6.66	0.32	0.54	0.54
	0.03	57.2	5.11	0.25	0.51	0.55

(b) Night shifts

G_L		Edges	Avg. degree	Density	Clust.	Trans.
t	d					
10	$d=0$	18.00	2.65	0.22	0.29	0.36
	$d=2$	31.78	4.72	0.39	0.59	0.59
	$d=4$	46.00	6.83	0.57	0.80	0.76
15	$d=0$	21.33	3.15	0.26	0.40	0.43
	$d=2$	33.89	5.02	0.41	0.62	0.59
	$d=4$	48.00	7.12	0.59	0.80	0.77
30	$d=0$	27.44	4.07	0.34	0.56	0.54
	$d=2$	39.44	5.84	0.48	0.74	0.66
	$d=4$	52.56	7.80	0.65	0.84	0.81
G_C	$w_C^{\min}$					
	0.01	14.67	2.17	0.18	0.37	0.37
	0.02	8.22	1.21	0.10	0.14	0.24
	0.03	4.56	0.68	0.06	0.08	0.21

Table I
AVERAGE PROPERTIES OF THE LOGIN AND CONTACT NETWORKS, OVER (A) 10 DAY SHIFTS AND (B) 9 NIGHT SHIFTS.

Name	Formula
Common neighbors	$CN_\alpha(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} w_L(u,z)^\alpha + w_L(z,v)^\alpha$
Common neighbors extended	Same as CN, but using Γ^+ instead of Γ
Adamic-Adar	$AA_\alpha(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{w_L(u,z)^\alpha + w_L(v,z)^\alpha}{\log(1+s_\alpha(z))}$
Adamic-Adar extended	Same as AA, but using Γ^+ instead of Γ
Jaccard's index	$JI_\alpha(u,v) = \frac{CN_\alpha(u,v)}{s_\alpha(u)+s_\alpha(v)}$
Jaccard's index extended	Same as JI, but using Γ^+ instead of Γ
Login hypothesis	$LH(u,v) = \begin{cases} 1, & \{u,v\}\in E_L \\ 0, & \sim \end{cases}$
Login hypothesis improved	$LH^\star(u,v) = w_L(u,v)$

Table II
SIMILARITY SCORES USED IN THIS WORK. THE FORMULAS ARE DEFINED FOR GRAPH $G_L = (V, E_L)$, AND VERTICES $u, v \in V$. NEIGHBORHOOD FUNCTION Γ IS DEFINED AS $\Gamma(v) = \{u \in V : \{u,v\} \in E_L\}$, AND $\Gamma^+(v) = \Gamma(v) \cup \{v\}$. WEIGHT w_L IS SUCH THAT $w_L(u,v) = 0$ WHEN $\{u,v\} \notin E$.

fold cross validation (10 folds for day shifts and 9 folds for night shifts).

IV. RESULTS

A. Performance of link prediction

Insets (a), (b) and (c) of Table III summarize the performance of the predictors during day shifts. For contact networks G_C generated with $w_C^{\min} = 0.01$, the *login hypothesis* predictor *LH* achieved the highest accuracy (0.68). This suggests that there is a large overlap between login-relations (G_L with $t = 15$, $d = 2$) and contacts (G_C with $w_C \geq 0.01$), specially if we consider that their densities are similar (0.45 for G_L and 0.47 for G_C). For contact networks with $w_C^{\min} = 0.02$ and $w_C^{\min} = 0.03$, the common neighbor scores *CN* and CN^+ performed best.

Insets (d), (e) and (f) of Table III summarize the performance of the predictors during night shifts. The best predictor, for all contact networks, is the unweighted Adamic-Adar score *AA*. However, we must state that these results are not auspicious. Adamic-Adar performs as well as predicting $E_C = \emptyset$, i.e. predicting no proximity contacts; the accuracies achieved can be concluded just by looking at the densities presented in Table I. In spite that the trained *AA* does not predict $E_C = \emptyset$ (but it does for several edge types), it seems that the densities of the login networks are too different from the densities of the contact networks for the scores to produce meaningful predictions.

When training the predictors, most relations were of homophily but there were several of heterophily. Heterophily was mainly present in Jaccard's index scores *JI* and JI^+, normally in *criticalcare-doctor* relations, together with negative α. Otherwise, most relations were of homophily. Also note that most link prediction scores had very different configurations for different types of edges, suggesting that interaction varies substantially among job types.

(a) Accuracy forecasting day G_C with $w_C \geq 0.01$

Score	Accuracy	G_L	t	d
AA	0.653		30	2
AA^+	0.653		30	2
CN	0.635		30	2
CN^+	0.642		10	2
JI	0.611		30	4
JI^+	0.611		30	4
LH	**0.680**		15	2
$LH^\star$	0.552		30	4

(b) Accuracy forecasting day G_C with $w_C \geq 0.02$

Score	Accuracy	G_L	t	d
AA	0.723		30	2
AA^+	0.733		30	2
CN	**0.734**		10	2
CN^+	**0.734**		10	2
JI	0.731		30	4
JI^+	0.731		30	4
LH	0.700		30	0
$LH^\star$	0.624		30	2

(c) Accuracy forecasting day G_C with $w_C \geq 0.03$

Score	Accuracy	G_L	t	d
AA	0.764		10	2
AA^+	0.792		10	2
CN	**0.795**		10	2
CN^+	**0.795**		10	2
JI	0.740		15	4
JI^+	0.783		30	4
LH	0.737		10	0
$LH^\star$	0.779		30	0

(d) Accuracy forecasting night G_C with $w_C \geq 0.01$

Score	Accuracy	G_L	t	d
AA	**0.832**		30	4
AA^+	0.831		15	4
CN	0.827		30	4
CN^+	0.829		30	4
JI	0.816		30	4
JI^+	0.815		30	4
LH	0.696		15	0
$LH^\star$	0.749		30	2

(e) Accuracy forecasting night G_C with $w_C \geq 0.02$

Score	Accuracy	G_L	t	d
AA	**0.893**		30	4
AA^+	0.888		30	4
CN	0.887		15	4
CN^+	0.888		30	4
JI	0.877		30	4
JI^+	0.877		30	4
LH	0.715		10	0
$LH^\star$	0.812		30	0

(f) Accuracy forecasting night G_C with $w_C \geq 0.03$

Score	Accuracy	G_L	t	d
AA	0.873		30	4
AA^+	0.869		30	4
CN	0.866		30	4
CN^+	0.871		30	4
JI	0.871		30	4
JI^+	**0.874**		30	4
LH	0.664		30	0
$LH^\star$	0.759		30	2

Table III
ACCURACY OF LINK PREDICTION SCORES IN DAY SHIFTS (INSETS (A), (B), (C)) AND NIGHT SHIFTS (INSETS (D), (E), (F)), OBTAINED THROUGH CROSS VALIDATION.

B. Login hypothesis

The login hypothesis predictors were chosen following our belief that login-relations do a good job of estimating contacts. The finding that *LH* is the best predictor for day shift G_C, $w_C^{\min} = 0.01$, seems to confirm that login-relations predict an important number of contacts. *LH* is not a flexible score. For an edge type, it can assume four configurations. Let $t_1(V)$ and $t_2(V)$ represent vertices of job types t_1 and t_2. Then, for every $u \in t_1(V)$ and $v \in t_2(V)$, *LH* either predicts: $\{u,v\} \in E_C$, $\{u,v\} \notin E_C$, $\{u,v\} \in E_L \Leftrightarrow \{u,v\} \in E_C$, and $\{u,v\} \in E_L \Leftrightarrow \{u,v\} \notin E_C$. This latter possibility never happened in practice. Thus, when predicting G_C, $w_C^{\min} = 0.01$, with G_L, $t = 15$ and $d = 2$, all the predictions were performed in either a direct translation or a default prediction. In addition, observe that for higher $w_C^{\min}$, $LH^\star$ became more predictive than *LH*. $LH^\star$ has the ability to apply thresholds to weights for performing predictions. Again, heterophily was never used together with $LH^\star$, so it predicted that more frequent login-relations predict longer contacts. This seems to be confirmed when predicting G_C, $w_C^{\min} = 0.03$, as $LH^\star$ achieved an accuracy of 0.779 (versus the maximum of 0.795 achieved by *CN*).

The link hypothesis predictors did not work during the night, probably because the login networks were much denser than the contact networks, and the proportion of job types varied significantly across shifts (see Fig. 1), making prediction difficult. (The other predictors were affected by this issue as well.)

V. CONCLUSIONS

Realistic contact networks are of great importance to the study and containment of hospital-acquired infections. In this work, we prepared a method to derive hospital-wide contact networks by using computer login data and direct measurements of workers' movement dynamics. Our results are preliminary, and can be improved by incorporating more information in link prediction (for example, distance).

An interesting finding of our work was that login-relations predict some contacts without making use of further transformations to perform the prediction, save for predicting always *yes* or *no* for edges connecting some job types. This suggests that simple cues from the work schedule partially predict encounter between workers. And besides, we found that predictions can improve by just using simple transformations, such as considering the number of common neighbors between two vertices.

Besides the benefits for generating contact networks, link prediction can also increase one's knowledge about the dynamics of the contact network under study. For instance, we found that the weak ties phenomenon holds in the MICU contact networks, and that the contact network dynamics are driven by both homophily and heterophily. Perhaps, by using more refined link prediction techniques, it could be possible to acquire a deeper understanding of the dynamics of in-hospital contact networks.

The main limitation of the work consisted in the generalizability of the patterns of the MICU to the whole hospital. Each day, roughly 4,000 health care workers work in the UIHC while only 30 to 40 do so in the MICU. The work routine in the different units is also different. For example, patients are admitted directly from outside of the hospital frequently to other units, whereas that does not apply to the MICU. Also, not all units need night shifts. Another important limitation is in regard of job types: the EMR data contains several additional job types that were not present in the MICU data. All of these limitations can be overcame

with the use of more contact data.

REFERENCES

[1] Klevens RM, Edwards JR, Richards CL Jr, Horan TC, Gaynes RP, Pollock DA, Cardo DM. Estimating Health Care-Associated Infections and Deaths in US Hospitals, 2002. Public Health Reports 122 (2007), pp. 160-166.

[2] L.A. Meyers, M.E.J. Newman, M. Martin, S. Schrag. Applying network theory to epidemics: control measures for Mycoplasma pneumoniae outbreaks. Emerging Infectious Diseases, 9 (2003), pp. 204-210

[3] LA. Meyers, Contact network epidemiology: Bond percolation applied to infectious disease prediction and control, Bulletin: American Mathematical Society 44 (2007), 63-86.

[4] T. Ueno, N. Masuda. Controlling nosocomial infection based on structure of hospital social networks. Journal of Theoretical Biology 254(3), pp. 655-666, 2008.

[5] P. Polgreen, T. Tassier, S. Pemmaraju, A.M. Segre. Prioritizing Healthcare Worker Vaccinations on the Basis of Social Network Analysis. Infection Control and Hospital Epidemiology 31(9): 893–900. 2010.

[6] D.E. Curtis, G. Kanade, S. Pemmaraju, P. Polgreen, A.M. Segre. Analysis of hospital health-care worker contact networks. 5th UK Social Networks Conference, 2009.

[7] D.E. Curtis, S. Pemmaraju, C. Hlady, J. Fries, T. Herman, A.M. Segre, and P. Polgreen. Vaccination strategies for healthcare workers based on social networks. In Proceedings of the International Meeting on Emerging Diseases and Surveillance (IMED), pp. 101-102, 2009.

[8] D.E. Curtis, C.S. Hlady, S. Pemmaraju, P. Polgreen, A.M. Segre: Modeling and estimating the spatial distribution of healthcare workers. IHI 2010: 287-296

[9] T. Hornbeck, D.E. Curtis, T. Herman, G. Thomas, A.M. Segre, P. Polgreen. Contact Patterns for HCWs: Not Everyone is the "Average". 21st Annual Scientific Meeting of the Society for Healthcare Epidemiology of America, 2011.

[10] M.E. Halloran, N.M. Ferguson, S. Eubank, I.M. Longini, D. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T.C. Germann, D. Wagener, R. Beckman, K. Kadau, C. Barrett, C.A. Macken, D.S. Burke, P. Cooley. Modeling targeted layered containment of an influenza pandemic in the United States. PNAS 105(12): pp. 4639-4644, 2008.

[11] T. Smieszek. A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread. Theoretical Biology and Medical Modeling. 2009; 6: 25.

[12] L. Lü, T. Zhou. Link prediction in complex networks: A survey. Physica A 390 (2011): 1150-1170.

[13] E. Acar, D.M. Dunlavy, T.G. Kolda. Link Prediction on Evolving Data using Matrix and Tensor Factorizations. CDMW'09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, pp. 262-269, 2009.

[14] D. Liben-Nowell, D. Kleinberg. The link prediction problem for social networks. CIKM 2003.

[15] M. Kim, J. Leskovec. The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. SIAM ICDM 2011.

[16] R.N. Lichtenwalter, J.T. Lussier, N.V. Chawla. New Perspectives and Methods in Link Prediction. KDD 2010, July 25–28, 2010.

[17] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, Y. Elovici. Link Prediction in Social Networks using Computationally Efficient Topological Features. 3rd IEEE SocialCom, 2011.

[18] P. Sarkar, D. Chakrabarti, A.W. Moore. Theoretical Justification of Popular Link Prediction Heuristics. IJCAI 2011: 2722-2727.

[19] T. Murata, S. Moriyasu. Link Prediction of Social Networks Based on Weighted Proximity Measures. 2007 IEEE/WIC/ACM International Conference on Web Intelligence.

[20] L. Lü, T. Zhou. Role of Weak Ties in Link Prediction of Complex Networks. CNIKM'09, November 6, 2009, Hong Kong, China.

[21] H. Rodrigues de Sá, R.B.C. Prudêncio. Supervised Link Prediction in Weighted Networks. IJCNN 2011, San Jose, California, USA, July 31 – August 5, 2011.

[22] T. Zhou, L. Lü, Y-C. Zhang. Predicting missing links via local information. European Physics Journal B 71(4): 623-630, 2009.

[23] D. Wang, D. Pedreschi, C. Song, F. Giannotti, A-L. Barabási. Human Mobility, Social Ties, and Link Prediction. KDD 2011, August 21-24, 2011, San Diego, California, USA.

[24] K. Jahanbakhsh, G.C. Shoja, V. King. Human Contact Prediction Using Contact Graph Inference. GREENCOM-CPSCOM '10. IEEE Computer Society, Washington, DC, USA, 813-818.

[25] K. Jahanbakhsh, V. King, G.C. Shoja. Predicting missing contacts in mobile social networks. WOWMOM 2011: 1-9