Temporal Distance Metrics for Social Network Analysis

John Tang
Computer Laboratory
University of Cambridge
jkt27@cam.ac.uk

Mirco Musolesi
Computer Laboratory
University of Cambridge
mm753@cam.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cm542@cam.ac.uk

Vito Latora
Dipartimento di Fisica
University of Catania
latora@ct.infn.it

ABSTRACT

The analysis of social and technological networks has attracted a lot of attention as social networking applications and mobile sensing devices have given us a wealth of real data. Classic studies looked at analysing *static* or *aggregated* networks, i.e., networks that do not change over time or built as the results of aggregation of information over a certain period of time. Given the soaring collections of measurements related to very large, real network traces, researchers are quickly starting to realise that connections are inherently varying over time and exhibit more dimensionality than static analysis can capture.

In this paper we propose new temporal distance metrics to quantify and compare the speed (delay) of information diffusion processes taking into account the evolution of a network from a local and global view. We show how these metrics are able to capture the temporal characteristics of time-varying graphs, such as delay, duration and time order of contacts (interactions), compared to the metrics used in the past on static graphs. As a proof of concept we apply these techniques to two classes of time-varying networks, namely connectivity of mobile devices and e-mail exchanges.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Network Topology; C.2.0 [**General**]: Data communications

General Terms

Measurement, Algorithms, Theory

Keywords

Temporal Graphs, Temporal Metrics, Temporal Efficiency, Social Networks, Complex Networks, Information Diffusion

WOSN'09, August 17, 2009, Barcelona, Spain.

1. INTRODUCTION

The appearance of abundant and fine grained data about social network interactions has sparked numerous investigations into the properties of human interactions [9, 10]. What has become increasingly clear is that the time dimension of these interactions have often been neglected or understated while developing analytical methods for social and complex network analysis.

We argue that static metrics such as path length, clustering coefficient and centrality [15], to name a few, are sufficient where temporal information is not inherent in the network but give a too coarse-grained view in networks where the temporal dynamics is an essential component of the phenomenon under observation such as human interactions over time.

Past research by Kempe et al. proposed a *temporal network* model with time labelled edges where paths need to obey the time order of the appearance of edges [8]. However, this model does not allow for analysis of frequency of contacts between nodes or groups, nor does it handle temporally disconnected nodes i.e., where there is no time respecting transitive path between two nodes over time. Similarly, in [11] Kostakos presented the concept of temporal graphs and an equivalent measure of delivery time between nodes of a temporal graph. However this provides a skewed indication of the global delay of the information diffusion process since it does not take into account pairs of nodes for which a transitive path does not exist. Also the lack of normalisation over nodes or time do not lend for easy comparison between networks. In [10] the authors analyse information dissemination processes focussing on identifying the diffusion of the most recent piece of information about a certain topic in a social network. We instead are interested in measuring the smallest delay path of generic information spreading processes. Spatio-temporal aspects have also been studied for the analysis of delay and data delivery in DTN networks [7, 2]. The Kempe-Kleinberg model has also been adapted for social networks analysis [5, 13, 1], however the focus of these works is on the local characteristics of time-varying networks; global aspects of the information processes in these networks are not captured.

In this paper we present new metrics related to *temporal distance* and evaluate how these are useful to capture properties at a fine granularity with a global and local view. The key measure that we propose is the average *temporal path length* of a network that gives us a global measure of how fast information spreads to all the nodes of the network by means of transitive connections between them. From this

measure, we derive others describing the temporal network efficiency (a static definition of which is contained in [12]) and temporal clustering.

Previous work on small world effects such as the analysis based on short path length and high clustering on static graphs obtained by aggregating all the links over a certain period of time indicated that these networks are good for data diffusion due to a few edges acting as shortcuts, connecting distant nodes together [15]. However, we show that since static graphs treat all links as appearing at the same time, they do not capture key temporal characteristics such as duration of contacts[1], inter-contact time, recurrent contacts and time order of contacts along a path. For this reason, they give us an overestimate of the potential paths connecting pairs of nodes and they cannot provide any information about the delay associated with the information spreading process.

We show that our metrics are able to quantify and compare in a compact way these temporal characteristics for the study of information diffusion processes. As a proof of concept, we apply temporal metrics on conference, campus and e-mail traces and we show that the network of e-mail exchanges exhibits different and slower efficiency characteristics for data diffusion than that of human contacts.

The rest of this paper is organised as follows. In Section 2 we present a formal definition of our model of temporal graphs and temporal metrics. In Section 3 we present preliminary results of the calculation of the metrics on the datasets before concluding with a discussion of the results and future work in Section 4.

2. TEMPORAL METRICS

A *temporal graph* can be represented by means of a sequence of *time windows*, where for each *window* we consider a snapshot of the network state at that time interval. The metrics we developed over this view of the temporal graph retain the time ordering, repeated occurrences of connections between nodes, contact time and deletions of edges.

We now formally introduce the definition of temporal graph $\mathcal{G}_t^w$. Given a real network trace starting at t_{min} and ending at t_{max} we define a contact between nodes i, j at time s with the notation R_{ij}^s. A temporal graph $\mathcal{G}_t^w(t_{min}, t_{max})$ with N nodes consists of a sequence of graphs $G_{t_{min}}$, $G_{t_{min}+w}$, ..., $G_{t_{max}}$, where w is the size of each window in some time unit (i.e., seconds). Then G_t consist of a set of nodes V and a set of edges E, such that $i, j \in V$, if and only if there exists R_{ij}^s with $t \leq s \leq t + w$.[2] We now introduce the temporal distance metric and then the global and local metrics which we have derived from this.

2.1 Temporal Distance

Given two nodes i and j we define a temporal path:

$$p_{ij}^h(t_{min}, t_{max}) \tag{1}$$

to be the set of paths starting from i and finishing at j that pass through the nodes $n_1^t \dots n_i^t$, where $t_{min} \leq t \leq t_{max}$ is the time window that node n is visited and h is the max hops within the same window t. There may be more than one shortest path. Given two nodes i and j we define the *shortest temporal distance*:

$$d_{ij}^h(t_{min}, t_{max}) \tag{2}$$

to be the shortest temporal path length. Starting from time t_{min}, this can be thought of as the number of time windows (or *temporal hops*) it takes for information to spread from a node i to node j. The *horizon* h indicates the maximum number of nodes within each window G_T which information can be exchanged. In the case of *temporally disconnected node pairs* q, p i.e., information from q never reaches p, then we set the temporal distance $d_{pq} = \infty$.

To compute $d_{ij}^h(t_{min}, t_{max})$ we have implemented a depth first search algorithm that gives the distance from a source node i to all other nodes. The algorithm assumes global knowledge of the temporal graph and keeps track of two global lists, D and R, indexed by node identifier. D keeps track of the number of temporal hops to reach a node and R keeps track of nodes that are reached. We initialise the value of every nodes of D to 1 and R to *False*. Starting with the first time window, we check that the source node i has been sighted. If so, we perform a depth first search (DFS) to see if any unreached nodes have a path to a node that was reached in a previous window. The maximum depth of DFS is dictated by the horizon h and if there are more than one path we choose the shortest. If a node j is reachable then we set $R[j] = True$ otherwise we increment the distance $D[j]$. If the source node i is not reachable then we increment all $D[j]$ since we cannot establish a transitively connected path from the source. We then repeat this for the next window.

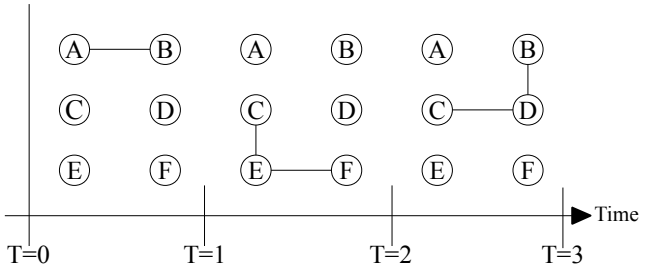


Figure 1: Example Temporal Graph, $\mathcal{G}_t(0,3)$,$h = 2$ and $w = 1$.

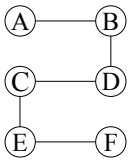


Figure 2: Example static graph based on the temporal graph in Figure 1.

2.2 Example

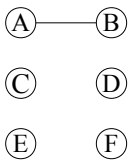
As pointed out in the introduction, we argue that aggregated (i.e., static) graphs are unable to model temporally rich networks since they assume contact between nodes occur all at once. Let us consider the temporal graph in Figure 1 and its static version in Figure 2 where all contacts

[1]Contact in this paper expresses the general concept of a node having some sort of interaction with another node such as physical proximity or exchange of a message.

[2]The limit case is a time window with duration equal to the minimum interval between the appearances of two consecutive contacts. By selecting this window size, there is no approximation in the calculation of the temporal metrics.

are aggregated into a single graph. If node A wanted a piece of information to reach F, according to the static graph it could do so via nodes B, C, D, and E. Also, reversing the path, if node F wanted to reach A it could do so i.e., suggesting paths are symmetric. In fact over time, contacts between A and F occur in the wrong time order to facilitate this. As we can see, the static graph incorrectly showed that information could spread between node A and node F. We now show how our algorithm calculates the true reachability and temporal distance between nodes in the network.

Starting with the first window we calculate the reachability of a message sent from node A. Figure 3 shows the snapshot of contact graph at $t = 1$ and the upper table shows the state of lists R and D after the initialisation phase. We first check if we can see the source node A. Since node A appears in this first window, $R[A]$ is set to *True*. We then iterate through every other node in the window to check for reachability. Since there is a path between A and B and also since A was reached already we update $R[B]$ to *True*. However for node C, there are no contacts to any other nodes so we increment the distance $D[C]$. The same applies to nodes D, E and F and the lower table shows the state of D and R after processing the first window. The second window is shown in Figure 4. We iterate through all unreached nodes C, D, E and F and perform DFS to see if they can be reached via already reached nodes i.e. A or B. As we can see there are contacts amongst the unreached nodes, however none are with A or B so again the distance D for nodes C, D, E and F are incremented. The state of D and R are shown in Figure 4 after processing the second window.



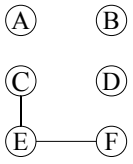
Initialization

	A	B	C	D	E	F
D	1	1	1	1	1	1
R	F	F	F	F	F	F

Variables

	A	B	C	D	E	F
D	1	1	2	2	2	2
R	T	T	F	F	F	F

Figure 3: Distance and Reachability of Window 1.

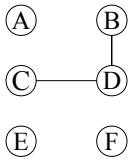


Variables

	A	B	C	D	E	F
D	1	1	3	3	3	3
R	T	T	F	F	F	F

Figure 4: Distance and Reachability of Window 2.

In the third and final window starting from node C, we check if there is a path to a previously reached node. In this case performing DFS gives us two nodes we can reach D and B in the current window, but only node B has been reached in a previous window. We only care that there is a valid path not the number of hops within the current window, so we set $R[C] = True$. Since the value of $D[C]$ is 3 and $R[C]$ is *True*, we now know that a message from node A can reach node C in 3 time windows. Therefore the temporal distance $d_{AC} = 3$. For node D there is a path to node C and node B, but since only node B was reached in a previous window we use this path and set $R[D]$ to $True$. For nodes E and F, a message from node A has still not arrived and so the final state shown in Figure 5 reflects this. For all values of R that are *False* we can treat the distance D as ∞.



Final

	A	B	C	D	E	F
D	1	1	3	3	4	4
R	T	T	T	T	F	F

Figure 5: Distance and Reachability of Window 3.

2.3 Global Temporal Metrics

Global temporal metrics capture the dynamics of the whole network, in particular how easy information flows from source to destination across the whole time space. In the spirit of static global efficiency E_{glob} [12], we define the *temporal efficiency* $E_{T_{ij}}$ between nodes i and j and between the time interval t_{min} to t_{max} as:

$$E^h_{T_{ij}}(t_{min}, t_{max}) = \frac{1}{d^h_{ij}(t_{min}, t_{max})} \quad (3)$$

where temporally disconnected nodes intuitively have $E_{T_{ij}} = 1/\infty = 0$. Therefore, given a horizon h, we can then define the characteristic *shortest temporal path length* L^h and *temporal global efficiency* E^h_{glob} for a temporal graph as:

$$L^h(t_{min}, t_{max}) = \frac{1}{N(N-1)} \sum_{ij} d^h_{ij}(t_{min}, t_{max}) \quad (4)$$

$$E^h_{glob}(t_{min}, t_{max}) = \frac{1}{N(N-1)} \sum_{ij} E^h_{T_{ij}}(t_{min}, t_{max}) \quad (5)$$

To fully characterise a temporal graph, temporally disconnected nodes are captured in the average. In the case of efficiency this is straightforward since temporally disconnnected node pairs have a zero efficiency. In the case of temporal path length we assume that information expires after a certain time period i.e. t_{max}. Therefore, the maximum temporal length that we consider is $(t_{max} - t_{min})$.

2.4 Local Temporal Metrics

Local temporal metrics capture the dynamics of each node and its neighbours across the whole time space. In particular the recurrent interactions between friends or clusters of nodes across time. Also transitivity is an important concept coming from social network analysis [14]. In other words, in a social system there is a strong probability that a friend of your friend is also your friend. As a measure of transitivity in a static graph G, Watts and Strogatz introduced the so-called graph clustering coefficient C defined as the fraction of links that exist between the neighbours k_i of node i, over the total possible number of edges $k_i(k_i - 1)/2$ [15].

The generalisation of the clustering coefficient and the local efficiency E_{loc} for temporal graphs we propose is as follows. We first define $\mathcal{N}_i(t_{min}, t_{max})$ as the set of all first-hop neighbours seen by node i at least once in the time interval

	INFOCOM	REALITY	EMAIL
Start	2005-03-13	2004-07-26	2001-07-29
Duration	4 days	280 days	112 Days
Times	day1:6pm-12pm day2:12am-12pm day3:12am-12pm day4:12am-5pm	12am-12pm	12am-12pm
No. of nodes	41	100	59812
Contacts	avg. 4817	avg. 231	avg. 4000
Granularity	120 secs.	300 secs.	1 sec.

Table 1: Experimental Data Sets.

$[t_{min}, t_{max}]$. We indicate as $k_i(t_{min}, t_{max})$ the number of nodes in the set $\mathcal{N}_i(t_{min}, t_{max})$. We then consider the sequence of subgraphs $G_t^{\mathcal{N}_i(t_{min},t_{max})}$, $t = t_{min}, t_{min+w}, \ldots, t_{max}$ where each $G^{\mathcal{N}_i(t_{min},t_{max})}$ is the neighbour subgraph of node i, considering only the nodes in $\mathcal{N}_i(t_{min}, t_{max})$ and retaining the edges from $G_{t_{min}}$. We define the *clustering coefficient* $C_i(t_{min}, t_{max})$ of node i as:

$$C_i(t_{min}, t_{max}) = \frac{\sum_{t=t_{min}}^{t_{max}} \#\text{ of edges in } G_t^{\mathcal{N}_i(t_{min},t_{max})}}{\tau \cdot \frac{k_i(t_{min},t_{max})[k_i(t_{min},t_{max})-1]}{2}} \quad (6)$$

where the maximum time to live of a message is $\tau = (t_{max} - t_{min})$. Analogously, we can define the local efficiency of node i in the time window $[t_{min}, t_{max}]$ as:

$$E_{loc_i}(t_{min}, t_{max}) = E_T\{G_t^{\mathcal{N}_i(t_{min},t_{max})} \quad t \in [t_{min}, t_{max}]\} \quad (7)$$

that is the efficiency of the time varying graph of the first neighbours of i in the time window $[t_{min}, t_{max}]$, i.e. the shortest-path for time-varying graphs are computed for $G_t^{\mathcal{N}_i(t_{min},t_{max})}$, $t \in [t_{min}, t_{max}]$. Note that by definition, for E_{loc} the horizon is always 1 since we are only considering the direct neighbours of node i.

Finally, the *characteristic temporal clustering coefficient* and the *temporal local efficiency* are defined as follows:

$$C(t_{min}, t_{max}) = 1/N \sum_i C_i(t_{min}, t_{max}) \quad (8)$$

$$E_{loc}(t_{min}, t_{max}) = 1/N \sum_i E_{loc_i}(t_{min}, t_{max}) \quad (9)$$

3. RESULTS AND EVALUATIONS

In our evaluation we use three traces that have been used in previous literature, Bluetooth traces of people at the 2005 INFOCOM conference [6], campus Bluetooth traces of students and staff at MIT [3] and email traces from Kiel University [4]. We shall refer to these as *INFOCOM*, *REALITY* and *EMAIL*, respectively. Table 1 describes the characteristics of each set of traces.

The *INFOCOM* traces were collected in a conference environment using Bluetooth colocation scanning every 2 minutes. With 41 nodes it is quite a small trace but temporally dense in that there are a high number of contacts per day. The *REALITY* traces were collected at the MIT campus between Bluetooth phones sightings of students, research staff and professors, with Bluetooth scanning every 5 minutes. The *EMAIL* traces contain email server logs for 56,969 student at Kiel university. Due to the size we only analyse 7 days of the trace during the Fall semester. To make the experimental results comparable we fix the window size, w to 5 minutes which is equal to the longest Bluetooth scanning rate of the *REALITY* trace. We discuss the effect of different values of window size in Section 3.4.

			Static				Temporal		
Day	N	$\langle k \rangle$	C	L	C_{rand}	L_{rand}	C	L^*	Disc
1	37	25.7	0.818	1.291	0.764	1.336	0.033	4.090	0.28
2	39	28.3	0.845	1.269	0.824	1.263	0.110	4.556	0.13
3	38	22.3	0.744	1.420	0.644	1.405	0.077	4.003	0.19
4	39	21.4	0.722	1.444	0.541	1.474	0.052	4.705	0.14

Table 2: *INFOCOM* Static and Temporal Metrics ($h = max, t_{min} = 12am, t_{max} = 12pm, w = 5min$).

	Temporal Metrics				Reshuffled		
Day	C	E_{loc}	L	E_{glob}	E_{loc}	L	E_{glob}
1	0.033	0.003	19h 39m	0.003	0.077	5h 29m	0.100
2	0.110	0.020	9h 6m	0.024	0.194	2h 45m	0.239
3	0.077	0.013	10h 32m	0.018	0.114	4h 6m	0.167
4	0.052	0.009	9h 55m	0.013	0.104	3h 3m	0.165

Table 3: *INFOCOM Temporal Metrics* ($h = 1, t_{min} = 12am, t_{max} = 12pm, w = 5min$), ($shuffledruns = 50$).

3.1 Comparison with Static Metrics

Firstly, as a comparison between the temporal and the static metrics we show the results calculated for the *INFOCOM* data set. As argued before, paths in static graphs ignore duration of contacts, inter-contact time, recurrent contacts and time ordering of contacts and so overestimate the number of connected node pairs and underestimate the path lengths. Table 2 shows calculations for both static and temporal clustering coefficient, C and path length, L. As a note, since our temporal L metric presented in Equation 4 is in real time, it is hard to compare with static L. To bridge the gap we show temporal L^* which is calculated as the average shortest node to node hop that obeys time ordering of edges. This is fair since temporal L uses the same time ordered path but measured in terms of elapsed time. For the static metric, we also calculate C and L on a random graph with the same average degree, $\langle k \rangle$ and number of nodes N, as prescribed in [15]. As we can see in the static results for Day 1, clustering is high and path length is low. Now looking at the temporal aspects, we have calculated the same metrics but obeying time ordering, duration and recurrence of contacts. The third column, *Disc* shows the ratio of disconnected node pairs. In the case of static graphs, there were no disconnected node pairs. As we can see temporal $L^* \gg$ static L and there are also much more disconnected node pairs due to the observed asymmetry and time ordering of paths. We can see that temporal $C <$ static C, due to the fact that static graphs assume edges always exist across time, when in fact they come and go. In other words, temporal C and temporal L give us a better understanding of the network with respect to the temporal dimension since they can provide us an accurate measure of the delay of the information diffusion process that is not possible with traditional static metrics.

3.2 Temporal Efficiency of Human Contacts

We now calculate temporal L from Equation 4 as a real time along with clustering C and efficiency E. Each data set is measured individually by day, processed by window

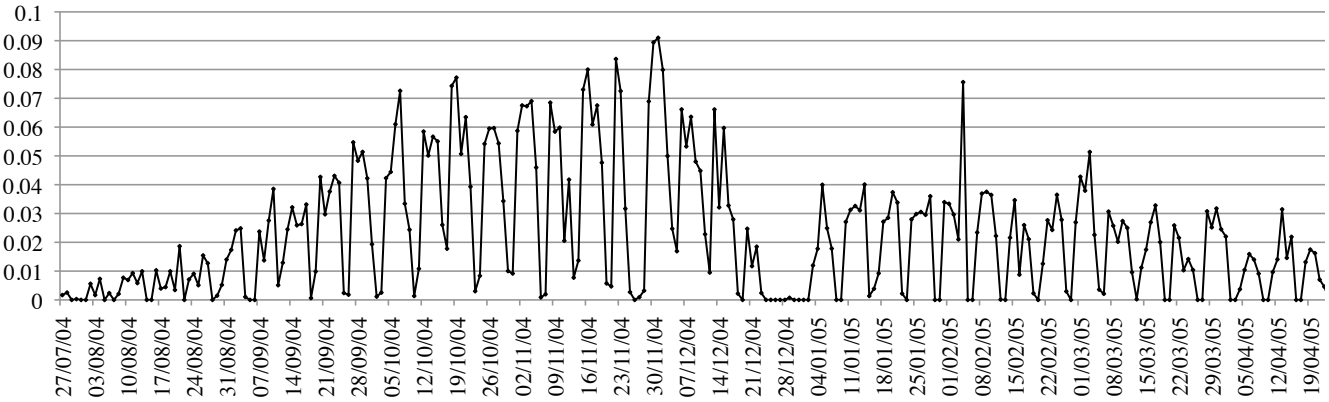


Figure 6: *REALITY* Temporal C ($h = 1, t_{min} = 12am, t_{max} = 12pm, w = 5min$).

Temporal Metrics					Reshuffled		
Date	C	Eloc	L	Eglob	E_{loc}	L	E_{glob}
08 Sep	0.014	0.000	23h 15m	0.000	0.003	21h 58m	0.010
15 Sep	0.060	0.000	22h 47m	0.001	0.007	19h 55m	0.024
22 Sep	0.061	0.000	22h 53m	0.001	0.007	20h 42m	0.019
29 Sep	0.060	0.001	22h 20m	0.001	0.009	17h 44m	0.037
06 Oct	0.026	0.000	22h 14m	0.001	0.011	16h 23m	0.041
13 Oct	0.038	0.000	21h 37m	0.004	0.013	14h 57m	0.055
20 Oct	0.067	0.001	21h 45m	0.003	0.007	17h 4m	0.031
27 Oct	0.050	0.002	22h 1m	0.001	0.013	15h 19m	0.050
03 Nov	0.051	0.001	21h 6m	0.004	0.012	16h 17m	0.043
10 Nov	0.051	0.000	20h 5m	0.004	0.015	14h 25m	0.061

Table 4: *REALITY Temporal Metrics 10 days* ($h = 1, t_{min} = 12am, t_{max} = 12pm, w = 5min$), ($shuffledruns = 50$).

Temporal Metrics					Reshuffled		
Date	C	Eloc	L	Eglob	E_{loc}	L	E_{glob}
27Oct	0	$3.1E^{-8}$	86397.94s	$9.3E^{-7}$	$7.7E^{-8}$	86396.91s	$1.6E^{-6}$
28Oct	$3.5E^{-7}$	$4.0E^{-8}$	86399.78s	$1.4E^{-7}$	$4.1E^{-8}$	86399.71s	$1.5E^{-7}$
29Oct	$2.5E^{-7}$	$3.9E^{-8}$	86399.03s	$3.9E^{-7}$	$7.2E^{-8}$	86398.59s	$7.3E^{-7}$
30Oct	0	$5.8E^{-8}$	86398.76s	$5.5E^{-7}$	$6.9E^{-8}$	86398.48s	$7.5E^{-7}$
31Oct	0	$4.7E^{-8}$	86398.92s	$4.9E^{-7}$	$6.5E^{-8}$	86398.64s	$6.9E^{-7}$
01Nov	0	$5.8E^{-8}$	86399.03s	$4.9E^{-7}$	$6.6E^{-8}$	86398.85s	$6.0E^{-7}$
02Nov	0	$4.3E^{-8}$	86398.68s	$5.4E^{-7}$	$6.5E^{-8}$	86398.67s	$6.8E^{-7}$

Table 5: *EMAIL Temporal Metrics 7 days* ($h = 1, t_{min} = 12am, t_{max} = 12pm, w = 5min$), ($shuffledruns = 50$).

size $w = 5$ minutes. The left hand side of Tables 3, 4 and 5 show the temporal metrics for *INFOCOM*, *REALITY* and *EMAIL*, respectively. The right hand side of the tables will be discussed in the next section.

First looking at *INFOCOM*, recall in Table 2 that static L and L^* only told us the average number of hops in a path but gave us no indication of how long each hop took. Our temporal metrics give us a value that takes account of time and also captures disconnected nodes. From Table 3 we can see L for Day 1, if two people started gossiping at the start of the day, it would take 19 hours to spread to all participants. We also see it is quicker to spread information in the second, third and final day of the conference at about 10 hours. From Table 1 this makes sense since on the first day participants did not start until 6pm (i.e., there is an initial delay equal to 18 hours).

What we see from the low values of E_{glob} and E_{loc} are that contacts between all participants, and contacts between acquaintances did not allow for a high capacity to spread information. Since temporal local efficiency E_{loc} measures how people you meet interact amongst themselves we can drill in and examine on a local view, if the interaction between such acquaintances are any better for spreading information. The fact that there are also low C values for each day reiterates how infrequent groups of people interacted with each other for long periods of time during the conference.

The *REALITY* data set has many more days so gives us a better overview of day to day trends. Plotting temporal C for each day in Figure 6, we can see there are more groups during the middle of each week, with a steady increase in the peaks of the Fall '04 semester (8th Sep to 9th Dec 2004)[3] levelling out during the Spring '05 semester (1st Feb to 12th May 2005). Focusing on these peaks in Table 4 we show 10 consecutive Wednesdays starting from the first day of term. For the first day we can see that it is slow for information to spread since $L = 23$ hours. Also C is 0.014 which tells us that there were groups of people forming infrequently perhaps for lectures or meetings but since both local and global efficiency are at zero they did not interact outside of these meetings. This makes sense since relationships are unlikely to have formed and so there are less contacts. During the subsequent Wednesdays the information spreading process is are quicker and there is also a steady increase in clustering and efficiency. This means that groups are forming more often and outside of these meetings the same people are in contact. However still compared to the conference environment, on a campus it is twice as slow for information to spread.

The final *EMAIL* dataset is the poorest for data diffusion as seen by the zero value clustering and extremely low efficiency and high temporal path length, shown in Table 5. Since there are close to 57,000 nodes we have to take this into consideration when examining these numbers as it contributes to the small normalised values. Classic metrics used on this dataset provide an overestimate of clustering since they assume that all links exist uniformly across time, when in fact in reality, e-mail exchanges take place at specific points in time. What differs from low values seen in

[3]http://web.mit.edu/registrar/www/calendar0405.html

REALITY are that now on some days C is zero and E_{loc} is non zero, albeit extremely small. From the zero C values we can say that email users do not stay in groups or, in other words, do not use email as quick exchanges of messages to each other which makes sense since there are delays between replies.

3.3 Effects of Cyclic Social Behaviour

As a null model, we compare the real data sets $\mathcal{G}_t$ with their randomised counterpart where we have randomly reshuffled the time windows $G_T \in \mathcal{G}_t$, destroying any inherent time order. By definition, temporal clustering coefficient is not affected by the time ordering of windows so we do not show any results for C. The right hand half of Tables 3, 4 and 5 show the metrics calculated on reshuffled temporal graphs for *INFOCOM*, *REALITY* and *EMAIL*, respectively. As we can see in all three traces the shuffled network gives a quicker data diffusion time and higher clustering and efficiency. The reason for this is down to the cyclic behaviour of humans contacts as observed in the previous experiment. Humans as a collective congregate during the working hours and are more sociable during mid week. This means that there is a denser number of contacts at certain times which limits the opportunity for transitive meetings between friends to certain times of the day and decreases the speed of data diffusion. Reshuffling leads to the introduction of heterogeneity of contacts throughout a time period and introduces more opportunity for contacts throughout the day.

3.4 Effects of Varying Window Sizes

We analysed how varying the window size affects the temporal metrics. By considering a larger window size the accuracy of the measurements decreases since by neglecting the order of edge appearances, the temporal path length is under-estimated as it considers links that cannot be exploited in reality. This is coupled with the higher granularity of the measurement units leading to a lower precision in the estimation of the temporal path length (which is over-estimated). However, the latter phenomenon is predominant in the traces taken into consideration, therefore we observe a higher temporal path length as window size increases. On the other hand, since it is inversely proportional to the temporal path length, temporal efficiency decreases as the window size increases.

4. CONCLUSIONS

We have presented a set of new temporal distance based metrics and have shown how they can be applied effectively to characterise the temporal dynamics and data diffusion efficiency of social networks. As a preliminary case study, we have provided comparable, quantitative results using three social network datasets. There are still open topics for future investigation. In this paper we have not shown the effects of increasing the horizon variable, but initial results show intuitively that the temporal path length decreases and global efficiency increase as the reach increases. There are also clear extensions to the temporal path length to capture node importance in the form of a temporal centrality measure, and to see how the maximum diffusion range evolves over time using temporal diameter.

Acknowledgments We would like to acknowledge our colleagues in the Networks and Operating Systems Group, Computer Laboratory. This work was supported through EPSRC grants EP/D077273, EP/C544773 and EP/F013442 and in part by the U.S. Army Research Laboratory and the U.K. MOD under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. MOD or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

5. REFERENCES

[1] A. Clauset and N. Eagle. Persistence and Periodicity in a Dynamic Proximity Network. In *Proc. of DIMACS Workshop on Computational Methods for Dynamic Interaction Network*, 2007.

[2] E. Daly and M. Haahr. Social network analysis for information flow in disconnected Delay-Tolerant MANETs. *IEEE Trans. Mob. Comp.*, 8(5):606–621, 2009.

[3] N. Eagle and A. Pentland. Reality Mining: Sensing Complex Social Systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 2006.

[4] H. Ebel, L. Mielsch, and S. Bornholdt. Scale-free Topology of E-mail Networks. *Phys. Rev. E*, 66(3):035103, 2002.

[5] P. Holme. Network Reachability of Real-world Contact Sequences. *Physical Review E*, 71(4):046119–8, Apr. 2005.

[6] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket Switched Networks and Human Mobility in Conference Environments. In *Proc. of ACM SIGCOMM WDTN '05*, pages 244–251, 2005.

[7] S. Jain, K. Fall, and R. Patra. Routing in a Delay Tolerant Network. In *Proc. of ACM SIGCOMM '04*, pages 145–158, 2004.

[8] D. Kempe, J. Kleinberg, and A. Kumar. Connectivity and Inference Problems for Temporal Networks. *J. Comp. Sys. Sci.*, 64(4):820–842, June 2002.

[9] J. Kleinberg. The Convergence of Social and Technological Networks. *Commun. ACM*, 51(11):66–72, 2008.

[10] G. Kossinets, J. Kleinberg, and D. Watts. The Structure of Information Pathways in a Social Communication Network. In *Proc. of ACM SIGKDD '08*, pages 435–443, 2008.

[11] V. Kostakos. Temporal Graphs. *Physica A*, 388(6):1007–1023, Mar. 2009.

[12] V. Latora and M. Marchiori. Efficient Behavior of Small-World Networks. *Physical Review Letters*, 87(19):198701, Oct. 2001.

[13] A. Mtibaa, A. Chaintreau, J. LeBrun, E. Oliver, A. K. Pietilainen, and C. Diot. Are you Moved by your Social Network Application? In *Proc. of ACM SIGCOMM WOSN '08*, pages 67–72, 2008.

[14] S. Wasserman and K. Faust. *Social Networks Analysis.* Cambridge University Press, 1994.

[15] D. J. Watts and S. H. Strogatz. Collective Dynamics of 'Small-world' Networks. *Nature*, 393(6684):440–2, June 1998.