Centrality prediction in dynamic human contact networks

Hyoungshick Kim *, John Tang, Ross Anderson, Cecilia Mascolo

Computer Laboratory, University of Cambridge, United Kingdom

ARTICLE INFO

ABSTRACT

Real technological, social and biological networks evolve over time. Predicting their future topology has applications to epidemiology, targeted marketing, network reliability and routing in ad hoc and peer-to-peer networks. The key problem for such applications is usually to identify the nodes that will be in more important positions in the future. Previous researchers had used ad hoc prediction functions. In this paper, we evaluate ways of predicting a node's future importance under three important metrics, namely degree, closeness centrality, and betweenness centrality, using empirical data on human contact networks collected using mobile devices. We find that node importance is highly predictable due to both periodic and legacy effects of human social behaviour, and we design reasonable prediction functions. However human behaviour is not the same in all circumstances: the centrality of students at Cambridge is best correlated both daily and hourly, no doubt due to hourly lecture schedules, while academics at conferences exhibit rather flat closeness centrality, no doubt because conference attendees are generally trying to speak to new people at each break. This highlights the utility of having a number of different metrics for centrality in dynamic networks, so as to identify typical patterns and predict behaviour. We show that the best-performing prediction functions are 25% more accurate on average than simply using the previous centrality value. These prediction functions can be efficiently computed in linear time, and are thus practical for processing dynamic networks in real-time.

1. Introduction

Over the past few years, network models have been developed that shed new light on patterns of association and interaction in human societies, with implications for real world applications. A significant problem is to measure the centrality of nodes (or edges) in networks; network centrality can be used to identify important nodes for many applications such as targeted advertisement and recommendation [1], routing protocols [2–4], content sharing [5], epidemiological modelling [6,7], network reliability [8,9], resource provisioning [10], and urban planning [11]. Depending on the application, the importance of a node can have different meanings and hence several network centrality measures have been proposed, namely degree, closeness and betweenness centrality [12]. Degree centrality measures how many connections each node has and has been used to attack networks; for example, the police often disrupt criminal gangs by going after the ringleaders [8]. The same models work in epidemiology, where doctors may first vaccinate those individuals who are likely to come into contact with most others. Closeness centrality measures the average geodesic distance to all other nodes in the network and has been applied to the study of influence; targeted advertisements can favour people who can spread information quickly to other nodes in the network [13]. Finally, betweenness centrality identifies nodes which act as bridges between different groups of nodes, taking into account alternative communication paths between pairs of nodes in a network, which is useful in identifying bottlenecks in traffic networks [14].

However, when calculating such centrality measures, the current analysis makes two simplifying assumptions. First,

* Corresponding author. Tel.: +44 1223 33 46 06.
E-mail addresses: hk331@cl.cam.ac.uk, hk331@cam.ac.uk (H. Kim), jkt27@cl.cam.ac.uk (J. Tang), rja14@cl.cam.ac.uk (R. Anderson), cm542@cl.cam.ac.uk (C. Mascolo).

past studies have focussed on analysing static networks that do not change over time or aggregated networks built by collecting information over a period of time; or in other words, where relationships between nodes are known *a priori* [15]; and second, many opportunistic and delay-tolerant communication protocols [3,10,16,17] are designed on the assumption of the stationary nature of human contacts. However, in real life many networks are inherently dynamic. For example, friends are added and removed in online social networks; the topology of the internet changes with time; and contacts between mobile devices depend on the time of day. Therefore it is not prudent to assume stationary human behaviour in the design of practical applications.

A salient point is that since the end users of technological systems are humans, clearly the evolution of such systems will be driven by natural social patterns. For example, a simple routine of travelling to work every day brings a regular pattern of email communications, wifi hotspot connections, mobile phone bluetooth contacts and online social network activity, which in turn provides the periodicities seen in the underlying technological communication processes. This paper makes a crucial insight: since an individual's schedule is regular, if they are an important node at some time point, then it is highly likely that their importance will be correlated in the future.

We therefore set out to predict the state of such networks taking into account realistic schedules of human contact networks. In this study we show in fact that intuitive and simple prediction functions can be designed which take advantage of the predictability of such networks. Our key contributions can be summarised as follows:

- First, we show that empirical human contact networks are predictable and in particular, that there are clear correlations in node centrality values corresponding to natural human periods (i.e. 24 h) and legacy effects (read Section 4).
- Second, based on this observation we design several intuitive and simple prediction functions, to predict a node's future network centrality. We here focus on three exemplary metrics that are used widely: degree, closeness, and betweenness centrality. We evaluate their performance on real human contact datasets and show that the best-performing prediction functions are 25% more accurate on average than just using the last centrality value. Moreover, our experimental results show that the best approximation method and the optimal training time must be selected depending on the conditions of the prediction problem (read Section 5).

Our approach has two key advantages: (1) it is simple to implement and deploy since we only require the past centrality values of nodes, rather than tracing the geometric positions of nodes; (2) they require linear time $O(r)$ to approximate network centrality where r is the number of training time windows used. Our strategies are thus useful for large-scale and online computation – training data can be frequently updated in real time. We envisage that this

work can be easily integrated into dynamically evolving technological systems that require predictive capabilities driven by social processes.

The rest of this paper is organised as follows. Related work and potential applications will be discussed in the next section. In Section 3.2 we formally define the prediction problem and notation. In Section 4 we explore the temporal characteristics (e.g. periodic patterns) of human contact networks to predict network centrality by analysing the correlation between past and future centrality values. In Section 5, we evaluate the performance of the proposed approximation methods, and recommend how they should be used depending on the conditions. Finally, we make final conclusions in Section 6.

2. Related work

A number of pioneering paers [18–20] focussed on contact traces in order to gain insight about human movement patterns. Chaintreau et al. [17] found that contact duration and inter-contact time between humans can be represented by power-law distributions. Many real human contact traces [20–22] support this observation. Karagiannis et al. [23] show that inter-contact time follows a power-law closely up to twelve hours, with an exponential cut-off after that. Such results have been used to model potential future contacts but do not provide much insight into the problem of predicting future network structure.

At the most basic level, prediction in complex networks can be described in terms of the well-studied link prediction problem [24–26]. However our study aims at the prediction of a higher level metric which abstracts the reliance on geodesic or contact information to compute these values. Instead we want to predict node centrality directly.

Important observations on the periodicity of human behaviour have been made recently. Clauset and Eagle [27], Kim and Kotz [28], and Hsu et al. [29] showed periodic behaviour of human contact networks. In particular, Clauset et al. [27] show that the periodic patterns of human contacts are characterised by external calendar cycles. Also, Hui et al. [2] discuss how human contacts are distributed by time of day. Scherrer et al. [30] analysed the statistical characteristics of human networks from real datasets. They observed many temporal aspects of human networks; for example, there are clear periods of one day and variations from days to nights. Scellato et al. [31] analysed the temporal patterns of human networks based on time series analysis to quantify the amount of the information about the periodic patterns of human behaviour over time.

Centrality prediction in complex networks has been applied to a wide range of social-based forwarding schemes [2–4]. It has been proposed for Delay Tolerant Networks (DTNs), where the connection between nodes in the network frequently changes over time: the basic idea is to use node centrality for relay selections, and the forwarding strategy is to forward messages to nodes which are more central than the current node. Daly and Haahr [3] proposed a scheme based on ego-centric betweenness [32]. Hui et al. [4] consider node centrality based on social communities, and suggested some approximation methods to predict

network centrality values. They believe that the number of contacts in the last time window or the average contacts number on all previous windows can be used as reasonable approximation solutions. They simply used six hours as unit time under the assumption that human daily life is divided into four main periods—morning, afternoon, evening, and night—each almost six hours. In this paper we will further discuss the validity of using the average or the last centrality value and suggest the optimal time window size following intensive empirical analysis. Furthermore we discuss the feasibility of several other reasonable approximation methods which are carefully designed from the observation from real human contact networks.

3. Preliminaries

In this section we first define notation and terminology for centrality metrics and dynamic graphs, and then introduce the generalised network centrality prediction problem which will be used in the rest of the paper.

3.1. Network centrality measures

Formally, we use the standard definition of the degree, closeness and betweenness centrality values of a node u as follows [12]:

3.1.1. Degree centrality

$$Deg(u) = \frac{\kappa(u)}{|V| - 1}, \tag{1}$$

where $\kappa(u)$ is the number of edges of node u and V is the set of nodes in the network.

3.1.2. Closeness centrality

Closeness centrality measures how near nodes are to each other or in practical terms how quickly a node can communicate with all other nodes in a network. This is calculated for a node u as the average shortest path length to all other nodes in the network:

$$Clo(u) = \frac{1}{|V| - 1} \sum_{v \neq u \in V} dist(u, v), \tag{2}$$

where $dist(u, v)$ is the number of hops in the shortest path from node u to node v and V is the set of nodes in the network.

3.1.3. Betweenness centrality

Betweenness centrality measures the sum of the fraction of shortest paths through u, where the fraction is normalized for each single pair of other nodes x and y. More formally this is defined as:

$$Bet(u) = \sum_{x \neq u, y \neq u \in V} \frac{q_{x,y}(u)}{Q_{x,y}}, \tag{3}$$

where $Q_{x,y}$ is the total number of shortest paths starting from source node x and destination node y, and $q_{x,y}(u)$ are the number of shortest paths starting which actually pass through node u (starting from source node x and destination node y).

3.2. Dynamic graph model

We assume that the time during which a network is observed is finite, from t_{start} until t_{end}; without loss of generality, we set $t_{start} = 0$ and $t_{end} = T$. A dynamic network $G_{0,T}^{\mathcal{D}} = (V, E_{0,T})$ on a time interval $[0,T]$ consists of a set of vertices V and a set of temporal edges $E_{0,T}$ where a temporal edge $(u, v)_{i,j} \in E_{0,T}$ exists between vertices u and v on a time interval $[i,j]$ such that $i \leqslant T$ and $j \geqslant 0$. In other words, a dynamic network has a static set of vertices V while the set of edges can change over time.

Most characterisations of dynamic networks discretise time by converting temporal information into a sequence of network "snapshots" to apply techniques derived from graph theory to the analysis of networks [15,33]. For simplicity, the time period is divided into fixed discrete steps $\{1, \ldots, n\}$. We use w to denote the size of each time window, T/n, expressed in some time unites (e.g., seconds or hours). In other words, a dynamic network can be represented as a series of static graphs at each time, G_1, G_2, ..., G_n. The notation G_t ($1 \leqslant t \leqslant n$) represents the aggregate graph which consists of a set of vertices V and a set of edges E_t where an edge $(u, v) \in E_t$ exists only if a temporal edge $(u, v)_{i,j} \in E_{0,T}$ exists between vertices u and v on a time interval $[i,j]$ such that $i \leqslant w \cdot t$ and $j > w \cdot (t - 1)$. In other words, G_t is the tth temporal snapshot of the dynamic network $G_{0,T}^{\mathcal{D}}$ during tth time window. For simplicity, we use $start(t)$ and $end(t)$ to denote the starting time and the ending time of G_t, respectively.

For clarity, we introduce the following example. When $t_{start} = 0$, $t_{end} = 3$ and $w = 1$, the dynamic network with the set of temporal edges in Table 1 can be represented as the aggregated graph where all edges are aggregated into a single graph, $G_{1,3}^{S}$, or the series of static networks, G_1, G_2 and G_3 as we explained. The visual representations are shown in Fig. 1. Unlike the aggregated view of the graph $G_{1,3}^{S}$ in Fig. 1(a), we can see that the series of static networks, G_1, G_2 and G_3 in Fig. 1(b), represents temporal edge relationships effectively.

From Fig. 1(b), we compute the network centrality values of the nodes for G_1, G_2, and G_3 to show how these values change over time. For comparison, we also compute the centrality values of the nodes for $G_{1,3}^{S}$ in Fig. 1(a).

In Table 2, the centrality values of the nodes generally change over time as the corresponding network topology changes. For example, the node A's closeness values (0.333, 0.444, and 0.000) are continuously changed. By comparing these values with the node A's closeness value

Table 1
Example contacts in dynamic network.

Edge	Time interval
(A,C)	[1,1]
(A,D)	[2,2]
(B,D)	[2,3]
(C,D)	[3,3]

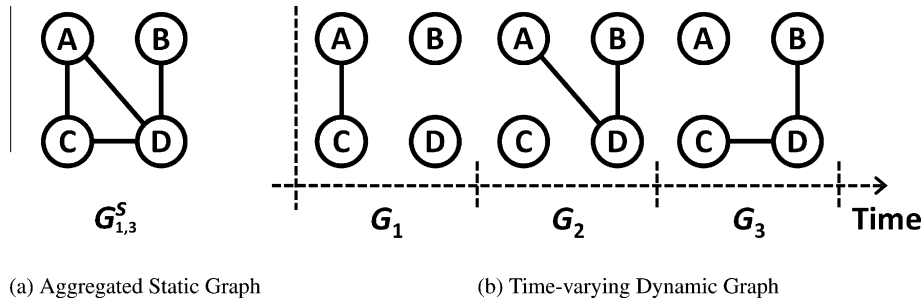(a) Aggregated Static Graph   (b) Time-varying Dynamic Graph

Fig. 1. Comparison of (a) aggregated representation and (b) time series representation of the contacts in Table 1.

Table 2
Network centrality of nodes in each graph.

Node	Graph	Deg (node)	Clo (node)	Bet (node)
A	G_1	0.333	0.333	0.000
	G_2	0.333	0.444	0.000
	G_3	0.000	0.000	0.000
	$G_{1,3}^S$	0.667	0.750	0.000
B	G_1	0.000	0.000	0.000
	G_2	0.000	0.444	0.000
	G_3	0.333	0.444	0.000
	$G_{1,3}^S$	0.333	0.600	0.000
C	G_1	0.333	0.333	0.000
	G_2	0.000	0.000	0.000
	G_3	0.333	0.444	0.000
	$G_{1,3}^S$	0.667	0.750	0.000
D	G_1	0.000	0.000	0.000
	G_2	0.333	0.667	0.333
	G_3	0.667	0.667	0.333
	$G_{1,3}^S$	1.000	1.000	0.667

(0.667) for $G_{1,3}^S$, we can see that the aggregated graph generally overestimates centrality values since it ignores disappearing edges.

3.3. Centrality prediction problem

We want to design a prediction function for the centrality value of a node in the dynamic network. For example, given a known historical dynamic network $G_{1,3}^D$ as the training input, how can we predict the node A's average closeness value in the future dynamic network, $G_{5,6}^D$ when we have a lag time of unknown interactions between time windows 3 and 6?

A reasonable solution is to use the average centrality value of the node A in G_1, G_2 and G_3. In other words, we can use 0.259 (=(0.333 + 0.444 + 0.000)/3) as the unknown, future closeness value of the node A in $G_{5,6}^D$. In fact, this idea is already used to select relay nodes for forwarding algorithms in Pocket Switched Networks (PSNs) [4].

We generalise the problem for predicting the average network centrality values of nodes as follows: given a dynamic network $G_{1,k}^D$ observed during k past time intervals, predict the average network centrality values of the nodes in the network during m future time intervals with l lagged time intervals, $G_{a,b}^D$, where $a = k + l$, $b = k + l + (m - 1)$, and $0 < k, l, m$. In this setting, the above example problem can be formulated with $k = 3$, $l = 2$, and $m = 2$. We represent the related variables visually in Fig. 2.
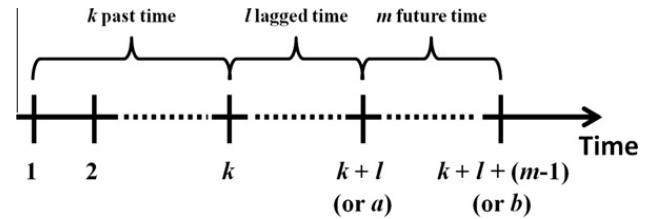

Fig. 2. The visual representation of k, l, and m.

The appropriate value of the parameter in this problem seems to depend on applications. For example, a small m is required to identify streets incurring temporal high traffic overhead during rush hours while in a model of disease spreading it seems more important to measure a node's long-term (or potential) centrality with a large m since overall central nodes may strongly affect the spreading of the disease.

We use $C_{i,j}(u)$ to denote the node u's average centrality value in $G_{i,j}^D = (V, E_{i,j})$ when $i \leqslant j$. In other words, $C_{i,j}(u) = (\sum_{t=i}^{j} c_t(u))/(j - i + 1)$ where $c_t(u)$ is the node u's centrality value such as $Deg(u)$, $Clo(u)$, or $Bet(u)$ in G_t. Similarly, we use $\widehat{C}_{i,j}(u)$ to denote the node u's predicted average centrality value between the time intervals i and j. With this notation, we formulate the problem on minimizing the average error between the guessed centrality values and the true centrality values. In other words, given $G_{1,k}^D$, l and m, find $\widehat{C}_{a,b}(u)$ where $a = k + l$, $b = a + (m - 1)$ for each $u \in V$ to minimise

$$Error(G_{1,k}^D, l, m) = \frac{\sum_{u \in V} |C_{a,b}(u) - \widehat{C}_{a,b}(u)|}{|V|}. \tag{4}$$

4. Human contact traces are predictable

4.1. Empirical dataset

We hypothesise that important nodes are more likely to be important at similar times in the future. To test this hypothesis we used three real human contact networks consisting of Bluetooth devices for detecting proximity devices through periodic Bluetooth scans. We summarise the datasets as follows:

(1) **MIT**: In the MIT Reality Mining project [34], 97 smart phones were deployed to students and staff at MIT over a period of 9 months. We here use the

human contact traces during the first week only. The Bluetooth scan interval is 5 min.

(2) **INFOCOM**: 78 iMotes, which are sensor boards equipped Bluetooth for detecting proximity devices, were deployed to the participants at the Infocom 2006 conference for 4 days. The Bluetooth scan interval is 2 min [17].

(3) **Cambridge**: In the Haggle project [35], 12 iMotes were deployed to the students for 6 days. The Bluetooth scan interval is 2 min.

Since a dynamic graph can be constructed using varying window sizes w, finding an appropriate w introduces a natural trade-off: by considering a larger w the accuracy of the measurements decreases since by neglecting the duration or the order of edge appearances, the temporal characteristics may be underestimated. However, the smaller we make w, the more expensive it is to analyse and to collect data. In real human contact networks, w should be reasonably small due to node mobility. The rate of topology change depends on many factors including node speeds and terrain. For these reasons, this section analyses the effects of increasing w from the finest granularity (equal to the scanning rate). In Section 5 we will discuss the effects of the size of w in predicting the centrality values.

4.2. Analysis of correlation between past and future centrality

Fig. 3 plots each nodes' closeness centrality value compared to its value in a past window, for the **MIT** dataset.

Notice three features: first, there is high correlation (0.565) between a node's closeness centrality value with its value 4 h ago which fits the intuition of a legacy effect (see the first scatter plot in Fig. 3); second, increasing the

difference decreases the correlation (-0.00832 at 12 h difference); and third, at 24 h difference the correlation rises again (0.432), which indicates possible periodic behaviour.

Generalising this analysis, we analyse the similarity between past and future centrality values, by calculating the average Pearson correlation coefficients among all possible pairs of C_t and $C_{t-\delta}$ for δ where $\delta \geqslant 1$ and $t \geqslant \delta + 1$, where C_t denotes the centrality values of nodes in the tth temporal network G_t. We note that the Pearson correlation is originally defined only if the standard deviations of the random variables are finite and are non-zero. However, in dynamic networks, the standard deviations can often be zero since the nodes often have the same centrality values (e.g. when the graph is totally disconnected). Here we assume that the Pearson correlation coefficient is zero when the standard deviation of a random variable is zero. This assumption helps exclude zero centrality values computed from disconnected networks.

For each dataset, the results of the average correlation coefficients are plotted in Figs. 4–6. In each figure, the X-axis, the Y-axis and the Z-axis are the time difference value δ, the time window size w, and the average correlation value, respectively.

With this, we make four key observations:

- First, as we might expect, recent past centrality values are strongly correlated compared with more distant values. For betweenness, however, this trend appears to be rather weak since the overall correlation coefficients are relatively very low compared with the other centrality metrics.
- Second, we can see the pattern of repeated peaks with 24-h time difference although this trend seems rather weak for **Infocom**. Probably, this is because people at
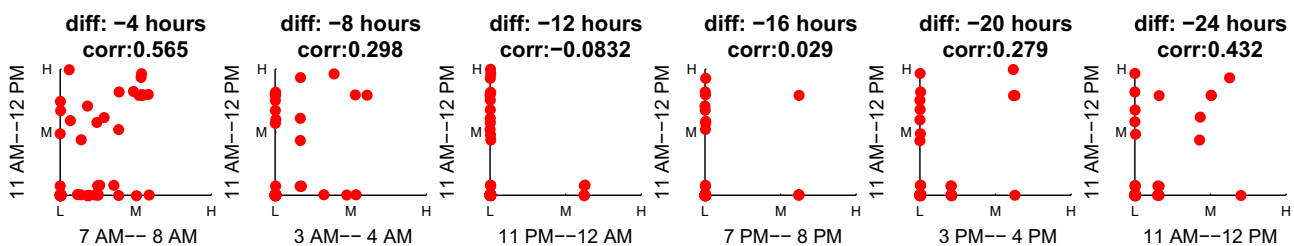

Fig. 3. Scatter plots depicting centrality correlation between a fixed window (y-axis) and an increasingly distant window from the past (x-axis) every four hours (left-to-right). We plot for `Closeness`, **MIT** (w = 1 h) for a typical weekday in September. The axis labels represent the low (L), medium (M), and high (H) centrality values.
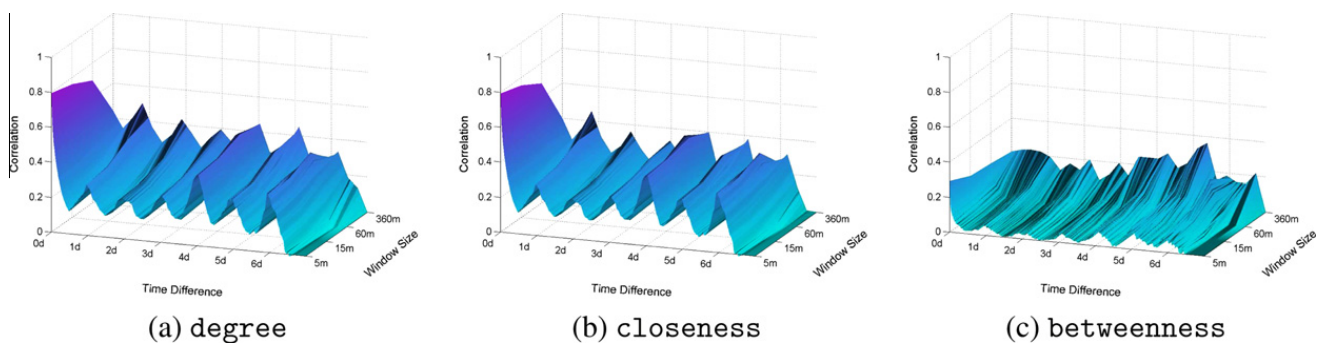

(a) degree   (b) closeness   (c) betweenness

Fig. 4. The average correlation analysis between C_t and $C_{t-\delta}$ for δ where $\delta \geqslant 1$ and $t \geqslant \delta + 1$ for **MIT**. For visualisation, we plotted the first week only.
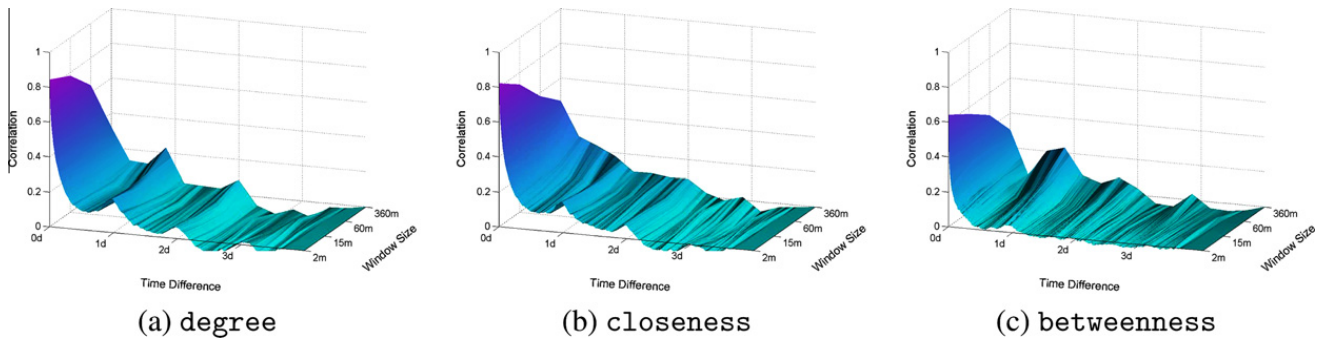
Fig. 5. The average correlation analysis between C_t and $C_{t-\delta}$ for δ where $\delta \geqslant 1$ and $t \geqslant \delta + 1$ for **Infocom**.
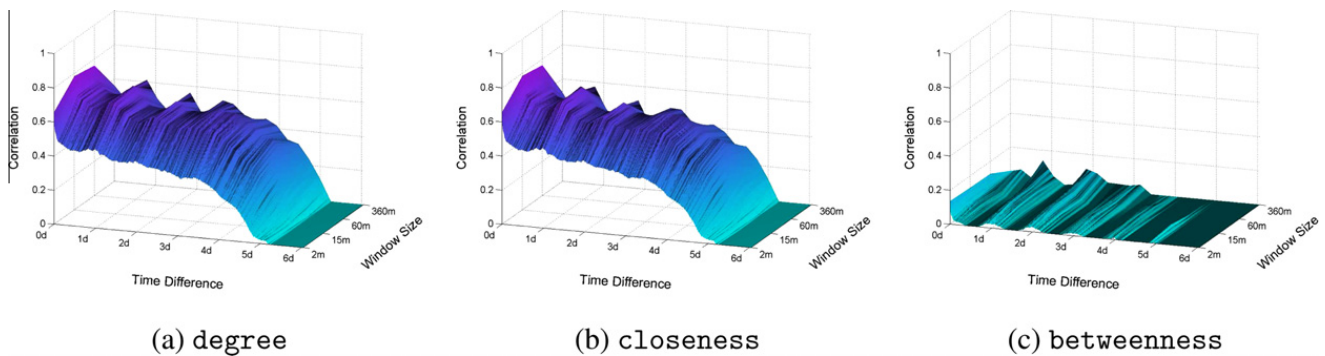
Fig. 6. The average correlation analysis between C_t and $C_{t-\delta}$ for δ where $\delta \geqslant 1$ and $t \geqslant \delta + 1$ for **Cambridge**.

a conference seek out new colleagues to talk to at the breaks between sessions, rather than socialising with the same people. However some academics are more sought after than others.

- Third, the average correlation coefficients for degree and closeness are much higher than those for betweenness. We assumed that the Pearson correlation coefficient is zero when the standard deviation of a random variable is zero. For betweenness, this exceptional case has been more often observed since indirect paths between nodes which are required for betweenness are very unlikely to be observed during the night time. So the computed average correlation coefficients may be rather underestimated.

- Fourth, the periodic patterns are clearly shown as a function of window size w, and longer windows give higher correlation, except in **Cambridge** where the best correlations correspond to a window size of 60 min. This presumably represents students sitting next to each other in lectures.

In Section 5, we will present prediction functions based on these observations.

4.3. Comparisons with a null model

To show the significance of the second and third observations we compare with a classic random network-evolution model, namely the edge-Markovian evolving graph (edge-MEG) [36]. An edge-MEG $G(n, p_i, r_b, r_d)$ is defined by four parameters, the number n of nodes, the edge probabil-

ity p_i of the initial graph, the edge birth-rate r_b, and death-rate r_d: start with the initial Erdös–Rényi random graph G_{n,p_i} [37], and at every time step, if an edge exists then it will die in the next time step with the probability r_d; while, if an edge does not exist, then it will appear at the next time step with the probability r_b. For comparison, we compute the average correlation coefficients for several edge-MEGs with $n = 100$ and $p_i = 0.2$. The results are shown in Fig. 7. We can see that there are no periodic patterns in them while the correlation coefficients are dramatically decreased up to about zero except for Fig. 7(a). Morevoer, the average correlation coefficients for betweenness (B.) are very similar to those for degree (D.) and closeness (C.) unlike human contact networks.

5. Predicting centrality values

5.1. Prediction functions

In dynamic networks, the network centrality values of nodes can change over time as new edges are created or existing ones removed. We want to predict these values from the node history. In practice, it is expensive to consider all mobile traces, so we evaluate simple approximation methods based on previous centrality values only.

We evaluate eight prediction functions summarised in Table 3, based on empirical observations seen in Section 4. The first observation is that the recent past network topologies are more similar to future network topologies than distant past network topologies. In other words, G_k may be more like G_{k-1} than G_{k-l} where $l > 1$. The second
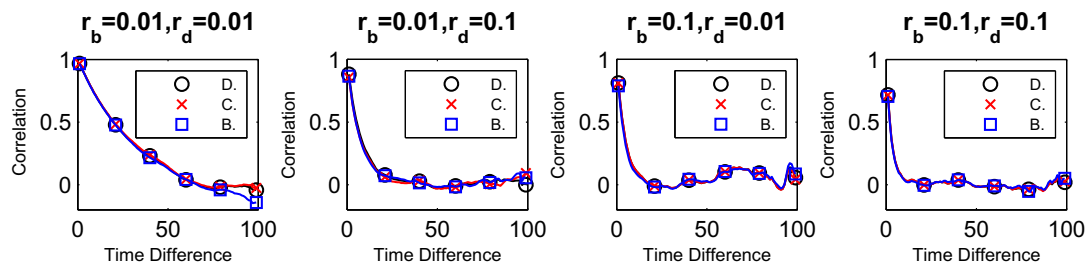
Fig. 7. The average correlation analysis between C_i and $C_{i-\delta}$ for δ where $\delta \geqslant 1$ and $i \geqslant \delta + 1$ for several edge-MEGs. While the recent past centrality values are highly correlated compared with more distant past centrality values, the subsequent peaks are rather irregular. In addition, the average correlation coefficients for betweenness are very similar to those for degree and closeness. These features are totally different from human contact networks.

Table 3
Summary of prediction functions.

Function	$\widehat{C}_{a,b}(u)$	Runtime	Description
Last	$C_k(u)$	$O(1)$	Last centrality
Uniform	$C_{k-(r-1),k}(u)$	$O(1)$	Uniform avg.
W-sqrt	$\sum_{i=1}^{r}((\sqrt{\delta})^{-1}/\omega_{total}) \cdot C_{k-(i-1)}(u)$	$O(k)$	Square root weighted avg.
W-linear	$\sum_{i=1}^{r}((\delta)^{-1}/\omega_{total}) \cdot C_{k-(i-1)}(u)$	$O(k)$	Linear weighted avg.
R-0.05	$\rho(k - 0.05)$	$O(k)$	Polynomial regression
R-0.15	$\rho(k - 0.15)$	$O(k)$	Polynomial regression
U-period	$\sum_{i=1}^{r}\left(\left(\left\vert\arcsin\cos\left(\frac{\pi \cdot \delta^{period}+\epsilon}{p}\right)\right\vert\right)^{-1}/\omega_{total}\right) \cdot C_{k-(i-1)}(u)$	$O(k)$	Uniform periodic avg.
W-period	$\sum_{i=1}^{r}\left(\left(\left\vert\arcsin\cos\left(\frac{\pi \cdot \delta^{period}+\epsilon}{p}\right)\right\vert \cdot \delta\right)^{-1}/\omega_{total}\right) \cdot C_{k-(i-1)}(u)$	$O(k)$	Weighted periodic avg.

observation is that human contact patterns are repeated periodically (e.g. 1 h, 24 h or 1 week). With three real datasets in Section 4, we will empirically analyse which predictor is really effective. We present several reasonable methods to minimise the objective function $Error(G_{1,k}^{\mathcal{D}}, l, m)$ discussed in Section 3.2. Let $a = k + l$, $b = a + (m - 1)$ for simplicity.

Last Centrality. As the first candidate, we just use the node's centrality value in the last temporal network (G_k) at time k. In other words, for $u \in V$, we use $C_k(u)$ as $\widehat{C}_{a,b}(u)$.

Uniform Average Centrality. In order to improve the accuracy of the prediction, we can use the node's r previous centrality values instead of one last previous centrality value. A reasonable idea is to use the node's uniform average centrality value between $G_{k-(r-1)}$, ..., G_{k-1}, G_k where $0 < r \leqslant k$ as the node's future centrality value. In other words, for $u \in V$, we use $C_{k-(r-1),k}(u)$ as $\widehat{C}_{a,b}(u)$.

We want to find the best r given the cost of computation and the accuracy of prediction, and will suggest values based on several real datasets in Section 3.3.

Note that although the runtime complexity of the uniform function is $O(k)$ for one prediction of $\widehat{C}_{a,b}(u)$, it gives an $O(1)$ amortised time per computation of $\widehat{C}_{a,b}(u)$ for m consecutive predictions where $m \geqslant k$. This property is practically useful since many applications [4,7,9] require the computation of consecutive centrality values over time.

Weighted Average Centrality. In order to consider the relative importance of the recent temporal networks, we can use the *weighted* average centrality value instead of the *uniform* average centrality value. Formally, $\widehat{C}_{a,b}(u)$ is com-

puted as $\sum_{i=1}^{r}(\omega_i/\omega_{total}) \cdot C_{k-(i-1)}(u)$ where $0 < r \leqslant k$ and $\omega_{total} = \sum_{i=1}^{r}\omega_i$. In fact, the *uniform* average centrality is a special case of the *weighted* average centrality when $\omega_i = 1/r$. We consider two reasonable weight assignments depending on the physical time difference δ between $G_{k-(i-1)}$ and $G_{a,b}^{\mathcal{D}}$. In other words, δ is the difference between $k - (i - 1)$ and $k + l$ ($=a$). Since time window size is w, δ is computed as $(l + i - 1) \cdot w$. We consider two weight assignment schemes, chosen empirically: square root $\left(\omega_i = \frac{1}{\sqrt{\delta}}\right)$ and linear $\left(\omega_i = \frac{1}{\delta}\right)$. The weighted assignments of square $\left(\omega_i = \frac{1}{\delta^2}\right)$ and cubic $\left(\omega_i = \frac{1}{\delta^3}\right)$ functions were also evaluated, but since they produce almost the same results of **Last**, we do not use them.

Polynomial Regression. An approximation is to use a polynomial regression model to predict a node's centrality value in the future. From a set of the node u's centrality values between $G_{k-(r-1)}$, ..., G_{k-1}, G_k, we can derive a finite sequence of r input–output pairs, $(1, c_{k-(r-1)}(u))$, ..., $(r - 1, c_{k-1}(u))$, $(r, c_k(u))$. From these data, we find the coefficients of a polynomial $\rho(\cdot)$ of degree m that fits the data, $\rho(i)$ to $c_{k-r+i}(u)$ where $1 \leqslant i \leqslant r$, in a least squares sense. This polynomial $\rho(\cdot)$ is used to predict the node u's centrality value at the target time. We here use $\rho(k - \epsilon)$ as $\hat{c}_{a,b}(u)$. We empirically tested several ϵ values and degrees of polynomials to find the best one. Our recommendation is to select degree 3. We also recommend using a small number for ϵ, less than 0.2. We also used 0.05 and 0.15 as representative values in our experiments.

Periodic Intervals. In general, human activities are repeated periodically and as we have shown in Section 4.2,

important nodes are also correlated with such periods; hence an intuitive method is to use these periodic patterns to improve the accuracy of the prediction.

For human contact networks, reasonable periods are a day or week. Given the period p of a day or a week, we consider an approximation as a special case of the *weighted average centrality*. The periodic physical time difference δ^{period} between $G_{k-(i-1)}$ and $G_{a,b}^{\mathcal{D}}$ where $1 \leqslant i \leqslant r$ is computed as follows:

$$\delta^{\text{period}} = \min_{j=a}^{b}\{(j - (k - (i-1))) \cdot w \bmod p\}. \tag{5}$$

In practice, δ^{period} can be efficiently computed in $O(1)$ time as follows:

$$\delta^{\text{period}} = \begin{cases} 0 & \text{if } \delta_{\text{end}}^{\text{period}} \leqslant (b-a) \cdot w, \\ \min\left\{\delta_{\text{start}}^{\text{period}}, \delta_{\text{end}}^{\text{period}}\right\} & \text{otherwise}, \end{cases} \tag{6}$$

where $\delta_{\text{start}}{}^{\text{period}} = (a - (k - (i-1))) \cdot w \bmod p$ and $\delta_{\text{end}}^{\text{period}} = (b - (k - (i-1))) \cdot w \bmod p$.

So we assign the relative weights with δ^{period} for each $c_{k-(i-1)}(u)$ where $1 \leqslant i \leqslant r$ and r is the number of centrality values used in predicting $\hat{c}_{a,b}(u)$ as follows: $\omega_i = \left(\left|\arcsin \cos\left(\frac{\pi \cdot \delta^{\text{period}} + \epsilon}{p}\right)\right|\right)^{-1}$ where ϵ is a very small number (close to zero). (ϵ is used to avoid divide-by-zero exceptions when $\delta^{\text{period}} = 0$.)

In addition, we consider both the relative importance of the recent temporal networks and the periodic intervals at the same time. As an approximation, we propose the following weight assignment function: $\omega_i = \left(\left|\arcsin \cos\left(\frac{\pi \cdot \delta^{\text{period}} + \epsilon}{p}\right)\right| \cdot \delta\right)^{-1}$ where ϵ is a very small number (close to zero) and $\delta = (l + i - 1) \cdot w$.

5.2. Evaluation of prediction functions

We analyse the performance of each approach on real human contact datasets used in Section 3.3. For each dataset, we calculated $Error(G_{1,k}^{\mathcal{D}}, l, m)$ discussed in Section 3.2 by varying l, m, w and r. We use the symbol E to denote this value in the resulting figures. These parameters are summarized in Table 4. The aim of the experiment is to evaluate feasibility and usefulness of each function and to find the optimal parameter values (e.g. r) of each prediction function at the same time.

The performance of all prediction functions except **Last** is primarily determined by the choice of r as well as l, m

Table 4
Summary of parameters in prediction functions.

Parameters	Description
k	The most recent observation time window for training input $G_{1,k}^{\mathcal{D}}$
l	The number of the (unknown) lag time windows
m	The number of future time windows to be predicted
w	The size of a time window for $G_{1,k}^{\mathcal{D}}$
r	The number of training time windows used

and w. However, choosing a suitable r value is not easy when some past centrality values are not strongly correlated. As r increases, moreover, the cost of computation increases. So we shall consider finding the optimal r by fixing some reasonable l, m and w values.

First of all, we select the time window size w as small as possible. We consider the finest granularity of temporal characteristics with the smallest w (2 min or 5 min). We will revisit the effects of the window size w later. In addition, we fix $l = (48\,\text{h})/w$ and $m = (48\,\text{h})/w$ hours. In many applications such as routing protocols and epidemic modelling, the centrality prediction for a larger m is more important. It also seems reasonable to consider some lagged time l since it is difficult to collect human contact traces in real time. We will discuss the effects of l and m later.

The prediction results by varying r from $(3\,\text{h})/w$ to $(72\,\text{h})/w$ are shown in Figs. 8–10. For improved visualisation, we use the same range on the y-axis between degree and closeness only per dataset since the levels of accuracy and precision are totally different between datasets and centrality types (e.g. betweenness). For example, prediction for the **MIT** dataset is capable of higher precision than **Cambridge** – which may have a significant error level because of the small sample size.

In **MIT** (see Fig. 8), the **U-period** prediction function produced the best results. This is because the periodic patterns of human contact traces are clearly shown in **MIT**. When around $(60\,\text{h})/w$ is used, **U-period** particularly achieved the best $Error(G_{1,k}^{\mathcal{D}}, l, m)$ values below 0.016 in degree and closeness and 0.0013 in betweenness and then the $Error(G_{1,k}^{\mathcal{D}}, l, m)$ values increased since this time interval. This may imply that the centrality value at a specific time of an average node was not significantly affected by the distant past centrality values 60 h ago in **MIT**.

Although the performance of **Uniform** is not as strong as that of the **U-period** function, it outperforms the other prediction functions. Considering that its computation cost of **Uniform** is also relatively cheap, we recommend using **Uniform** as an alternative. However, we would not recommend using **Last** because its relative accuracy is not enough. Overall, the $Error(G_{1,k}^{\mathcal{D}}, l, m)$ values decreases in all prediction functions when a larger r is used.

By contrast, in **Infocom** (see Fig. 9), the **U-period** and **W-period** prediction functions are not good options when a larger r is particularly used. Instead, our recommendations would be to use **W-sqrt**, **W-linear** or **Uniform**. We already observed that there is no noticeable periodic patterns while the recent past centrality values are strongly correlated in **Infocom** – Fig. 5 illustrates this.

Another interesting observation in **Infocom** is that all of the prediction functions for degree significantly outperformed those for closeness unlike the other human contact traces from campuses. Relations among neighbours are maintained well in the conference but the overall network topologies change continuously over time.

In **Cambridge** (see Fig. 10), the accuracy of **U-period** and **W-period** prediction functions is not good with a larger r for degree and closeness, which is similar to **Infocom**. However, for betweenness, **U-period** outperforms the other prediction functions. The **U-period** with around $r =$
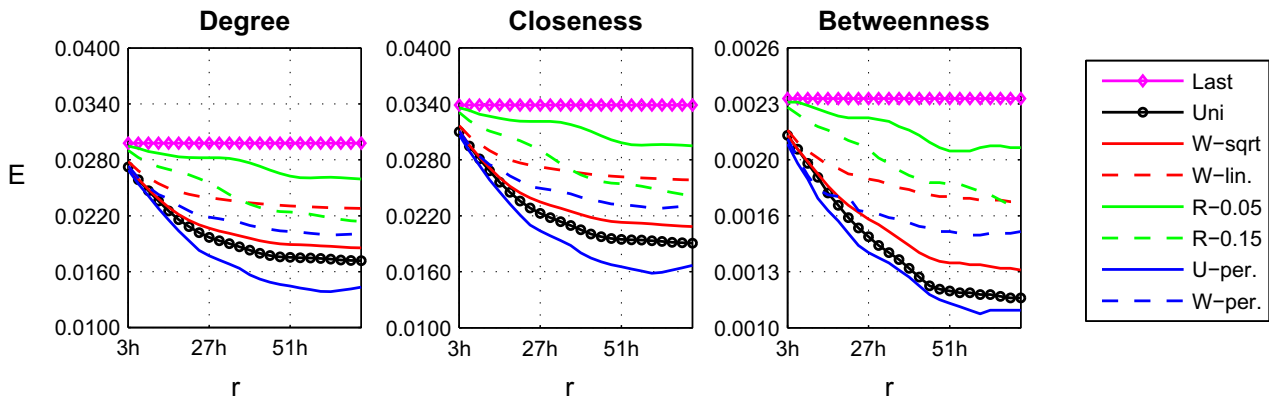
Fig. 8. The prediction results by varying r for **MIT** ($l = (48\,\text{h})/w$, $m = (48\,\text{h})/w$ and $w = 5$ min).
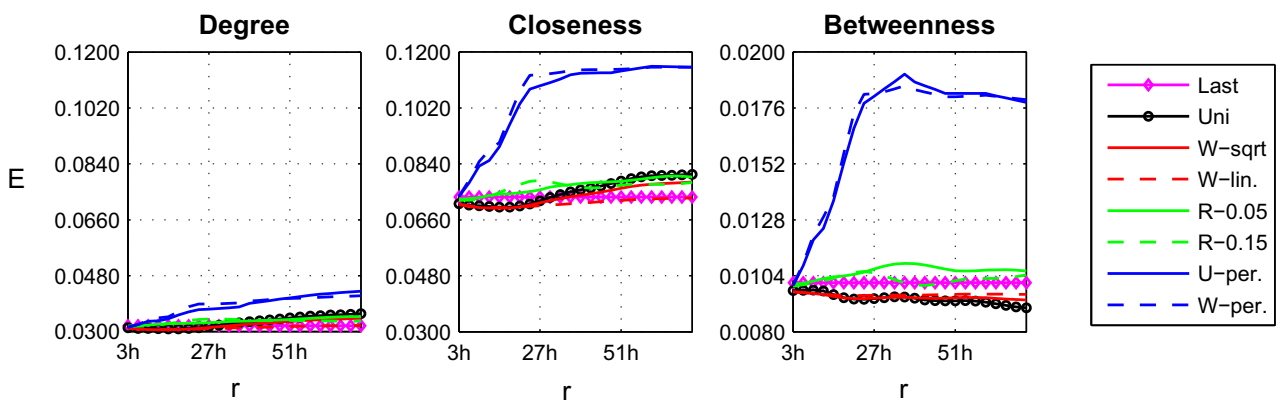

Fig. 9. The prediction results by varying r for **Infocom** ($l = (48\,\text{h})/w$, $m = (48\,\text{h})/w$ and $w = 2$ min).
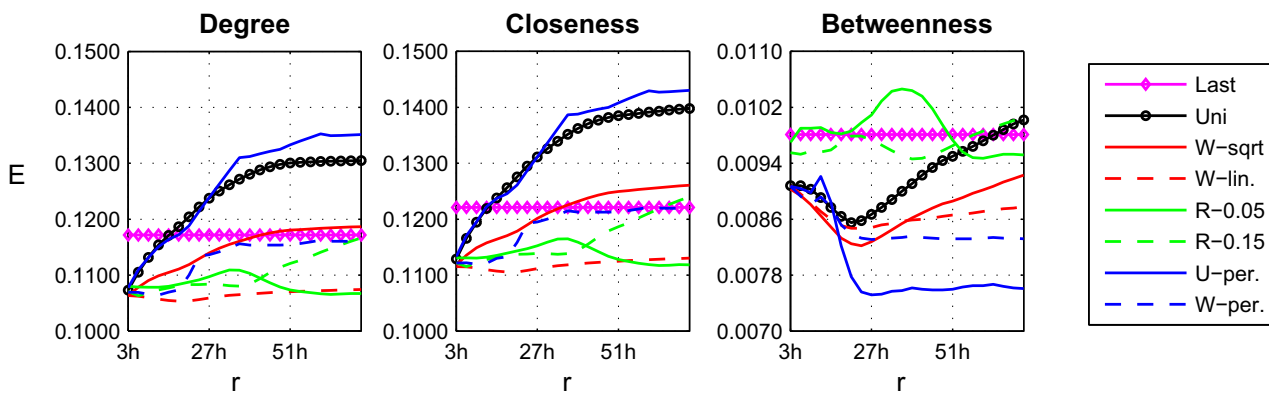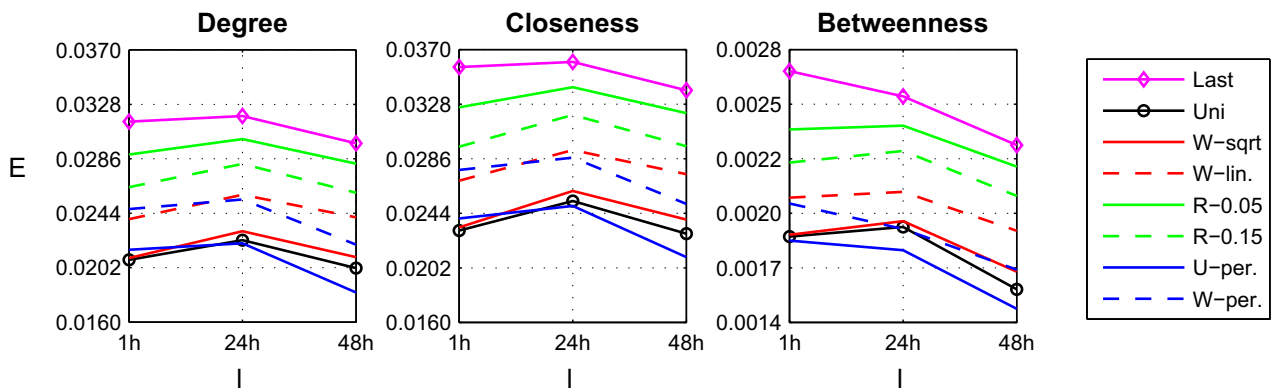

Fig. 10. The prediction results by varying r for **Cambridge** ($l = (48\,\text{h})/w$, $m = (48\,\text{h})/w$ and $w = 2$ min).

$(27\,\text{h})/w$ achieved the $Error(G_{1,k}^{\mathcal{D}}, l, m)$ values below 0.00078 in betweenness. For degree and closeness, **W-linear** with around $r = (27\,\text{h})/w$ hours seems a good choice.

In order to show this more effectively, we computed the ratio of the **Best** prediction function and the **Last** prediction function. The results are shown in Table 5. We show that the best-performing prediction functions are 25% more accurate on average compared to **Last**. In particular, in **MIT**, the $Error(G_{1,k}^{\mathcal{D}}, l, m)$ value for `betweenness` of **U-period** is only about 46.8% of **Last**. When we try to predict the network centrality values in human contact traces like **MIT** is a realistic dataset, this result is very encouraging.

We now discuss the effects of the time lag l (see Figs. 11–13). To demonstrate this we fix $r = (24\,\text{h})/w$, $m = (48\,\text{h})/w$ and $w = 2$ min for **Infocom** and **Cambridge** (or 5 min for **MIT**).

Interestingly, the $Error(G_{1,k}^{\mathcal{D}}, l, m)$ values generally increase with the time lag l for degree and closeness in **Infocom** and **Cambridge**. In particular, **U-period** and **W-period** are not effective as l increases. However, the results were somewhat inconsistent in the other cases. We can see that an increase in l does not significantly affect the $Error(G_{1,k}^{\mathcal{D}}, l, m)$ values of the prediction functions for all centrality types in **MIT** and betweenness in **Cambridge**,

Table 5
Ratio of "the results of the **Best** prediction function" and "the results of the **Last** prediction function". Here, *h* represents hour (time).

Network	Centrality	Best			Last	Ratio
		Function	*r*	Error	Error	
MIT	degree	U-period	63 h	0.0139	0.0298	0.4653
	closeness	U-period	60 h	0.0158	0.0339	0.4672
	betweenness	U-period	60 h	0.0011	0.0023	0.4675
Infocom	degree	W-sqrt	12 h	0.0306	0.0319	0.9596
	closeness	W-sqrt	15 h	0.0697	0.0733	0.9505
	betweenness	Uniform	72 h	0.0090	0.0101	0.8932
Cambridge	degree	W-linear	18 h	0.1054	0.1172	0.8991
	closeness	W-linear	18 h	0.1106	0.1221	0.9055
	betweenness	U-period	27 h	0.0075	0.0098	0.7666

Fig. 11. The prediction results by varying *l* for **MIT** (*r* = (24 h)/*w*, *m* = (48 h)/*w* and *w* = 5 min).
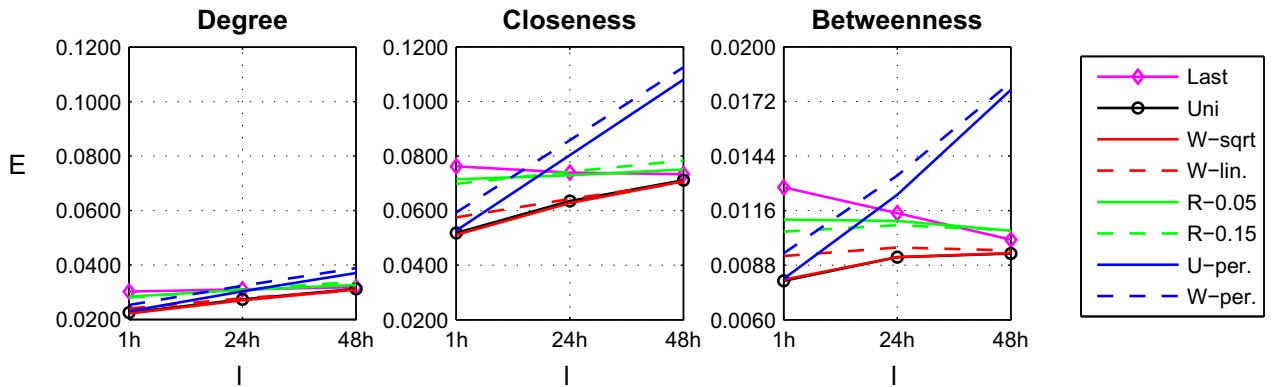
Fig. 12. The prediction results by varying *l* for **Infocom** (*r* = (24 h)/*w*, *m* = (48 h)/*w* and *w* = 2 min).
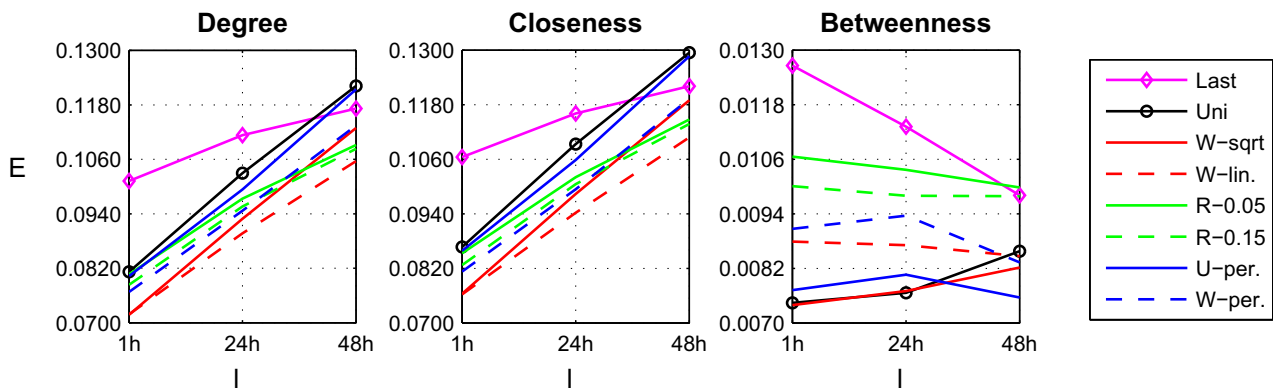
Fig. 13. The prediction results by varying *l* for **Cambridge** (*r* = (24 h)/*w*, *m* = (48 h)/*w* and *w* = 2 min).

which have apparent periodic patterns of human contact traces. Intuitively, we expect that **U-period** (or **W-period**) performs well even for a large l (e.g. $l = 48$ h) if the periodic patterns exist in human contact traces. Overall, when $l = (1\ \text{h})/w$, **W-sqrt** also performs well. So our recommendation would be to use **W-sqrt** for a small l.

Turning our attention to the case of m future time intervals to be predicted, see Figs. 14–16. We fix $r = (24\ \text{h})/w$, $l = (1\ \text{h})/w$ and $w = 2$ min for **Infocom** and **Cambridge** (or 5 min for **MIT**). We here select $l = (1\ \text{h})/w$ to minimise the effects of l. As m increases, the $Error(G^{\mathcal{D}}_{1,k}, l, m)$ values of **Last** generally increase while those of **U-period**, **W-period**,

W-sqrt and **Uniform** commonly decrease. In particular, our analysis show a marginally significant change in these values between $m = (1\ \text{h})/w$ and $m = (24\ \text{h})/w$. This in fact matches our intuition: the relative importance of the recent centrality values is reduced as m increases. Therefore our recommendations would be to use **U-period**, **W-period**, **W-sqrt** and **Uniform** for a large m. However, we recommend that **Last** should be used for degree and closeness and **W-linear** for `betweenness`, respectively, as alternatives for a small m. These results are natural consequences since the relative importance of recent network topologies decreases when m increases.
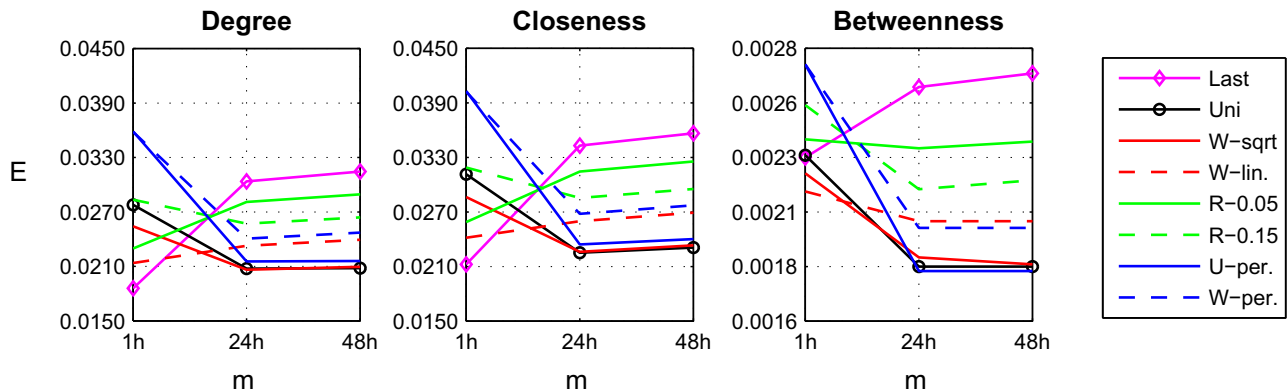

Fig. 14. The prediction results by varying m for **MIT** ($r = (24\ \text{h})/w$, $l = (1\ \text{h})/w$ and $w = 5$ min).
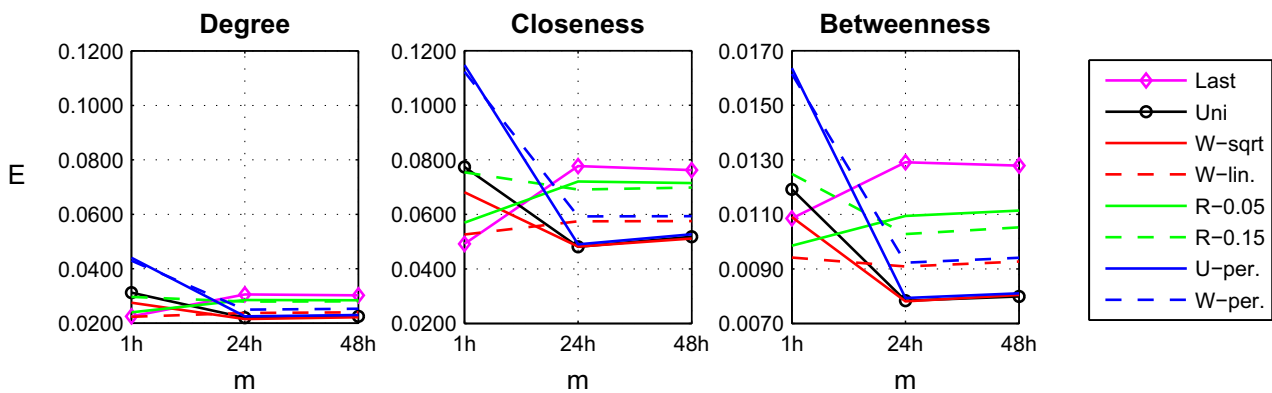

Fig. 15. The prediction results by varying m for **Infocom** ($r = (24\ \text{h})/w$, $l = (1\ \text{h})/w$ and $w = 2$ min).
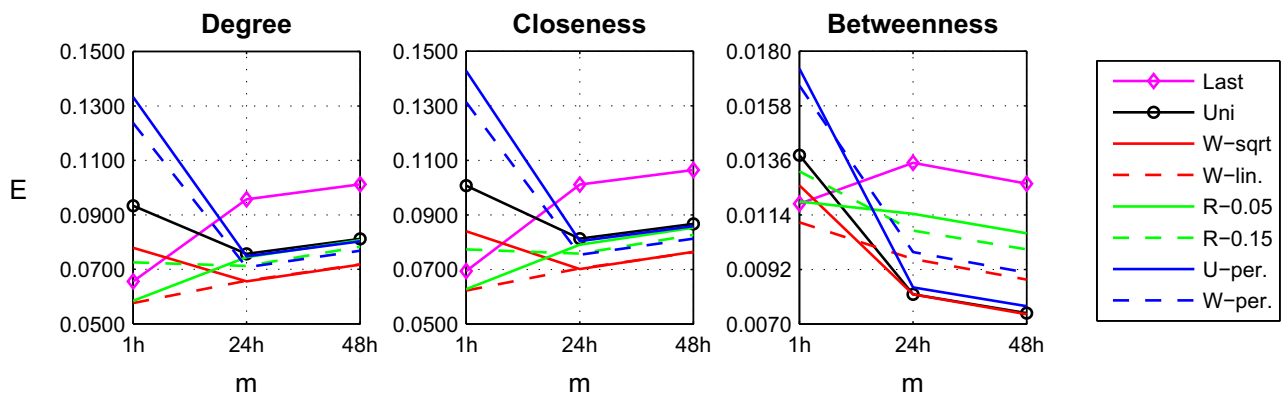

Fig. 16. The prediction results by varying m for **Cambridge** ($r = (24\ \text{h})/w$, $l = (1\ \text{h})/w$ and $w = 2$ min).

Finally, we discuss how the performance of prediction functions may change with the window size w. As w increases, the temporal characteristics of human contacts are underestimated, but the costs of collecting and analysing traces decrease. Figs. 17–19 show the effects of varying w from 2 min (or 5 min) to 60 min. To demonstrate this we fix $r = (24\ \text{h})/w$, $l = (48\ \text{h})/w$ and $m = (48\ \text{h})/w$.

As w increases, the $Error(G_{1,k}^{D}, l, m)$ values of all prediction functions (except **U-period** and **W-period** for `between-ness` in **Infocom**) increase. It is natural; a finer granularity may be desired to improve the accuracy. In particular, the performance of **U-period** and **W-period** is apparently deteriorated. We can see this trend in **MIT** (see Fig. 17). So we

recommend that **Uniform** should be used as alternatives for a large w. As we already discussed in Fig. 9, the **U-period** and **W-period** performed badly for `betweenness` in **Infocom**. Therefore the prediction results can be rather improved when we reduce the effects induced by periodic weights.

In summary, our recommendation would be to use **U-period** for human contact networks which can be assumed a relatively stable environment with the periodic patterns of human contacts. However, we would not recommend using **U-period** when the network topology (e.g. participant contacts in a conference) is unstable and changes rapidly; more obvious recommendations would be to use the
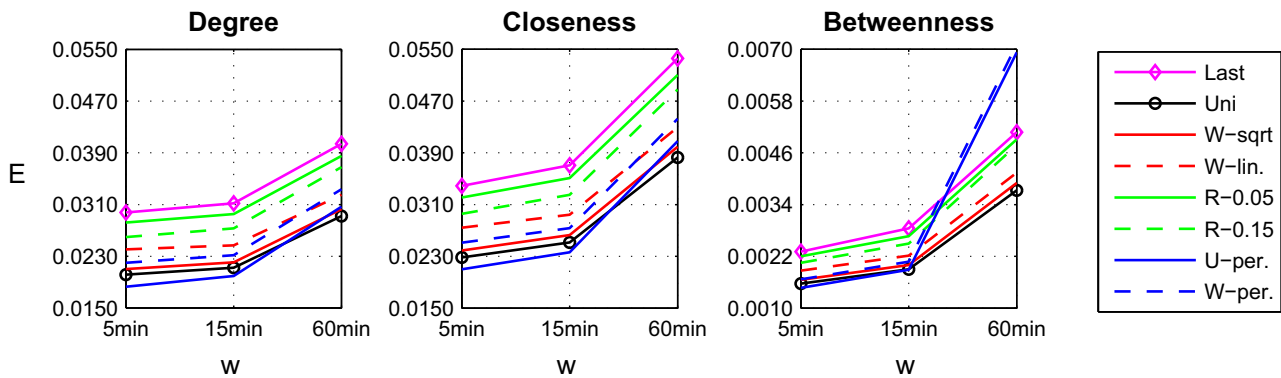

Fig. 17. The prediction results by varying w for **MIT** ($r = (24\ \text{h})/w$, $l = (48\ \text{h})/w$ and $m = (48\ \text{h})/w$).
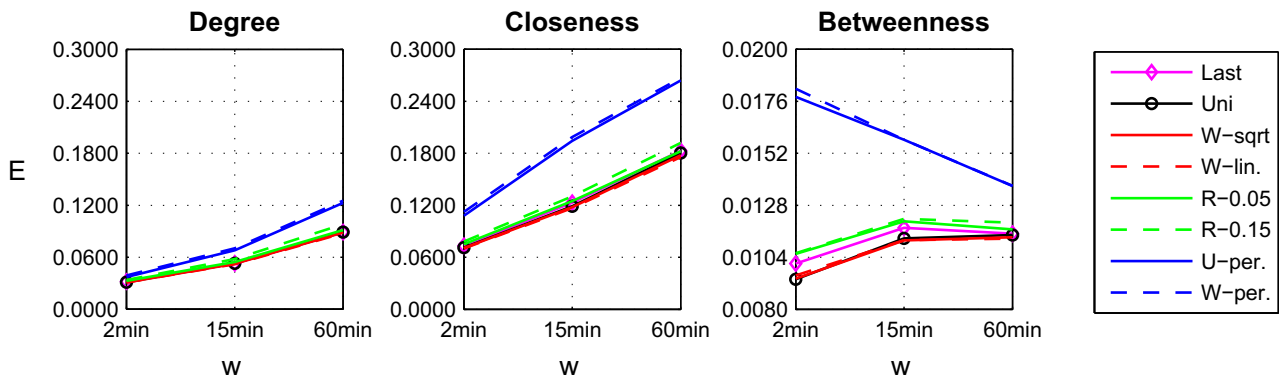

Fig. 18. The prediction results by varying w for **Infocom** ($r = (24\ \text{h})/w$, $l = (48\ \text{h})/w$ and $m = (48\ \text{h})/w$).
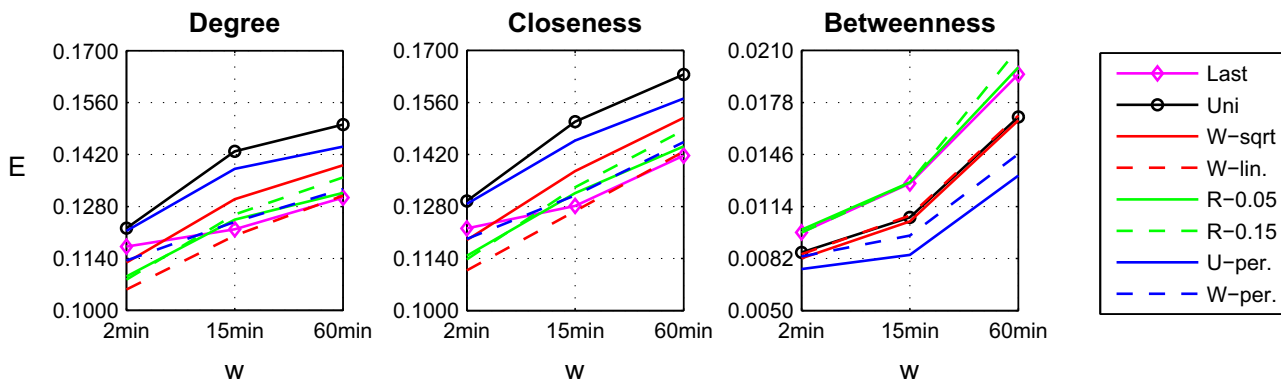

Fig. 19. The prediction results by varying w for **Cambridge** ($r = (24\ \text{h})/w$, $l = (48\ \text{h})/w$ and $m = (48\ \text{h})/w$).

average of a few previous centrality values with relative weights. We can see that **W-sqrt** or **W-linear** performed well even with a relatively small r in **Infocom**. Inherently, the recent centrality values will become more important as m decreases. Therefore we would recommend using **W-linear** or **Last** when m is very small.

6. Conclusion and future work

Measuring network centrality is an important problem for many applications. Most existing studies have focussed on analysing static networks, while in reality this assumption is not reliable since many networks are inherently dynamic; connections are added or removed over time. Previous writers had used ad hoc methods to predict centrality; we studied this intensively using empirical data from three human contact networks.

We presented eight prediction functions and explored their feasibility. Our design principles were based on two empirical observations: the relative importance of centrality values with elapsed time and the periodic repeatability of human contact patterns. We analysed centrality prediction functions by computing the difference between observed and predicted values. We discussed which prediction functions are generally recommended under which conditions. When human contact traces are clearly repeated, the accuracy of prediction can significantly be improved by taking this into account.

Our approach is simple to implement and deploy since only past centrality values are required for prediction, not node position, and since our predictors can be computed in time linear in the number of training time windows used. These strategies are thus amenable for large-scale, online and real-time computation.

As part of this ongoing study, we plan to analyse additional human contact traces to study their common characteristics and also employ more advanced techniques from signal processing, such as matched filters, to improve centrality prediction accuracy. Another interesting problem is the question of the sensitivity of centrality values to the underlying edge process which we intend to consider in future work.

Acknowledgements

References

[1] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering 17 (6) (2005) 734–749.

[2] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, C. Diot, Pocket switched networks and human mobility in conference environments, in: Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking, WDTN'05, ACM, New York, NY, USA, 2005, pp. 244–251.

[3] E.M. Daly, M. Haahr, Social network analysis for routing in disconnected delay-tolerant MANETs, in: Proceedings of the 8th ACM International Symposium on Mobile Ad hoc Networking and Computing, MobiHoc'07, ACM, New York, NY, USA, 2007, pp. 32–40.

[4] P. Hui, J. Crowcroft, E. Yoneki, Bubble rap: social-based forwarding in delay tolerant networks, in: Proceedings of the 9th ACM international symposium on Mobile Ad hoc Networking and Computing, MobiHoc'08, ACM, New York, NY, USA, 2008, pp. 241–250.

[5] L. McNamara, C. Mascolo, L. Capra, Media sharing based on colocation prediction in urban transport, in: Proceedings of the 14th ACM International Conference on Mobile Computing and Networking, MobiCom'08, ACM, New York, NY, USA, 2008, pp. 58–69.

[6] J. Kleinberg, The wireless epidemic, Nature 449 (2007) 287–288.

[7] J. Tang, C. Mascolo, M. Musolesi, V. Latora, Exploiting temporal complex network metrics in mobile malware containment, in: Proceedings of the 12th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM 2011), 2011, pp. 1–9.

[8] R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks, Nature 406 (6794) (2000) 378–382.

[9] S. Nagaraja, R. Anderson, The topology of covert conflict, in: Proceedings of the 5th Workshop on The Economics of Information Security (WEIS 2006), 2006.

[10] L. Song, U. Deshpande, U.C. Kozat, D. Kotz, R. Jain, Predictability of WLAN Mobility and its Effects on Bandwidth Provisioning, in: Proceedings of the 25th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), Barcelona, Spain, 2006.

[11] J. Reades, F. Calabrese, A. Sevtsuk, C. Ratti, Cellular census: explorations in urban data collection, IEEE Pervasive Computing 6 (2007) 30–38.

[12] S. Wasserman, K. Faust, Social Networks Analysis, Cambridge University Press, 1994.

[13] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of ACM SIGKDD'03, 2003, pp. 137–146.

[14] P. Holmes, Congestion and centrality in traffic flow on complex networks, Advances in Complex Systems 6 (2) (2003) 163–176.

[15] J. Tang, M. Musolesi, C. Mascolo, V. Latora, Temporal distance metrics for social network analysis, in: Proceedings of ACM WOSN'09, 2009.

[16] A. Lindgren, A. Doria, O. Schelén, Probabilistic routing in intermittently connected networks, ACM SIGMOBILE Mobile Computing and Communications Review 7 (2003) 19–20.

[17] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, J. Scott, Impact of human mobility on opportunistic forwarding algorithms, IEEE Transactions on Mobile Computing 6 (2007) 606–620.

[18] A. Balachandran, G.M. Voelker, P. Bahl, P.V. Rangan, Characterizing user behavior and network performance in a public wireless lan, ACM SIGMETRICS Performance Evaluation Review 30 (2002) 195–205.

[19] M. Balazinska, P. Castro, Characterizing mobility and network usage in a corporate wireless local-area network, in: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, MobiSys'03, ACM, New York, NY, USA, 2003, pp. 303–316.

[20] T. Henderson, D. Kotz, I. Abyzov, The changing usage of a mature campus-wide wireless network, in: Proceedings of the 10th Annual International Conference on Mobile Computing and Networking, MobiCom'04, New York, NY, USA, 2004, pp. 187–201.

[21] J. Su, A. Chin, A. Popivanova, A. Goel, E. de Lara, User mobility for opportunistic ad-hoc networking, in: Proceedings of the Sixth IEEE Workshop on Mobile Computing Systems and Applications, IEEE Computer Society, Washington, DC, USA, 2004, pp. 41–50.

[22] M. McNett, G.M. Voelker, Access and mobility of wireless pda users, ACM SIGMOBILE Mobile Computing and Communications Review 9 (2005) 40–55.

[23] T. Karagiannis, J.-Y. Le Boudec, M. Vojnović, Power law and exponential decay of inter contact times between mobile devices, in: Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking, MobiCom'07, ACM, New York, NY, USA, 2007, pp. 183–194.

[24] D. Liben-Nowell, J. Kleinberg, The link prediction problem for social networks, in: Proceedings of the Twelfth International Conference on Information and Knowledge Management, ACM, New Orleans, LA, USA, 2003, pp. 556–559.

[25] L. Lü, T. Zhou, Role of weak ties in link prediction of complex networks, in: Proceeding of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management, CNIKM'09, ACM, New York, NY, USA, 2009, pp. 55–58.

[26] L. Getoor, C.P. Diehl, Link mining: a survey, SIGKDD Explorations Newsletter 7 (2005) 3–12.

[27] A. Clauset, N. Eagle, Persistence and periodicity in a dynamic proximity network, in: Proceedings of Workshop on Computational Methods for Dynamic Interaction Networks (DIMACS), 2007.

[28] M. Kim, D. Kotz, Periodic properties of user mobility and access-point popularity, Personal and Ubiquitous Computing 11 (2007) 465–479.

[29] W.-J. Hsu, T. Spyropoulos, K. Psounis, A. Helmy, Modeling spatial and temporal dependencies of user mobility in wireless mobile networks, IEEE/ACM Transactions on Networking (TON) 17 (2009) 1564–1577.

[30] A. Scherrer, P. Borgnat, E. Fleury, J.L. Guillaume, C. Robardet, Description and simulation of dynamic mobility networks, Computer Networks 52 (2008) 2842–2858.

[31] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, On nonstationarity of human contact networks, in: Proceedings of the 2nd Workshop on Simplifying Complex Networks for Practitioners (SIMPLEX), 2010, pp. 105–111.

[32] P.V. Marsden, Egocentric and sociocentric measures of network centrality, Social Networks 24 (4) (2002) 407–422.

[33] J. Tang, S. Scellato, M. Musolesi, C. Mascolo, V. Latora, Small-world behavior in time-varying graphs, Physical Review E 81 (5) (2010) 055101.

[34] N. Eagle, A. (Sandy) Pentland, Reality mining: sensing complex social systems, Personal and Ubiquitous Computing 10 (4) (2006) 255–268.

[35] Haggle Project, 2009. <http://www.haggleproject.org>.

[36] S. Wasserman, Analyzing social networks as stochastic processes, Journal of the American Statistical Association 75 (370) (1980) 280–294.

[37] B. Bollobas, Random Graphs, Cambridge University Press, 2001.

Hyoungshick Kim is a Ph.D. candidate in the Computer Laboratory at the University of Cambridge as a Ph.D. student. He received the B.S. degree from the Department of Information Engineering at Sungkyunkwan University in Korea and M.S. degree from the Department of Computer Science at KAIST in Korea, in 1999 in 2001, respectively. He previously worked for Samsung Electronics as a senior engineer from May 2004 to September 2008. He also served as a member of DLNA and Coral standardization for DRM interoperability in home networks. His current research interest is focused on privacy and anonymity in complex networks and distributed systems.

John Tang is an EPSRC Doctoral Student at the Computer Laboratory, University of Cambridge. His research lies in temporal extensions to complex network analysis motivated by the need for more realistic models to analyse many technological, social and biological systems which change over time. Prior to starting his Ph.D., John graduated in Computer Science from UCL in 2006 and worked for two years as a technologist in the banking industry.

Ross Anderson is Professor of Security Engineering at Cambridge University. He is one of the founders of a vigorously-growing new academic discipline, the economics of information security. Ross was also a seminal contributor to the idea of peer-to-peer systems and an inventor of the AES finalist encryption algorithm "Serpent". He also has well-known publications on many other technical security topics including hardware tamper-resistance, emission security, copyright marking, and the robustness of application programming interfaces (APIs). He is a Fellow of the Royal Society, the Royal Academy of Engineering, the IET and the IMA. He also wrote the standard textbook "Security Engineering-a Guide to Building Dependable Distributed Systems".

Cecilia Mascolo is a Reader in Mobile Systems in the Computer Laboratory, University of Cambridge. She has published extensively in the areas of mobile computing, mobility and social network modelling. She is and has been PI of various national and international projects on social network analysis, sensing and mobility modelling. She has served as a programme committee member of many mobile, sensor systems and networking conferences and workshops. She has taught various courses on mobile and sensor systems and a Masters course on social and technological network analysis.