Periodic patterns in human mobility

Matthew James Williams

A thesis submitted in partial fulfilment of the requirement for the degree of
Doctor of Philosophy
in
Computer Science

Cardiff University

School of Computer Science & Informatics

2013

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed ..... (candidate)

Date .....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed ..... (candidate)

Date .....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

In loving memory of Mum
and with profound gratitude to Dad.

Abstract

The recent rise of services and networks that rely on human mobility has prompted the need for tools that detect our patterns of visits to locations and encounters with other individuals. The widespread popularity of location- and encounter-aware mobile phones has given us a wealth of empirical mobility data and enabled many novel applications that benefit from automated detection of an individual's mobility patterns.

This thesis explores the presence and character of periodic patterns in the visits and encounters of human individuals. Novel tools for extracting and analysing periodic mobility patterns are proposed and evaluated on real-world data. We investigate these patterns in a range of datasets, including visits to public transport stations on a metropolitan scale, university campus WLAN access point transitions, online location-sharing service checkins, and Bluetooth encounters among university students. The methods developed in this thesis are designed for decentralised implementation to enable their real-world deployment.

Analysing an individual's visit and encounter events is a challenging problem since the data are often highly sparse. In order to study visit patterns we propose a novel inter-event interval (IEI) analysis approach, which is inspired by neural coding techniques. The resulting measure, IEI-irregularity, quantifies the weekly periodic patterns of an individual's visits to a location. To detect encounter patterns we propose and compare methods based on IEI analysis and periodic subgraph mining. In particular, we introduce the novel concept of a periodic encounter community; that is, a collection of individuals that share the same periodic encounter pattern. The decentralised algorithms we develop for periodic encounter community detection are of particular relevance to human-based opportunistic communication networks. We explore these communities in terms of their opportunistic content sharing performance.

Our findings show that periodic patterns are a prominent feature of human mobility and that these patterns are algorithmically detectable.

Acknowledgements

First of all, I would like to thank my supervisors Roger Whitaker and Stuart Allen. The journey towards becoming an independent researcher would have been nowhere near as enjoyable or rewarding without Roger and Stuart's enthusiasm and advice. I am especially fortunate to have received Roger's guidance from as early as my undergraduate dissertation and am grateful for the many opportunities he has afforded me since. In addition to my thanks, I should also offer my apologies to Stuart for subjecting him to far too many discussions on the minutiae of nomenclature and notation.

The School of Computer Science & Informatics has been a stimulating and supportive environment throughout my PhD study. I would like to thank Martin Chorley and Walter Colombo for our enjoyable collaborations, Dafydd Evans for insights which have benefited this thesis, Diego Pizzocaro for many conversations about research and for our other collaborations within the School, Jonathan Quinn for our valuable discussions at the start of my PhD study, Nick Fiddian for his encouragement, Christine Mumford for my first research experience, and the mobile and social computing research group for frequent enlightening debates. I should also acknowledge Martin a second time for kindly proof reading this thesis. In addition, I am grateful to Konrad Borowiecki, Lorenzo Moncelsi, Will Webberley, Chris Gwilliams, Rich Coombs, Ian Cooper, Rob Bareš, Phil Smart, Mark Hall, Mark Greenwood, and Alysia Skilton for their advice and friendship, and without whom my time at Cardiff would have not been as entertaining.

I would like thank the European Commission for providing financial support during my PhD study, through FP7 FET projects SOCIALNETS (grant 217141) and RECOGNITION (grant 257756). I am grateful for the opportunity to have contributed to these projects and for the engaging interactions with project partners.

Thanks also to Nathan Eagle at MIT for providing the full Reality Mining dataset; *Transport for London* (TfL) and, in particular, Mark Roberts, Andrew Gaitskell, and Duncan Horne, for their help in providing the London Oyster data; and the Advanced Research Computing at Cardiff (*ARCCA*) division for the use of their HPC cluster.

Finally, greatest thanks go to my family, who have shaped the person I am today and instilled in me the determination and curiosity required to pursue this thesis. To Kathy, Lloyd, Rebecca, Richard, Hetty, Lillian, Lynda, Robert, and Julian I extend my deepest gratitude.

Contents

List of figures

List of tables

List of algorithms

List of acronyms

GPS global positioning system

AFC automated fare collection

WLAN wireless local area network

AP access point

IC-MANET intermittently connected mobile ad hoc network

HEN human encounter network

VANET vehicular ad hoc network

RFID radio-frequency identification

IEI inter-event interval

PEC periodic encounter community

REC regular encounter community

Chapter 1

Introduction

The electronic footprints left by our real-world routines have become wide and varied. The technological systems we interact with and the electronic devices we carry are capable of recording our patterns of behaviour and adapting to our needs. Such systems and devices are increasingly able to detect the context in which they are invoked, enabling them to learn from user behaviour and react to a user's current and future context. Recent hardware advances have made it cheap to manufacture devices with the ability to determine their location and detect other nearby devices and systems. This has led to the proliferation of services and networks that either modify their operation based on the location of the user or whose operation depends on the movement of users. Examples include online social networks (such as Google+, Twitter, and Facebook), intelligent digital personal assistants (such as Google's *Google Now* and Apple's *Siri*), human-based opportunistic networks, and online location-sharing services (such as Foursquare and the now-defunct Gowalla and Brightkite).

The trend towards location awareness has driven interest in two types of mobility context: the places that an individual *visits* and the people with whom an individual comes into proximity (his or her *encounters*). Factors that influence an individual's patterns of mobility can be social, professional, psychological, biological, cultural, and environmental. From the combination of these factors, daily and weekly routines emerge, leading to encounters and visits that periodically recur over time. This thesis is concerned with extracting and analysing these periodic patterns. Our objective is twofold. Firstly, we develop methods for automated detection of periodic encounter and visit patterns that are suitable for application in existing real-world systems. Secondly, we use these methods to explore the character and prevalence of periodicity in human

encounters and visits.

In this chapter we expand on our objective and formally state the contributions made by this thesis, along with the potential implications of this research.

1.1 Human mobility in real-world systems

Opportunistic networks, location-sharing services, location-aware online social networks, and context-aware mobile devices are a selection of real-world services and networks that are capable of recording an individual's visits to locations and his or her encounters with other people. Recent trends have seen these systems attempting to model the temporal patterns that exist within the mobility data they are capturing.

A key factor in the proliferation of these systems is the rapid adoption of mobile phones. The increasing affordability and extensive capabilities of these devices has made them almost ubiquitous in highly developed countries, to the extent that they can now often be regarded as proxies for their owners. It is the now-common ability of mobile phones to geolocate themselves (by using, for example, GPS) that has led to location-awareness being built into online social networks and prompted the rapid growth of location-sharing services.

In the case of human encounters, the short-range device-to-device communication capability of modern mobile phones (through technologies such as Bluetooth and Wi-Fi direct) has made them the most-viable candidate for large-scale deployment of opportunistic networks. An opportunistic network is an extreme case of a mobile communication network where content is communicated exclusively through short-range encounters between the devices in the network. Although other realisations of opportunistic networks exist, human-based opportunistic networks (also referred to as human encounter networks and pocket-switched networks) are the most prolific.

Many advances in opportunistic network protocol design have arisen through the development of methods that detect patterns in devices' encounters and then use these patterns as context for making forwarding decisions. More recently, protocols with basic models of periodicity have emerged, resulting in further improvements to network

performance. This thesis contributes to the field of opportunistic networking through a dedicated study of periodicity in human encounters and the introduction of a number of methods for their automated detection.

Our interest in periodic visiting patterns is primarily related to location-sharing services and location-aware online social networks. These services are predicated on the association between their invocation and the mobility of the user. Many of their features benefit from modelling and mining user mobility patterns. A particularly salient example is that of place and activity recommendation. A recommendation algorithm equipped with the ability to make accurate inferences about a user's visit patterns is able to provide more-personalised recommendations, and even proactively offer suggestions based on the user's current location, previous patterns of behaviour, and the current time. The method and analysis of periodic visit patterns introduced in this thesis adds to the context-aware capabilities of these recommender systems.

The final real-world example that motivates this thesis is context-aware mobile computing. It is estimated that in 2013 over 50% of adults in the United Kingdom owned a smartphone[1]. The two operating systems used by the majority of these devices are Google's *Android* and Apple's *iOS*, both of which have recently added powerful context-aware features to their implementations. Detecting the periodic patterns that exist within records of visits and encounters enables mobile phones to generate even richer context for applications, such as intelligent digital personal assistants. In the future, by a mobile phone being able to infer its user's periodic visits and encounters it can take proactive action such as prefetching information, issuing spatiotemporal notifications, and automatically generating location- and encounter-aware reminders.

1.2 Thesis objectives

We now draw on the motivating scenarios from the previous section to formally state the scope, assumptions, and objective of this thesis as follows:
This thesis explores the presence and character of periodic patterns in the visits and en-

[1] UK Office of Communications (Ofcom) Communications Market Report 2013.

counters of human individuals for use as context in a variety of decentralised context-aware applications by proposing methods that operate on an event stream representation of data.

Event stream data: The most-common representations of visit and encounter data are as a set of time intervals or as a stream of zero-duration events. For example, Foursquare records the checkins of users to venues as events and does not capture any information about length of stay. In this thesis we develop methods specifically for the event stream representation of encounter and visit data. The data in most systems are either inherently event-based, or can be easily translated to an event-stream representation. Although this data can be challenging, building methods specifically to deal with event streams makes them widely applicable and allows us to explore visit and encounter periodicity in a variety of domains.

Decentralisation: Not all systems have consistent access to centralised infrastructure. An extreme example of this is an opportunistic network, which is by definition infrastructureless. Any methods we develop for this scenario must therefore be amenable to a decentralised implementation. More generally, a decentralised approach is also of benefit in scenarios relying on mobile phones to gather individuals' visit and encounter patterns. First, this keeps the control of personal information, and any sensitive patterns that may be revealed, with the user. Users participate in the system by communicating with other peers, rather than via any central authority, and can therefore choose to withhold personal data from users they do not trust. Second, if the methods for identifying visit and encounter patterns are computable at the phone itself, there is no need to wait for access to infrastructure, rely on a cellular network connection, or defer processing to a remote server. In many situations, cellular infrastructure is either unavailable, costly to access, bandwidth-constrained, or overloaded, and so it is necessary to offload communication work on to more-efficient short-range technologies. Other types of challenged network that require decentralised approaches include those in limited-power environments, such as sensor and wildlife monitoring networks, and those in settings that lack infrastructure such as rural, remote, or less economically developed areas.

Context at the individual scale: This thesis seeks to extract and analyse the periodic

patterns of each individual. Many interesting results have been found by analysing aggregated mobility patterns; however, the literature concerning periodic patterns at the individual scale is limited, and this is where our focus lies. To provide individual-specific context the methods we develop must be at this scale.

In summary, this thesis has two overriding themes: the development of decentralised methods to detect periodic patterns and the application of these methods to explore the presence and character of periodicity in human visits and encounters through real-world data.

1.3 Thesis contributions and outline

This thesis contributes novel methods for quantifying and detecting periodic patterns in human mobility. As this is a nascent topic of research, we also find it necessary to introduce new frameworks to define and analyse periodic human mobility patterns, and then apply our methods to study the existence of periodic patterns in the real world. The two features of mobility we study are visits and encounters, and we develop methods to extract periodicity in each of these cases.

In addition to the detection and study of periodic mobility patterns, this thesis also makes a key contribution in providing decentralised implementations of the methods we introduce. The need for decentralised methods is motivated by the potential application scenarios for periodic mobility pattern detection, including opportunistic routing protocols, wildlife monitoring and sensor networks, mobile peer-to-peer content sharing, and pervasive advertising systems. Aside from decentralised scenarios, more-general applications include user modelling, customer profiling, and mobility context for intelligent digital personal assistants.

The remaining chapters of this thesis and their contributions are summarised as follows.

In **Chapter 2** we discuss existing methods used to analyse human mobility and encounter patterns and the insights and applications that these methods have enabled. Particular focus is given to time-aware models of behaviour as these are most relevant to this thesis. To navigate the literature a novel classification system is also introduced.

Using this system we place this thesis in the literature to highlight the contributions it makes.

Chapter 3 discusses the datasets available to explore and evaluate the methods developed in this thesis. This discussion covers both encounter and visit datasets. A variety of data sources, collected via various methods, are open to us, all of which are inherently event-based or can be reduced to an event stream representation. The chapter evaluates the suitability of these various data sources before detailing those selected for use in this thesis. Part of our analysis considers Foursquare checkin patterns, necessitating the collection of a dataset for use in this thesis. This dataset is also detailed in Chapter 3, along with our collection method. The selected datasets are used throughout subsequent chapters.

In **Chapter 4** we tackle the problem of measuring regularity in visit patterns. We begin by highlighting the weekly patterns of visit behaviour at the collective scale. This motivates our subsequent study of regularity at the individual scale, which forms the primary contribution of this chapter. To deal with the event stream data we introduce a novel method of measuring regularity by inter-event interval analysis. This approach is inspired by techniques used in the field of neurophysiology.

In **Chapter 5** we move our focus from visit patterns to encounter patterns. We formulate the problem of periodic encounter community detection, which seeks to extract not only pairwise periodic encounter patterns, but also communities of individuals that share the same periodic encounters. Using a data mining approach and a discrete-time representation we develop a decentralised algorithm that allows individuals to detect the periodic encounter communities they belong to. An important feature of this algorithm is that the periods with which the individuals of a community encounter one another are automatically detected. Our application scenario for this method is opportunistic networking, and we dedicate part of our analysis to the communication dynamics within these communities.

In response to a number of limitations of the periodic encounter community detection algorithm detailed in Chapter 5 we develop an alternative approach in **Chapter 6**. Our findings prior to this chapter show that one-day and seven-day periods are the strongest

periods for human encounter behaviour, and therefore we can develop a single-period method that overcomes the limitations of the strict discrete-time representation used in Chapter 5. This method extends the inter-event interval methods developed in Chapter 4 to allow us to detect weekly encounter communities and is tolerant to uncertainty and noise in encounter times. The algorithm introduced in Chapter 6 re-uses the decentralised approach used for periodic encounter community detection in Chapter 5.

Finally, **Chapter 7** concludes the thesis. In this chapter the key insights and implications of the thesis are summarised and directions for future work are discussed.

1.4 List of publications

The work in this thesis has contributed to the following refereed publications:

[WWA12a] M. J. Williams, R. M. Whitaker, and S. M. Allen. Decentralised detection of periodic encounter communities in opportunistic networks. *Ad Hoc Networks*, 10(8):1544–1556, 2012.

[WWA12b] M. J. Williams, R. M. Whitaker, and S. M. Allen. Measuring individual regularity in human visiting patterns. In *Proc. 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (SOCIALCOM-PASSAT)*, 2012.

[CCW+12] G. B. Colombo, M. J. Chorley, M. J. Williams, S. M. Allen, and R. M. Whitaker. You are where you eat: Foursquare checkins as indicators of human mobility and behaviour. In *Proc. 2012 IEEE Pervasive Computing and Communications (PERCOM) Workshops*, 2012.

We also note that this thesis contributed to the following refereed paper whose workshop proceedings were not published:

[CCW+11] M. J. Chorley, G. B. Colombo, M. J. Williams, S. M. Allen, and R. M. Whitaker. Checking out checking in: observations on foursquare usage patterns. In *Proc. International Workshop on Finding Patterns of Human Behaviors in Network and Mobility Data (NEMO) (ECML-PKDD Workshops)*, 2011.

More specifically, these papers draw from this thesis as follows. Features of the Foursquare dataset we collect in Chapter 3 and use to analyse visit patterns in Chapter 4 are explored in [CCW+12]. The collective-scale analysis of weekly visit patterns in this dataset detailed at the beginning of Chapter 4 appears in [CCW+11]. [WWA12b] also draws on Chapter 4 and, in particular, presents the method and analysis of visit pattern regularity. Finally, [WWA12a] is based on Chapter 5.

Chapter 2

Models, methods, and analyses of human mobility patterns

Introduction

Many different fields of research have found it necessary to model dynamic real-world human mobility. There is a broad spectrum of research, ranging from work investigating universal rules of behaviour [GHB08, SQBB10] to systems that incorporate models of dynamic behaviour [MM09, MMC08]. Often the work is rooted in static (non-dynamic) analysis which has been extended to incorporate dynamic features of the behaviour being studied [TMML09].

This chapter provides a background to the methods used in human mobility and encounter research, along with the insights into human behaviour these methods have enabled. The focus of this thesis is on dynamic approaches and, in particular, periodicity in human behaviour; however, in order to place this area of research in the literature this chapter also briefly discusses static methods. We also note that while this thesis is directed at event-based data representation (as mentioned in Chapter 1), in this chapter we also consider methods that use other representations.

Chapter outline

In Section 2.1 an overview of the related fields of research is presented using a classification scheme based on scale and temporal context. In Section 2.2 we place this thesis with respect to the classification scheme, which serves to highlight the contributions of

this thesis in the body of relevant literature. The various findings and analyses in the relevant fields of mobility pattern research are discussed in Section 2.3. Section 2.4 discusses existing methods and models for extracting periodic patterns in human encounter and mobility behaviour, which is the focus of this thesis. Finally, Section 2.5 concludes this chapter.

2.1 Classification by temporal context and scale

To navigate the many fields of related research we propose a classification scheme of two dimensions: the *scale* and the type of *temporal context*.

2.1.1 Temporal context

Temporal context refers to the extent to which temporal information is encoded by a model. This dimension is divided into AGGREGATE, RECENT, and PERIODIC categories.

An AGGREGATE perspective is where temporal information is not explicitly encoded. These are typically static models of mobility and encounter behaviour where the underlying temporal processes is ignored and data are viewed as a single aggregate. There is still rich information that can be extracted from a static analysis, but inferences regarding temporal structure (i.e., patterns and periodicities) are not possible. Methods that, for example, calculate frequencies or rates aggregated over time, without reference to the ordering or timing of the underlying events, fall into this category.

Very often, the representation of data indicates the amount of temporal context that is available. Figure 2.1 illustrates how the same data can be viewed either temporally or statically. In this example, only AGGREGATE temporal context is available in the static representation, whereas the continuous- and discrete-time representations are amenable to RECENT and PERIODIC models.

A RECENT perspective uses conditions or patterns that have occurred recently as temporal context. Most approaches based on sequence analysis fall into this category. Se-

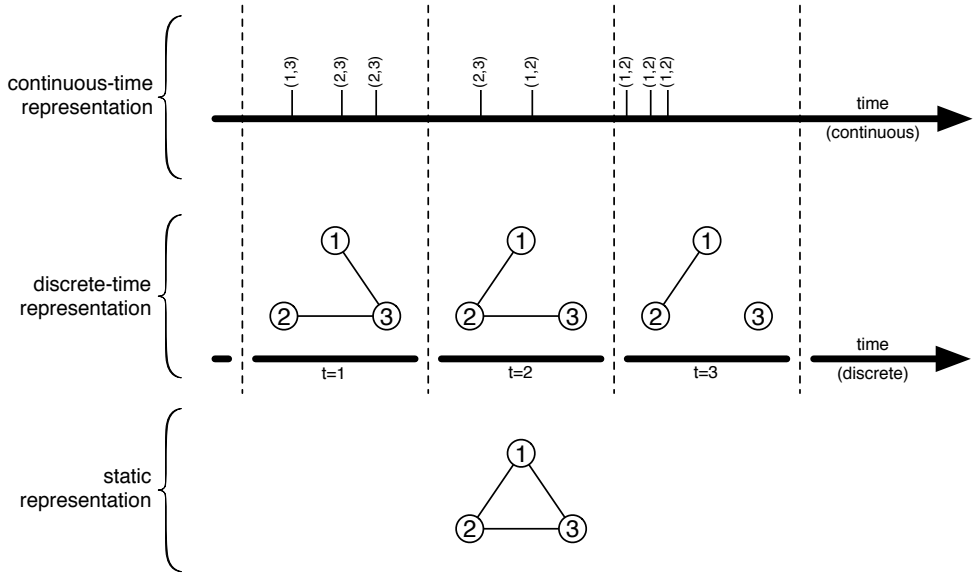


Figure 2.1: Three representations of encounters among a set of individuals. The *continuous-time representation* consists of a time-ordered sequence of timestamps. The label (v, u) indicates an encounter event between individuals v and u. By partitioning time into equally sized bins a *discrete time representation* can be derived from the original continuous-time representation. The *static representation* aggregates all events over the whole duration, producing a representation that does not encode any temporal information.

quence analysis methods consider an immediate sequence of events that have occurred, and their relative ordering, and compare them to other (e.g., historical) sequences of events to make inferences. For example, a frequent sequence approach to location prediction (such as [AS02]) might look for sequences of location visits (e.g., a visit to location l_1 is often followed by a visit to location l_2) that occur frequently with reference to the relative ordering of those visits (but not the specific times they occur) and compare them to the most-recent sequence of visited locations.

PERIODIC temporal context refers to the proposition that conditions (e.g., visit or encounter behaviour) at a particular instance are likely to be similar to conditions at some regular interval in time prior to that instance, and therefore there is repeated behaviour according to one or more periodicities. In less abstract terms, an example is a commuter's travel between 08:00 and 09:00 every Monday morning. The individual's movement is likely to be very similar to that of the same hour and day in the previous week, and the week before that, and so on. In this example we are assuming the

individual has a seven-daily periodic behaviour. A method leveraging periodic temporal context would look through the individual's history at multiples of this seven-day period and use the conditions at these instances to make inferences.

When comparing PERIODIC and RECENT, an important distinction is that RECENT temporal context does not discriminate *when* events occurred in time, and therefore ignores factors such as calendar-driven human behaviour.

2.1.2 Scale

Scale refers to whether the method and/or analysis extracts behaviour at an INDIVIDUAL or COLLECTIVE level. INDIVIDUAL methods act on data for a particular individual; i.e., a person's encounters with others or a person's visits to locations. Methods at this scale are amenable to local computation; e.g., by a user's mobile device. They also reveal user-specific behaviours, and therefore are useful as context for user-centric applications. Existing research in related areas has found it necessary to differentiate between collective scale and individual scale analysis [CGW+08]. Both scales have their respective uses depending on the characteristics (population or individual) one seeks to extract and the choice affects the methods that are necessary and how they can be applied.

2.2 Overview and thesis position

Given the thesis objectives stated in Chapter 1, the methods, models, and analyses in the literature that are most relevant to this thesis are INDIVIDUAL-PERIODIC. Although more recently a number of areas of encounter and mobility research have found utility in incorporating periodic temporal context, the overall amount of research in this area is limited. Figure 2.2 places the fields of mobility research according to our classification scheme. The contributions of this thesis lie in the bottom-right corner of the figure. To date, the most-significant contributions in the INDIVIDUAL-PERIODIC category are in the fields of location prediction and mobile communication networks. We will elaborate on the advances in these fields, along with relevant contributions in a

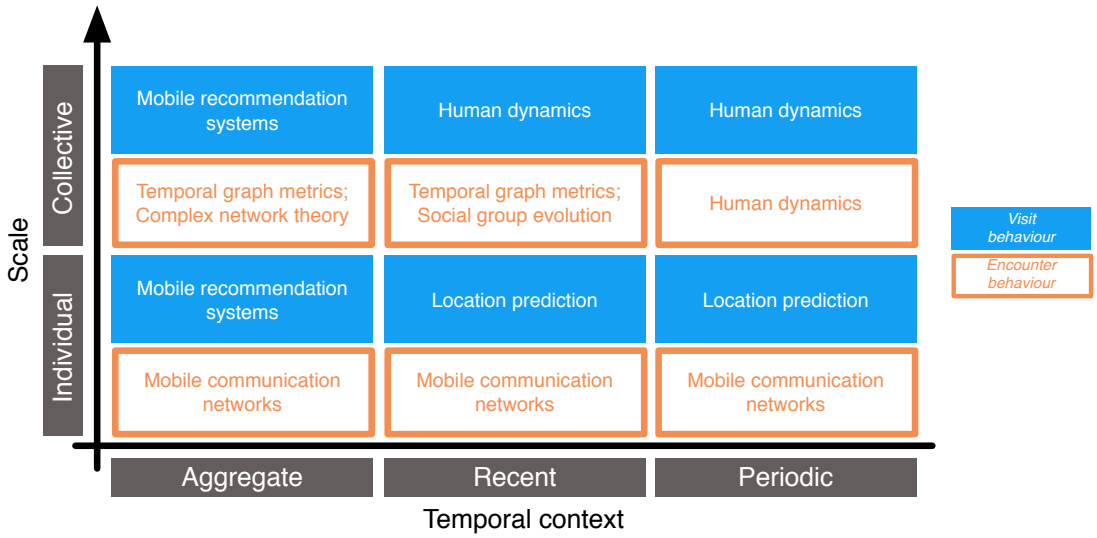


Figure 2.2: Classification of visit and encounter analysis by temporal context and scale.

number of other categories, in the rest of this chapter. We include discussion of a number of approaches in other categories as these may have the potential to be adapted to INDIVIDUAL-PERIODIC or highlight the evolution towards extracting richer temporal context.

2.3 Mobility and encounter patterns

In the following sections we discuss literature relevant to RECENT and PERIODIC temporal context, with particular reference to how periodic patterns in mobility and encounters have been observed and exploited. We also briefly consider AGGREGATE context, but only to provide background for the temporally richer approaches that are most relevant to this thesis. Each of the following sections elaborates on a research area classified in Figure 2.2; namely, mobile communication networks (Section 2.3.1), mobile recommendation systems (Section 2.3.2), community structure in complex networks and social group evolution (Section 2.3.3), temporal graph metrics (Section 2.3.4), human dynamics (Section 2.3.5), and location prediction (Section 2.3.6).

2.3.1 Mobile communication networks

The broad field of communication networks has been concerned with human mobility since the advent of wireless communications. The connectivity of portable devices such as mobile phones, personal digital assistants, and laptops through wireless technologies such as Bluetooth, 802.11, and ZigBee is subject to the device's proximity to its neighbours, and therefore the owner's location and movement.

2.3.1.1 Human encounter networks

Traditionally, computer communication networks and their protocols are designed to enable synchronous connection to a wider-scale network such as the Internet. However, a more-extreme class of networks are human encounter networks (HENs), which are a specific type of opportunistic network [PPC06] where content and messages are shared over occasional short-range connections between devices carried by individuals. A HEN relies on human mobility for data to travel from one location to another. We should note that this type of network is also sometimes referred to in the literature as a human-based opportunistic network or pocket-switched network.

In addition to opportunistic networks formed by mobile devices carried by humans, there are a variety of scenarios in the literature that envisage communication networks that rely on human mobility. These include intermittently connected mobile ad hoc networks (IC-MANETs) [MHM05] and vehicular ad hoc networks (VANETs) [LW07]. The extent to the impact of human mobility depends on the scenario, but in all cases understanding and detecting patterns in movement provide valuable context for routing and content sharing protocols.

2.3.1.2 Context for opportunistic forwarding protocols

Human mobility patterns are responsible for the sporadic contacts between nodes in HENs. This movement is driven by the users' underlying behaviour and social relations. Intuitively, there is structure and routine in human mobility, and thus we expect that encounters will also exhibit structure and routine. This assumption has driven re-

search into opportunistic forwarding protocols that use human behaviour as context in making forwarding decisions.

Pelusi et al. [PPC06] provide a taxonomy for the classification of forwarding techniques in opportunistic networks. The authors define context-based (or *context-aware*) protocols as infrastructureless protocols where nodes asses the usefulness (or *utility*) of proximate nodes as the next hop in communication, based on context information.

The types of context that can be utilised are wide and varied. Examples that have been exploited in the literature include local network metrics such as a node's rate-of-change of connectivity (demonstrated in the CAR protocol [MM09]) and personal data such as name, place of work, and place of residence (demonstrated in the HiBOp protocol [BCP08a]). Context can also extend beyond a node's immediate neighbourhood to include broader network structures that are relevant to it. In protocols such as SimBet [DH07] and Island Hopping [SPG06] nodes collaborate to learn topological features over which they can perform content forwarding. These approaches require nodes to build network knowledge over time, but do not explicitly model the temporal dynamics of encounter behaviour. By introducing methods to detect temporal features such as regularity in movement and encounters this thesis provides network engineers with more tools to extract context at nodes. This context can be used to find further improvements in the performance of HEN dissemination algorithms.

Many of the concepts in human context-aware systems are drawn from the domain of *user modelling*; research in this field attempts to develop machine-computable models for learning and understanding particular aspects of individual human behaviour. One area of user modelling research highly relevant to HENs is *user mobility prediction* [AS02] [SK05]. Human encounters occur due to users co-locating for a period of time, and this in turn is a result of the users' mobility, so the relationship with opportunistic communication protocols that operate in HENs is apparent.

It is clear, therefore, that predicting future locations has potential for application in information diffusion; however, for application in opportunistic networks we would require continuous-time high-precision location data for all individuals. The incompleteness, inaccuracy, and unreliability of existing location-inference techniques (GPS

tracking, cell tower signal strength [LRT04], etc.) mean that it is often difficult to use location context in practice. For this reason many protocols only focus on encounters between nodes, leaving the location-based patterns of nodes as a hidden layer.

2.3.1.3 HEN knowledge constraints

In a HEN, each node has the ability to maintain a history of its contacts for each of its acquaintances. In practice this can be implemented by humans carrying short-range radio devices which log and timestamp the sightings of other devices. Such systems have already been deployed in empirical studies. Examples are the MIT Reality Mining dataset [EP06], which used Bluetooth-enabled smart phones, and the Haggle dataset [SGC+06], which used Bluetooth-enabled iMotes. From the pairwise contact histories available at a node it is possible to extract the periodic patterns of encounters with the node's acquaintances and project these into the future to estimate forthcoming contacts.

For HEN communication protocols (i.e., protocols enabling the routing and sharing of content across a HEN) it is important to distinguish between having complete knowledge of the whole network and only local knowledge. This distinction is referred to in [LPdR06] as the difference between an *omniscient view* of the evolution of the network and an *egocentric view*. Certain systems allow for an omniscient view, the most obvious example being online social networks. Encounter networks, however, are inherently decentralised. Only local contact data is readily available at a node in such networks, therefore any truly deployable periodic pattern detection algorithm should be egocentric.

The egocentric and omniscient views of HENs correspond to the INDIVIDUAL and COLLECTIVE scales we introduced in Section 2.1. Much of the HEN communication protocol literature is concerned with INDIVIDUAL-scale extraction of encounter patterns as it is difficult to achieve global knowledge in an inherently decentralised network, and inferences regarding whole populations are not directly useful to individual nodes of the network. Human encounter networks are one of this thesis's motivating scenarios, and therefore the INDIVIDUAL scale is highly important to the periodic encounter pattern methods we develop in subsequent chapters.

2.3.1.4 Human encounter modelling for HEN communication protocols

Encounter prediction models have been developed and employed in content forwarding protocols. A subset of prediction techniques require the segmentation of time into slices and attempt to predict future quantities within such slices (e.g., the Content Source Selection protocol [MMC08]). For these models the size of slices is an important parameter; for example, an estimate of a dyad's in-contact duration within 10 minute chunks is more useful than estimates for 6 hour chunks. A narrower slice width equates to higher temporal resolution, and ideal models will provide sufficiently accurate estimates at high temporal resolution.

For HEN communication protocols there are a few specific attributes of encounters one may wish to predict. Many models focus on the start and end times of a given dyad's encounters. From these data two quantities readily follow:

- *Contact duration*: The duration of a particular contact.
- *Intercontact time*: The gap between two consecutive contacts.

Prediction can be carried out by forming a *time series* of these quantities (e.g., a time series of intercontact times) and extrapolating according to some prediction model. Two further attributes, each of which respectively relate to those above, can be obtained by slicing time into intervals:

- *In-contact fraction*: The duration within a time slice that the pair were in contact.
- *Out-of-contact fraction*: The duration within a time slice that the pair were out of contact.

These two quantities are less valuable than their non-time-sliced counterparts since they introduce ambiguity into the measure. The measures do not specify exactly how the encounters are distributed within the slice, providing only a generalised abstraction of the original measure.

The Content Source Selection (CSS) algorithm in [MMC08] is a type of *in-contact fraction* predictor. In particular, the CSS algorithm produces estimates of the fraction of in-contact duration per hour as a way to model the time-of-day dependency of human encounter behaviour. The Habit protocol [MMC09] uses a similar model, in

which a regularity weight is computed as the frequency of encounters between two nodes at a given time-of-week. These protocols based on a temporally varying mean fall into the category of INDIVIDUAL-PERIODIC models. By leveraging periodicity in appropriate scenarios these protocols have been able to improve routing and content sharing performance with respect to periodicity-naïve approaches; most notable is the CSS algorithm, which exploited weekly patterns among commuters to deliver content by opportunistic sharing.

More common in the mobile communications literature are *naïve mean* and *refresh-and-decay* models, which we categorise as INDIVIDUAL-RECENT.

Naïve mean models, such as [BCP08a], are simpler versions of temporally varying mean approaches. Instead of modelling time-of-day variation in encounter behaviour, naïve mean models estimate the likelihood of a future meeting by averaging encounter frequency over a finite history.

The refresh-and-decay class of models, such as drop-least-encountered [DFL01] and PRoPHET [LDS04], operate on the principle that a high number of encounters with a node in the immediate past implies high chance of meeting again in the future. An individual maintains a *meeting likelihood* value for each node it has sighted; the algorithms function by updating these values as follows:
When a pair of nodes are co-located they increment their respective meeting likelihoods by some predefined value (this is the *refresh* component of the protocol). Periodically, nodes decrease their meeting likelihoods for non-proximate nodes according to some *decay factor*.

Refresh-and-decay protocols have two main weaknesses. First, the meeting likelihoods have no well-defined meaning; these likelihoods are used as relative quantities for the purpose of ranking nodes to make forwarding decisions. Second, the protocols do not account for the periodic variation of encounters; more-recent encounters are favoured without consideration of more-complex temporal patterns (e.g., diurnal cycles).

The temporally varying mean model is more temporally aware than the naïve mean and refresh-and-decay approaches, but it too has its limitations. The predictor only models one period (i.e., 24-hour), even though other cycles, such as weekly, also exist. Also,

the period is predefined and does not account for individual behaviour; for example, some relationships may have a stronger diurnal component while other relationships may have stronger weekly component.

The state of the art in encounter-aware protocols is the context-aware adaptive routing (CAR) protocol [MM09], which predominantly operates at the INDIVIDUAL scale and incorporates RECENT and PERIODIC context. CAR represents the co-location time series as binary values, each element indicating co-location or absence during a particular time slice. The protocol uses a Kalman filter estimator [BD02], which is a well-established analytical prediction technique in which seasonal and trend components can be modelled. In this way the protocol models both periodic patterns (seasonal) and recent (trend) patterns. Through synchronous multi-hop communication, nodes also share their context information within their partition of the network. CAR is therefore another example of a PERIODIC protocol and partly relies on the assumption that periodic encounter patterns exist. The assumption of detectable periodic encounter behaviour is a hypothesis that this thesis specifically explores.

2.3.2 Mobile recommender systems

The rise of mobile phone usage has enabled the use of location in making personalised location-aware recommendations, allowing recommender systems to move from online systems into the real-world. Location can be used as context for many recommendation applications. Examples include recommending tourism services [Kab10], recommending social events [QLC+10], location-based content provision [BBG12], and location-based advertising [SQC12].

Recommender systems typically act at the individual scale. The system should recommend content that a user wishes to consume, and user-specific context provides a way to personalise recommendations to that user. Collaborative filtering methods introduce other users' behaviours and preferences to the recommendation process [MHN07].

Of particular interest is the use of a user's visit history to recommend venues to visit. A simple method is to consider the user's distance to candidate locations in the ranking process. More-sophisticated approaches consider the set of places a user has vis-

ited, from which other similar venues can be suggested as recommendation candidates [ZCZ+10], or using the social characteristics of users to generate candidates [YYL10]. We categorise these approaches as INDIVIDUAL-AGGREGATE.

In the context of human encounters, recommender systems have been built to automatically infer and suggest friends of an individual based on his/her Bluetooth proximity [QC09]. This work also falls in the INDIVIDUAL-AGGREGATE category as it relies on aggregated encounter characteristics.

2.3.3 Communities in complex networks

Community detection is a well-studied problem in the field of network science and has recently been applied to networks of physical proximity between human individuals. In general, community detection seeks to identify highly clustered components in large real-world networks. Many community detection methods have been proposed, but most are intended for offline analysis of networks (see [For10] for a comprehensive survey of community detection methods). Furthermore, most methods analyse static networks; i.e., where interactions have been aggregated into a single graph regardless of their time and order. The most-relevant community detection algorithms to this thesis are those of Hui et al. [HYCC07]. These algorithms are notable as they offer a decentralised approach for nodes to detect the static encounter communities they belong to over time. However, the algorithm considers aggregated graphs rather than any temporal or periodic trend in the encounter patterns.

In the context of centralised offline analysis, recent work has generalised community detection from a single-graph representation to a multi-slice network representation [MRM+10], where the network is taken as a collection of slices whose nodes are coupled by inter-slice edges. When applied to discrete-time data, this representation is similar to that of a time-varying graph (discussed in more detail in Section 2.3.4). This multi-slice network approach has so far not explored temporal dynamics within human encounter networks. Other recent research into the dynamics of community structure, such as that of Palla et al. [PBV07], has analysed the evolution of social groups over time, and falls into our COLLECTIVE category. There is little prior research into *peri-*

odic community behaviour, however.

The closest example is the Habit [MMC09] communication protocol mentioned in Section 2.3.1.4, which attempts to merge both multi-node encounter behaviour and periodicity. Habit begins with node-centric pairwise analysis of regularity patterns between familiar strangers and, subsequently, nodes exchange their regularity patterns to construct a regularity graph. This is an interesting direction for extending HEN routing into the PERIODIC category. As previously noted in our discussion of other protocols that presuppose periodicity in human encounter patterns, this area of protocol design can benefit from more-detailed empirical analysis of the presence and prevalence of periodic encounters.

2.3.4 Time-varying graphs

A common recent approach to studying the temporal properties of networks is to use a dynamic graph representation, sometimes also referred to as a *time-varying graph* [TMML10, NTM+12, TSM+10, TMML09]. A dynamic graph retains a network's temporal information by segmenting time into equally sized timesteps, with the encounters in a given timestep being collected together to form a single graph. The result is, therefore, a sequence of graphs representing the time-varying nature of the network, as illustrated in Figure 2.1.

We note that a dynamic graph representation introduces discrete timesteps. By discretising time, we lose some information on the ordering of events within a particular timestep. Discretisation is both a limitation, as it reduces temporal precision, and a benefit as it smooths uncertainty and noise in the timing of events. The degree of quantisation is governed by the timestep width, which we refer to as the temporal granularity Q.

Using a dynamic graph representation, traditional INDIVIDUAL and COLLECTIVE static graph methods have been extended to temporal methods. Temporal distance and reachability metrics [TMML10] have found application for quantifying information diffusion among humans. Temporal counterparts to path length, network efficiency, and connected component size [LM01, WS98] are presented and related to the spread of

information within encounter networks (and, less relevant to this thesis, in online social networks). The concept of small-world networks [WS98] has also been applied in the context of dynamic graphs [TSM+10].

We note that these analyses focus on temporal, but not periodic, encounter behaviour. This also applies to [NTM+12], which is nevertheless of interest to this thesis as it considers temporal multi-nodal community behaviour.

A unique example of analysis of encounter periodicity in dynamic encounter graph research is periodic subgraph mining, introduced by Lahiri and Berger-Wolf [LB09, LB08]. Unlike the other dynamic graph methods discussed in this section, periodic subgraph mining does not extend work in the static network science literature, and instead tackles the problem of identifying periodic encounters from a data mining perspective. The authors present *PSE-Miner*, a single-pass algorithm for extracting all periodic subgraphs embedded in a dynamic graph. PSE-Miner automatically extracts the recurrence period, constituent nodes, and constituent edges of each periodic subgraph. Automatic detection of the period is particularly novel, as other methods often require the expected period to be set *a priori* either explicitly or implicitly. As such, this work falls into the PERIODIC category of temporal context. Applying the PSE-Miner approach to a human encounter dataset, the authors find strong periodicities at 24 hours, 48 hours, and seven days.

We note that application of dynamic graph approaches are not limited to encounter networks. The temporal properties of dynamic graphs is also of interest in animal encounter networks (e.g., wild zebra association patterns [LB09]), biological networks (e.g., cortical networks [TSM+10]), and virtual networks (e.g., an online social network [TMML10]). While the behavioural insights of these analyses are not directly relevant to human visits and encounters, the methods they employ to conduct their analysis are of interest.

2.3.5 Human dynamics

The field of human dynamics attempts to model human behaviour as a *complex system* and to understand the collective behaviours that emerge from individual behaviour and

interaction. Through the abundance of large empirical datasets collected from technological systems, such as mobile phone logs [CGW+08], e-mail logs [Bar05], and social network logs, human dynamics is able to investigate hypotheses and make inferences about large populations. Although originally focused on virtual systems (briefly discussed in Section 2.3.5.1), human dynamics has recently expanded into the analysis of human mobility patterns (we discuss encounter dynamics in Section 2.3.5.2 and movement dynamics in Section 2.3.5.3), making this field of research particularly relevant to this thesis.

2.3.5.1 Human dynamics in virtual systems

Initial human dynamics research focused on virtual human behaviour, such as the inter-event times of e-mail communications [Bar05] and activity patterns of users on social networks [GWH07]. [GWH07] is of particular relevance as it highlights periodicities in social network usage. Human routine gives rise to these patterns, and similar patterns manifest in real-world human behaviour, since the users follow the same multi-scale calendar cycles in both cases. In the case of encounter behaviour, one important difference is that an online interaction network will not include interactions between *familiar strangers*, which would exist in encounter networks. A familiar stranger is a person that an individual encounters regularly, but with whom there is no intentional interaction or explicit relationship. These interactions are of note, despite the lack of a conscious relationship between the individuals.

2.3.5.2 Encounter dynamics

The cheapness of equipping users with encounter and location monitoring hardware and software has enabled researchers to directly obtain data about the movements and meetings of individuals. One salient example of this is the nine-month Reality Mining project [EP06] which followed 100 users and generated over 350,000 hours of encounter and visit data (along with a plurality of other behavioural data).

Two local metrics frequently used in the analysis of human and social networks are *clustering coefficient* and *node degree* [AB02]. Typically, these have been applied

to time-independent encounter networks, and in this case we classify them as INDIVIDUAL-AGGREGATE. These networks ignore the time at which an encounter occurred and ignore the patterns of repeat encounters on the same dyad. Static analysis with these two metrics has revealed some interesting properties (e.g., transitivity of friendships and scale-freeness of degree distributions [MGC+07, AB02]) but the omission of time-varying behaviour makes them less relevant to our work.

Chaintreau et al. [CHC+07] present a rigorous analysis of intercontact times (the gaps between consecutive contacts of a particular dyad) in human encounter networks. The authors study the distribution of intercontact times for both the aggregated nodes and individual dyads, showing that both can be characterised with a power law relationship. Their model makes the assumption that the sequence of a dyad's intercontact times is independent identically distributed and that there is no dependency between dyads. Given this, the time-varying nature of the network is ignored, and so this work falls outside of PERIODIC and RECENT. In reality, we would expect dependency in the series of intercontact times of a particular dyad due to changing behaviour throughout the day. For example, an individual is more likely to meet colleagues during work hours, resulting in shorter intercontact times during the day. We would also expect dependence among the individuals a person meets; for example, an individual is more likely to see commuters along with other commuters.

Our focus in this thesis is dynamic analyses. These capture the evolution and time-varying nature of human behaviour and fall in the PERIODIC and RECENT categories.

Clauset and Eagle [CE07] have adapted clustering coefficient and the node degree, two individual-scale metrics, to a dynamic network and show that they vary periodically with diurnal and weekly cycles. The authors also define *nodal adjacency correlation* and *network adjacency correlation* metrics. These local metrics attempt to quantify the similarity of a node's neighbourhood between two discrete timesteps. It is found that the amount of variation between consecutive snapshots is large during weekday days, and smaller on weekends and evenings; meaning that people tend to consecutively stay with the same group more in the evenings and weekends than during weekdays.

Analysis of the variation in encounters is also carried out in [MGC+07]. Instead of

looking at consecutive snapshots, the authors consider differences between encounters at the same time-of-day over different days. The number of encounters per hour for a particular node does not change greatly for the same hour of the day. Again this corroborates the proposition of cyclic encounter behaviour. We should note, however, that only the number of contacts is measured, not whether the neighbour encountered was the same.

Also at the individual scale is the *life entropy* metric presented in [EP06] which measures the strength of the patterns in a user's high-level daily activities and hourly encounter rates. Life entropy quantifies the level of uncertainty (randomness) in an individual's routine, but does not go to the finer scale of evaluating pairwise patterns between individuals. [EPL09] provides insight into periodic encounter behaviour of friends and non-friends. For both types of relationship, time-of-week encounter probability show diurnal cycles, with especially strong periodicities for encounters between friends.

At the COLLECTIVE scale, one property that has been studied is the evolution of the volume of interactions per unit time (equivalent to the number of edges in each snapshot). [MGC+07] examines the average number of encounters per day, showing that there are much fewer encounters on weekends as opposed to weekdays. In [EP06] the same metric is examined, but at a finer granularity using hourly buckets instead of whole days. The power spectra of this time series exposes 24 hour and 7 day periodicities, as one would intuitively expect.

Other COLLECTIVE-scale properties include aggregate network metrics such as average degree, average clustering coefficient, global efficiency, and largest connected component size [WS98]. The temporal and periodic behaviour of these properties have been studied by Scellato et al. [SMML10]. Wavelet decomposition reveals daily and weekly encounter cycles; i.e., multi-scale periodicity of human encounter patterns.

One application of the study of encounter dynamics is in analysing the spread of mobile phone malware; for example, [WGHB09]. This spread is similar to that of content sharing in opportunistic communication discussed in Section 2.3.1.1, and the findings have implications for both. We note, however, that this work studies the spreading

properties over time from an initial infection; it does not consider the influence of temporal properties such as periodicity.

2.3.5.3 Mobility dynamics

Human dynamics is also concerned with quantifying patterns in human mobility. Analysis of mobile phone logs has revealed deep insights into the movements of individuals.

The work of Gonzalez et al. in [GHB08] reveals both spatial and temporal regularity in human movement. Within a population, average travel distance is diverse. By normalising each individual's movement history by his/her characteristic travel distance, the authors find a high degree of similarity in the mobility of individuals across the whole population. More relevant for this thesis is the authors' identification of periodicity in the probability of individuals to return to the locations they visited before; in particular, strong periodicities at 24 hours and 48 hours, and a notable peak at 168 hours (seven days).

Further attention is given to temporal and periodic movement behaviour in [SQBB10], leading to two key contributions. First, the authors investigate the concept of regularity as the probability that an individual is found at his or her most-visited location at a particular time-of-week, noting a clear diurnal cycle and differences in weekday and weekend behaviour. Furthermore, they find that higher regularity is inversely related to the variety of places an individual typically visits during a particular time-of-week; in other words, the times when an individual is likely to be at his/her most-visited location, he/she is less likely visit a number of different locations. This is clear and unsurprising evidence that individuals do not move randomly between their locations, and movement is characterised by regularity.

Second, the authors of [SQBB10] study the entropy of mobility patterns to find that a significant amount of predictive information is encoded in the sequence and ordering of visits. More-recent work [MSRJ12] investigates this further by considering the variation in entropy by time-of-week. This shows that an individual goes through periods of high predictability and low predictability depending on the time of week; in par-

ticular, night-time hours have low uncertainty (high predictability), weekdays are less predictable during working hours and evenings, and weekend afternoons are highly unpredictable.

We should also acknowledge work identifying higher-level location context; i.e., in [EP06, EP09, Pen07] where regular behaviour is identified in the times of visits to home and work locations. In this case information entropy is used to quantify the predictability of mobile phone users' patterns of transition between home and work and used to predict a user's future activities with 79% accuracy. Jiang et al. [JFG12] also investigate mobility at the level of location-based activities (e.g., work, home, school, shopping, etc.). Using a dataset of individuals' self-reported visits, Jiang et al. study the time-of-week and time-of-day variations in the types of locations individuals visit at a collective scale. Furthermore, they perform a cluster analysis to group individuals according to their daily patterns of activity, identifying behaviour archetypes such as students, regular workers, early-bird workers, afternoon workers, the stay-at-home, and morning adventurers.

2.3.6 Location prediction

The ability to predict a user's future visits to locations has many applications, including context-aware content provision, digital assistants, and analysis of virus spreading patterns. Being able to do this accurately and at fine temporal and spatial resolution is a challenging task. In this area of research, one is interested in predicting an individual's future locations, time to arrival, and/or staying time. The focus is commonly on the next location an individual will visit, so-called *next place prediction* [LGA+12]. While the scale of location prediction is INDIVIDUAL, the temporal context used in existing techniques ranges from AGGREGATE to PERIODIC.

2.3.6.1 Data mining for location prediction

Yavas et al. [YKUM05] adapt a frequent sequential pattern mining [AS95, NKM03] approach for prediction of a user's inter-cell movement in a cellular radio system. Similar to the requirement we outline in Section 1.2, the prediction method in [YKUM05]

operates on an event stream of visits. The dwell time (cell staying time) is ignored except in a pre-processing step to filter out brief visits. Prediction is carried out by comparing the user's recent temporal context (their most recent sequence of visits) to previously mined sequences, and therefore we categorise this as an INDIVIDUAL-RECENT approach.

In [YKUM05] sequences are treated equally, regardless of the time-of-day or time-of-week they appear. Jeung et al. improve on this work with the *Hybrid Prediction Algorithm* [JLSZ08], which predicts location by considering not only the previous sequences, but also previous sequences occurring at a given period in the past; for example, comparing the user's current trajectory to the trajectories at the same time on previous days. This work extends an existing periodic pattern mining approach presented in [MCK+04]. By combining recent context and periodic context, the Hybrid Prediction Algorithm is both INDIVIDUAL-RECENT and INDIVIDUAL-PERIODIC. A key contribution of this work is the algorithm's differentiation between *near time* and *distant time* predictions. The algorithm operates differently in these two cases, preferring recent trend data for short-term predictions (short prediction length) and restricting pattern matching to periodic context for long-term predictions (long prediction length). With this approach the algorithm is able to maintain low prediction error as prediction length increases, unlike the non-periodic model used as comparison.

We also note the *Periodica* algorithm [LDH+10] among data mining approaches. Although it is not a prediction algorithm, it presents a data mining approach to extract periodic behaviour in moving objects. The approach can be categorised as INDIVIDUAL-PERIODIC. Periodica is interesting as it first attempts to automatically detect the most-prominent periodicity in an individual's movement patterns. Period detection identifies the strongest periodicity in the individual's visits to each location, with a periodicity being associated with each location. A discrete Fourier transform is used to extract the strongest period; this contrasts with the PSE-Miner's period detection algorithm ([LB09], discussed in Section 2.3.4), which uses a pattern tree to store repeated events and their periods in memory while performing a single pass over each timestep. Two advantages of PSE-Miner over Periodica are that PSE-Miner is able to extract and retain multiple periodicities involving the same objects. Of course, these two algorithms

are intended for different domains: PSE-Miner extracts periodic encounter patterns, Periodica extracts periodic visiting patterns. Furthermore, Periodica outputs a probability distribution describing the likelihood of a particular individual visiting a particular location in a particular time offset according to the identified periodicity; in other words, it generates a time-of-day, time-of-week, or otherwise (depending on the period) probability, similar to the time-varying mean methods discussed in Section 2.3.1.4.

2.3.6.2 Other location prediction approaches

In this section we consider location prediction techniques developed outside the field of data mining.

Markov (or, equivalently, n-gram) methods for location prediction are sequence-based approaches and have been explored by various authors [AS02, GKdPC12, MRM12, SKJH06]. These Markov prediction algorithms are INDIVIDUAL-RECENT. They are probabilistic sequence-based methods, where the user's recent sequence of visits is compared to previous sequences of visits and a distribution of transition probabilities built for subsequently visited locations. In an evaluation by Song et al. [SKJH06], Markov predictors were shown to outperform other RECENT-based techniques for predicting wireless local area network (WLAN) access point (AP) visits; namely, prediction by partial matching (PPM), sampled pattern matching (SPM), LZ-compression.

The NextPlace algorithm [SMM+11] is a key contribution in the field of next place prediction. NextPlace uses a time-delay embedding method from the field of nonlinear time series analysis [KS99]. This involves mapping sequences in an individual's series of visits to a location to an embedding space, and then finding neighbour sequences in this space that are closest to the most-recent sequence. It is notable that the data points used in [SMM+11] are the daily start times of visits to the given location. Therefore some periodic information about the visits is captured by the algorithm, indicating that NextPlace incorporates some elements of PERIODIC temporal context, rather than solely RECENT. In addition, predicting on this data distinguishes NextPlace from many other prediction algorithms, which often use location transition sequences (e.g., loca-

tion A is followed by location B is followed by location C), and allows it to produce predictions about the time of day when a user will make his/her next visit in addition to the location itself.

NextPlace's implementation in [SMM+11] uses sequential uniform time-delay coordinates; in other words, the algorithm builds sequential patterns, similar to other approaches (e.g., frequent sequences mining and probabilistic Markov modelling) we have discussed in this section, but with the key difference that NextPlaces uses a distance metric to find similar context as opposed to direct matching of sequences. The nonlinear time series method is open to other interesting extensions, such as non-sequential and non-uniform time-delay coordinates, which would capture other periodicities (e.g., seven-day recurrence) and nonlinearities in human visiting behaviour on an individual basis. We note that even without such extensions, NextPlace's prediction performance improves upon the state of the art, and by incorporating periodicity is also able to achieve superior precision for long prediction lengths.

Finally, particularly relevant contributions in this field are the Periodic Mobility Model (PMM) and Periodic & Social Mobility Model (PSMM) algorithms presented by Cho et al. [CML11]. Not only does this work explicitly model periodicity in an individual's visiting patterns (categorising it as INDIVIDUAL-PERIODIC), it also models the influence of the visiting patterns of friends on an individual's visits. Furthermore, the approach deals with the same data type as this thesis; that is, zero-duration event streams such as check-ins on online location-sharing services. Modelling temporally periodic movement in an individual's transitions between home and work is shown to outperform prediction accuracy when compared to naïve baselines and even frequent-location models. We note that there is temporal context to be leveraged beyond what is used in [CML11], as there are candidates for periodic behaviour other than the *home* and *work* latent states considered by the authors. Indeed, a variety of other location types (such as gyms, sports venues, and lecture theatres) are likely to exhibit their own periodic visit patterns with certain individuals.

2.4 Methods and models for detecting periodic patterns

In this section we consider the tools and techniques used to study and extract temporal context. The methods and models most relevant to this thesis are INDIVIDUAL-PERIODIC and we therefore direct most of our attention here. We note, however, that the amount of research in this area is limited.

There are two primary motives in the literature for the development of periodic models and methods. The first is to *observe and investigate* periodic behaviour in human visits and encounters. This research seeks to study the existence of periodic behaviour, and make inferences about its prevalence and character independent of any particular application domain. These observational studies provide insight into encounter and mobility behaviour, and are typically at the COLLECTIVE scale. The methods developed in these studies do not necessarily directly lead to methods that can be used in particular application domains. The second motive, on the other hand, is to *extract and exploit* temporal structure, such as periodicity, for use in particular applications. These methods are INDIVIDUAL-PERIODIC in nature.

A common approach to investigate collective periodic behaviour is to view how a measure, such as visit rate, encounter rate, distance from home, or average clustering coefficient, changes over time [CE07, CGW+08]. To further highlight periodic behaviour, these can be aggregated by time-of-day or time-of-week to analyse the daily or weekly profile of the measure [SQBB10, CGW+08]. Work in the literature has highlighted periodic aggregate behaviour at this scale; for example, in [SMML10] a wavelet decomposition of various time series of collective statistics shows strong daily and weekly periodicities in human encounter behaviour.

At the individual scale the data are more sparse and different methods are used. Methods based on sequence analysis are a common INDIVIDUAL-RECENT approach and include Markov models [AS02, GKdPC12, MRM12, SKJH06], frequent sequence mining [YKUM05], and pattern matching (compression-based prediction, prediction by partial matching, and sampled pattern matching) [SKJH06].

The aforementioned approaches do not model PERIODIC temporal context, but a minority of sequence-based methods have been extended to incorporate periodic characterist-

ics. NextPlace [SMM+11] uses a similar-sequence matching method, but incorporates daily behaviour by taking the visits' time-of-day values. The Hybrid Prediction Algorithm [JLSZ08] adapts a sequence mining method by considering the time-of-day that previous sequences occurred.

Signal processing techniques for modelling periodicity are available, as demonstrated by the application of a Kalman filter [Har90] for predicting pairwise encounters in the CAR protocol [MM09, MHM05]. CAR models both trend and seasonal components, and can therefore theoretically capture dependence of encounters on periodic temporal context, although the extent to which the predictor utilises seasonality in practice is not investigated.

A number of methods have been developed to assess predictability of visits [SQBB10, EP06, MSRJ12] and encounters [MM06b], which have utility in both studies and application domains. Life entropy measures the overall uncertainty in an individual's visits to locations and encounters with other Bluetooth devices [EP06]. This is a useful measure to understand an individual's overall behaviour, but does not reveal predictability in the individual's visits to specific locations (which is likely to vary by location) or other individuals (which is likely to vary by user). The work in this thesis seeks to extract and understand individual-at-location and individual-to-individual patterns, rather than overall behaviour. Instantaneous entropy is used in [MSRJ12] to study the variation entropy through the day, but at a COLLECTIVE rather than INDIVIDUAL scale.

Other relevant literature includes methods for automatically detecting periodicities in location and encounter data. Periodica [LDH+10] does so by selecting the strongest frequency in the discrete Fourier transform in the binary sequence of visits of an individual to a location. PSE-Miner [LB09], on the other hand, is able to detect multiple periodicities in a stream of individual-to-individual encounters at the cost of extra computation time.

It is clear from the literature that there is utility in periodic temporal context, and the increase in papers in this area indicates interest in, and the need for, periodicity-aware methods. When we narrow the scope to methods dealing with zero-duration event streams, the periodicity-aware literature is very limited. This is a more-challenging

data type due to sparsity, making it more difficult to apply approaches that require densely sampled continuous values. However, encounter and visit events are often the natural representation of these data.

Point process models (e.g., [CHC+07]) conform to this data type, but assume time-independence. Among the periodicity-aware models of mobility we have considered, NextPlace [SMM+11] is the closest to our scenario. The model deals with daily arrival times, and partially incorporates periodic temporal context. (We note that NextPlace also separately models staying times.) In terms of human encounter research, PSE-Miner [LB08] is capable of handling encounter event data by bucketing events into discrete time slots. It is also particularly relevant because it explicitly models and extracts periodic encounter behaviour.

2.5 Conclusions

We have presented a range of research, motivated by different domains and objectives, related to observing and extracting temporal context in human mobility and encounter patterns. We have discussed the emerging body of work that is finding value in periodic temporal context for a variety of applications. Incorporating periodic context has led to improved visit and encounter prediction algorithms and a deeper understanding of human behaviour. Existing analyses of human visits and encounters find strong evidence of fundamentally dynamic, and often periodic, behaviour on both an individual and collective scale.

We have classified the related work according to scale and temporal context. The methods developed in this thesis intend to extract periodic context concerning an individual's visits to locations and an individual's encounters with other people. Furthermore, these methods should be applicable on an individual basis, without the need for centralised infrastructure. Given these objectives and constraints, we have placed this thesis's contributions in the category of INDIVIDUAL scale and PERIODIC temporal context.

The most-related techniques in INDIVIDUAL-PERIODIC lie in location prediction (e.g.,

NextPlace [SMM+11]) and mobile communication protocols (e.g., CAR [MM09]). Other algorithms and protocols such as HiBOp ([BCP08a]) and Habit ([MMC09]) have also modelled periodicity to some degree, or made the assumption that periodic patterns exist. These models, however, are evaluated at the application level; that is, in terms of performance measures such as prediction accuracy, delivery time, and resource usage.

The performance improvements that existing approaches have gained by modelling periodicity motivate this thesis to focus on empirically investigating the presence of the periodic patterns that these models presuppose. This allows us to gain insights into the characteristics and prevalence of periodic mobility patterns, which can then inform future application-level methods. We have noted that the amount of literature exploring these INDIVIDUAL-PERIODIC patterns is limited. Furthermore, when narrowing our scope to data represented as zero-duration event streams, we have found that there are few techniques designed for this representation.

In the next chapter we will discuss the datasets that have been used in the literature, and those that will be used in this thesis. Visits and encounters are both the result of human mobility and we consider both types of dataset in the next chapter. Since encounters emerge from visits, this thesis first considers regularity in individual visiting patterns before moving on to encounter patterns in subsequent chapters. This allows us to explore the assumption of periodicity in the case of individual's visits to individual locations before exploring periodicity in encounters.

Chapter 3

Datasets of visit and encounter event streams

Introduction

Event streams of visits and encounters emerge from a variety of existing real-world systems and can also be generated synthetically. This representation of data is fundamental to this thesis.

This chapter serves two purposes. First, it provides a background to the empirical and synthetic data employed in the analysis of human visits and encounters. Second, this chapter introduces the four empirical datasets that we explore in this thesis. These datasets will be used to evaluate the methods developed in subsequent chapters. We select a variety of datasets to allow us to apply our methods in different contexts and study the presence of periodicity in a variety of scenarios. We note that not all datasets are relevant to all our analysis, and each may only be applied in a subset of cases.

An introduction to technological systems that collect human visits and encounters was presented in Section 1.1; in particular, we named opportunistic networks, location-sharing services, location-aware online social networks, and context-aware mobile phones as the scenarios of primary interest. The features captured in each of these systems are varied. Some systems record *interval data*; i.e., the start and end time of an individual's visit to a location or the start and end time of an encounter between a pair of individuals. In many cases, however, the application does not require interval data, and therefore does not need to record it, or other constraints (e.g., hardware limitations, battery life, and privacy concerns) make the task of recording interval data too

challenging.

The commonality in the systems mentioned in Section 1.1 is that they either directly record *event streams* or can be readily reduced to one. An event stream encodes the information that *individual* v *was at location* l *at time* t or, in the case of encounters, *individual* v *was near individual* u *at time* t.

Chapter outline

Section 3.1 provides an overview of existing datasets used in the literature and highlights the dataset requirements for use in this thesis. Existing empirical approaches are most-relevant to this thesis, and we discuss how these have been applied in related fields of research. This section also briefly considers synthetic methods for generating visit and encounter data, including mobility models and direct generators, and explains why these methods are not applicable in our work. Section 3.2 presents the datasets to be applied throughout this thesis. The context of each dataset is discussed, and their respective limitations are highlighted. Furthermore, we detail any sanitisation that was carried out to prepare the data for use. Conclusions are presented in Section 3.3.

3.1 An overview of mobility datasets

This section provides a background for the mobility datasets commonly used in the literature. These datasets fall into two broad categories: empirical and synthetic. Synthetic approaches (discussed in Section 3.1.1) use models of human behaviour to generate artificial data for evaluation. Empirical datasets, on the other hand, contain real-world visit and encounter data.

The worldwide popularity of mobile phones and online social networks has given rise to massive databases of individual visit and encounter behaviour, recorded by organisations such as cellular network operators (e.g., used in [PKK+12]), WLAN operators (e.g., [HKA08]), and social network providers (e.g., Foursquare and Gowalla [NSLM12]). Through datasets made available by these organisations, or by crawling their services using public APIs, scientists have access to rich real-world encounter and

visit data.

In addition to existing organisations and services that record mobility data as a by-product, encounter and visit data has also been collected as part of formal experiments investigating human behaviour. These typically involve equipping subjects with a device to monitor their movements and interactions. Mobile phones are desirable candidate devices as modern smartphones can record visit data using GPS and cell tower signal strength, and encounter data by Bluetooth proximity sensing. Furthermore, participants are likely to routinely carry their mobile phone with them, providing measurements on many of their daily activities. The MIT Reality Mining project [EPL09] is an example of a mobile phone dataset. Bespoke devices have also been used for data collection, particularly for recording human encounters. The Haggle datasets contain encounter data collected from participants carrying Bluetooth-enabled iMote devices [CMMD07, HCS+05]. While Bluetooth sensing represents proximity in the order of tens of meters, other work has found it necessary to capture face-to-face encounters, for which radio-frequency identification (RFID) badges have been used [ISB+11].

Regardless of the scenario a dataset represents, the methods presented in this thesis operate on timestamped sequences of individual-at-location visits and individual-to-individual encounters. These are the fundamental data structures we deal with in our work, and we will refer to them as *chronologies*.

Definition 3.1

A **chronology** is an ordered sequence of timestamped events. In the context of visit data, we let the **visit chronology** of an individual v's visits to a particular location l be denoted by the ordered sequence of times $\mathcal{S}_{v,l} = \{t_i \,|\, i = 1, \ldots, L\}$, where L is the number of v's visits to l. In the context of encounter data, we let the **encounter chronology** of an individual v's encounters with another individual u be denoted by the ordered sequence of times $\mathcal{S}_{v,u} = \{t_i \,|\, i = 1, \ldots, L\}$, where L is the number of v's encounters with u.

Dataset characteristics relevant to our work are:

- Geographic region: The extent of the geographic area in which data are collected.
- Sample size: The number of individuals for whom data are recorded.

- Chronology fidelity: The degree to which a particular encounter or visit chronology captures all real-world events. Fidelity can suffer due to unavailability (of an online service or of a mobile device), undersampling, and misreporting.
- Localisation: In the case of mobility data, this refers to the geographic locality of a visit. For example, in a WLAN trace, visits are localised to wireless APs, whose indoor range in ideal conditions is up to 40m [HKA08]. These locations correspond to wireless propagation areas such as a room or small building. For encounter data, localisation refers to the range threshold for which two individuals are recorded as proximate.
- Duration: The uninterrupted duration for which data was collected.

As our focus is on repeating patterns, datasets we use in experiments must be of sufficient duration to allow behaviour at various periodicities to emerge. Chronology fidelity is also particularly important, as missing visits and encounters can preclude the detection of periodicities that may have been present in reality. In this section we provide an overview of the data sources most relevant to this thesis, focusing on these two requirements as well as the secondary criteria.

3.1.1 Synthetic trace generation

A large amount of research, particularly in the field of mobile communications, relies on synthetic generation of human visit and encounter traces, rather than empirical traces. Synthetic methods are necessary in these fields to evaluate algorithms (such as communication protocols) under a variety of conditions where obtaining sufficiently large real-world datasets matching the desired test criteria would be impractical.

Many models of human mobility have be formulated, each incorporating particular characteristics of human movement. Such models can be used to generate visit events and encounter events. To generate visit events one can discretise the movement area into regions and regard entry into a region as a visit event. Encounters can be extracted by detecting node proximity. Features that have been modelled include statistical properties of human mobility and encounters [LHK+09], social and group mobility [MM06a, MM07, BCP08b], activity-based travel [BCP08b, ZBY+12, EKKO08], and

time-variant behaviour [MCK+04, EKKO08]. Direct encounter trace generation methods also exist, such as [CMRM07], where synthetic encounter traces are generated that replicate the statistical properties (e.g., inter-contact time distribution and node degree distribution) from input empirical traces.

Synthetic approaches such as these are out of the scope of this thesis. Our focus is to detect and quantify periodic behaviour in real-world encounters and visits. Existing synthetic encounter and visit generators do not sufficiently model periodicity for use in this thesis. While two aforementioned time-variant mobility models, the Periodic Trajectory Generator (PTG) [MCK+04] and the Working Day Movement Model (WDMM) [EKKO08], go towards modelling periodic behaviour, they are limited in the behaviours they produce. Furthermore, the nature and prevalence of periodicity in mobility is still not fully understood, and therefore there are real patterns that may not manifest in existing artificial models, which the methods developed in this thesis may capture. We wish to be sure that the results of our experiments reflect true human behaviour, rather than artefacts of a synthetic model, and therefore we restrict ourselves to empirical datasets.

3.1.2 Empirical visit data

Visit data used in related work includes GPS tracks, cellular network call logs, WLAN traces, location-sharing network traces, and fare collection in public transport systems.

3.1.2.1 GPS and cellular network data

global positioning system (GPS) logs have been useful in the study of mobility. For example, the GeoLife dataset covers 182 users over a period of three years [ZZXM09]. Despite being more than sufficient in duration, this dataset (and other GPS trajectory logs) is not suitable for our experiments. GPS logs alone do not provide the locations an individual visited, such as the places visited, and so we would need to perform significant-location mining before generating visit traces. This may potentially introduce false positive location visits. Furthermore, GPS is predominantly limited to outdoor scenarios, which precludes identification of different locations within the same

building, such as a restaurant and shop in the same shopping centre.

Anonymised datasets of mobile phone call logs have been made available to mobility researchers, such as the dataset explored in [GHB08]. These datasets are constructed from the locations of cell towers when they route calls or SMS messages for a mobile phone within their cell. The reliance on call and SMS events means the fidelity of these datasets is insufficient for analysing periodic patterns, since users are unlikely to consistently make calls or send SMS messages in each location they visit.

3.1.2.2 Wireless LAN traces

WLAN datasets consist of logs of wireless device registrations to WLAN access points (APs). When individuals carrying WLAN-enabled electronic devices, such as mobile phones and laptops, connect to an AP, the access is logged, thus providing a record of their visit to that location. Logs can be collected either from users' devices (e.g., [MV05]) or from APs (e.g., [HKA08]). When logs of multiple APs over a region are combined we can reconstruct a partial record of individuals' movements over that region. Many of the available datasets have been collected by universities (Dartmouth College [HKA08], UNC [CLP04], and UCSD [MV05]) and correspond to campus-wide geographic areas.

University campuses are a particularly interesting testbed for our experiments as they feature a variety of location types, such as residences, lecture theatres, cafes, and restaurants, and therefore allow us to study how periodicity is affected by the types of activities individuals perform at different times and locations. Since campus WLANs are managed networks and are often dense, these datasets do well to satisfy our two main requirements; i.e., duration and fidelity. We should note, however, that these networks have the limitation of being restricted to a specific campus and to a predominantly student population.

3.1.2.3 Location-sharing service traces

As briefly introduced in Section 1.1, location-sharing services are a class of location-based social networks focused on enabling users to share their location with one an-

other. Examples of services which are exclusively, or partly, location-sharing services include Facebook, Google Latitude, Gowalla, Foursquare, and BrightKite. These services have in common the ability for users to record that they visit a particular location. In the parlance of location-sharing services, a visit is often referred to as a *check in* and locations correspond to *venues*. Most services record checkins as time-stamped visit events and do not attempt to explicitly capture staying durations.

Location-sharing services can capture information on many individuals' daily routines over long periods of time, making them very attractive sources of data for human mobility research. Large checkin datasets have already been collected and explored in the human mobility research community. [SNLM11] presents an analysis of user behaviour in BrightKite, Foursquare, and Gowalla from datasets collected over multiple weeks, resulting in recordings for 54,190, 258,706, and 122,414 users, respectively. Still larger datasets have also been studied, such as [CCLS11], containing 22 million checkins by 220,000 users over five months.

In addition to valuable movement data, some services also provide rich information about venues, which can be obtained by crowdsourcing or obtaining a curated database of places of interest. For example, Foursquare provides a hierarchy of venue categories, varying from broad (e.g., *Food* and *Shops*) to specific (e.g., *College Tennis Court* and *Paella Restaurant*). This allows us to investigate not only the locations of checkins, but also the types of place a user is visiting.

We should note that these services require a user to check in using their mobile device when they visit a location. Visits in location-sharing service datasets are therefore self-reported, making them liable to under-reporting, misreporting, and different reporting behaviours among the population.

3.1.2.4 Automated fare collection in public transport systems

Automated fare collection systems in public transport systems provide convenience to both passengers and transport authorities for the payment of transport fares and have seen widespread adoption in metropolitan areas across the world. These electronic systems require each passenger to carry a smart card that he/she uses to access the

transport network and record his/her journeys. Each card is associated with an account from which fares are deducted according to the journeys of the corresponding passenger.

While some automated fare collection (AFC) systems use magnetic stripe cards and readers, many modern systems now use passive RFID smart cards for contactless interactions, allowing passengers to simply touch their cards against sensors placed on the transport network to register a journey. In most AFC implementations, each passenger is required to touch his/her card at least twice: once at his/her point of entry (‘*touch in*’) to the transport system and again at his/her point of exit (‘*touch out*’). Some implementations also track intermediate transfers, thereby recording yet more detail on passenger travel.

The daily touch-ins and touch-outs of AFC users generate large amounts of data on the movements of passengers, including where each user has been (e.g., at a bus stop, tram stop, rapid-transit station, or train station) and when they visited. Furthermore, since smart cards are intended for long-term individual use, from AFC logs one could extract a trace of an passenger's visits over a long duration (e.g., months and years), representing a rich image of movement over a large metropolitan region. As with location-sharing service checkins, AFC systems do not record staying times, only the event of an individual visiting a location at a particular time.

Existing large-scale AFC implementations include Hong Kong's Octopus card[1], London's Oyster card[2], and New York's metrocard[3]. Common applications of AFC data include urban transport planning, transport system performance monitoring, and ticket price analysis [Jan10, LC11]. In the field of human mobility, these datasets have proved useful in evaluating content sharing protocols and personalised transport systems [LFC10, MMC08]. Regarding periodicity, previous studies of transport systems have suggested that passenger movement is highly regular. For example, in [LFC10] Lathia et al. find two surges in journey activity, one between 06:30 and 09:30, and another between 16:30 and 20:00. The authors also identify a high degree of predict-

[1] `http://www.octopus.com.hk`
[2] `http://oyster.tfl.gov.uk`
[3] `http://www.mta.info/metrocard`

ability in users' trip times and future visits to stations. Given the large amount of commuter activity, we expect many individuals in transport networks to exhibit strong periodic visit behaviour.

AFC data is of course specific to the movements of passengers, and therefore only represents a subset of the population in the region. Furthermore, chronologies can only be obtained for passengers participating in the AFC system, and data on these passengers' visits to locations outside of the transport system are unavailable. However, the geographic scale, long duration, and large sample size makes these AFC datasets a very interesting test case for this thesis.

3.1.3 Empirical encounter data

Encounter traces have been predominantly collected by one of two methods: direct sensing and proximity inference.

Direct encounter sensing makes use of short-range wireless communication technologies to detect proximity between individuals. Direct sensing experiments collect encounter information by equipping individuals with mobile devices that periodically scan their vicinity via one of these wireless technologies to detect other devices nearby. Over time a device constructs a trace of the other devices (carried by other individuals) its owner comes into contact with.

Proximity inference encounter datasets use movement traces of individuals, such as those discussed in Section 3.1.2, to extract encounter events. By comparing the visit traces of two individuals, we can infer when they were in proximity by finding instances where they were at the same location at roughly the same time.

3.1.3.1 Direct encounter sensing

Among the short-range wireless communication technologies used for direct proximity sensing in the literature, Bluetooth has been the most commonly used. The widespread adoption of Bluetooth in modern mobile phones as a device-to-device data transfer mechanism means that many individuals carry a device capable of encounter sensing.

Bluetooth-enabled smart phones were used to collect direct encounters in the 2004-2005 Reality Mining experiment at the Massachusetts Institute of Technology (MIT) [EPL09]. 100 staff and students were involved in this experiment, each equipped with a mobile phone loaded with data logging software. The nine-month duration of this experiment makes it highly relevant to our work. Data from the Reality Mining project has already been used to study aspects of human encounter patterns and periodicity, such as in [CE07], [EPL09], and [SMML10]. Furthermore, as subjects were asked to carry their devices as much as possible, the dataset captures a large fraction of individuals' daily and weekly encounters.

Other Bluetooth direct sensing experiments have been carried out using a variety of devices, including embedded devices (e.g., Intel iMotes) and personal digital assistants. Overviews of existing direct encounter datasets are presented in [CHC+07] and [HYCC09], including iMote experiments (conducted at Cambridge [LLS+06], Infocom 2005 [HCS+05], and Infocom 2006) and personal digital assistants (conducted at Toronto [SCM+06]). However, unlike the Reality Mining experiment, these experiments are less than five days in duration, therefore precluding the study of weekly encounter patterns.

Proximity sensing approaches have the advantage of detecting only true encounters; i.e., only situations where the two devices are in range. Bluetooth range varies depending on the device's transceiver and surrounding environment, but in practice mobile phone scanning typically acquires devices within 10 metres [PD09], equating to true proximity events. Such events can represent opportunities for device-to-device content sharing, malware transfer, and, if proximity is especially close, transmission of a biological contagion.

RFID badges have also been used as an alternative proximity sensing method to Bluetooth [ISB+11, ASC+09, CVdBB+10, VdBCB+10]. This approach requires each participant to wear an active RFID badge on his/her chest which is able to detect other badges in its line-of-sight. These allow specific detection of face-to-face interactions, rather than the general proximity events that Bluetooth uncovers. Although this enables interesting avenues of research in human behaviour, the scope of this thesis is on the broader scenario; i.e., where encounters are localised to between 10 and 20

metres, rather than restricted specifically to face-to-face encounters. In addition, the existing RFID proximity datasets are not suitable for our experiments because they do not follow particular individuals for longer than a few days.

Despite the advantages of Bluetooth proximity detection, we should note that it is prone to missed encounters. The main source of this problem is the Bluetooth scanning interval. Most experiments must use a long interval (2+ minutes) to balance battery usage. The interval in the Reality Mining project was five minutes, long enough for an encounter between two nodes to be omitted if one or both of the nodes are moving. This loss of fidelity is a necessary compromise for the longer duration of these datasets. The iMote experiments used a shorter scan interval (two minutes), but were carried out over five or less days.

3.1.3.2 Inferred encounter traces

Reliable direct encounter sensing for large sample sizes is challenging due to the power demand on devices and the need for users to install monitoring software. The abundance of large-scale mobility datasets (some of which were discussed in Section 3.1.2) offer an alternative means of generating encounter data. By inferring when two individuals are in proximity from their visit logs we can extract a chronology for that pair's encounters, at the expense of potentially introducing false-positive encounter events. Inference errors can occur due to two individuals moving quickly through the location, and therefore never being proximate at the same instant, or due to a granular localisation that allows two individuals to be designated as at the same location while being separated by a significant distance.

The advantage of inferred datasets is their size. The largest sample size among direct encounter datasets (Section 3.1.3.1) was 100 individuals. On the other hand, the WLAN visit datasets range from 200 to 7,000 individuals.

An overview of encounter datasets, including both direct and inferred methods, is presented by Chaintreau et al. in [CHC+07]. The Reality Mining Bluetooth dataset collected 54,667 internal encounters (i.e., encounters among individuals participating in the experiment) among the 100 participants over 246 days, compared to the 4,058,284

internal encounters over 114 days generated by inferring encounters in the Dartmouth College WLAN dataset. Analyses have shown that the inter-contact time distributions of these datasets are well approximated by a truncated power law, regardless of whether the encounters are by Bluetooth sightings or inferred from WLAN AP logs [FWYC10]. Further investigation comparing the Reality Mining Bluetooth and Dartmouth College WLAN datasets in terms of the properties of encounter chronologies finds they have similar heterogeneity among their pairwise inter-contact distributions. These results indicate that both collection methods produce similar pictures of human encounter behaviour, at least in terms of their static properties, making them both useful means of studying human encounter patterns.

3.2 Visit and encounter datasets used in this thesis

In this thesis we explore the following four datasets:

- DARTMOUTH: Dartmouth College WLAN AP accesses.
- FOURSQUARE: checkins to venues on Foursquare.
- REALITY: Bluetooth encounters in the MIT Reality Mining project.
- UNDERGROUND: journeys by passengers on the London Underground recorded by the Oyster card AFC system.

These datasets and their derivatives will be used in this thesis's experiments. Of the four datasets, the FOURSQUARE visit was collected specifically for these experiments. The other three are existing datasets that were obtained from online data repositories or directly from the data collector; in particular, REALITY was obtained from Nathan Eagle at MIT (now available online[1]), UNDERGROUND was obtained from *Transport for London* (TfL), and DARTMOUTH was obtained from the CRAWDAD online repository[2].

In this section these datasets are introduced, including the circumstances of their collection, preprocessing that we carry out, and their limitations. We note that not every

[1] `http://realitymining.com`
[2] `http://crawdad.cs.dartmouth.edu/`

dataset will be applied in each experiment, and that some of these datasets include both visit and encounter traces.

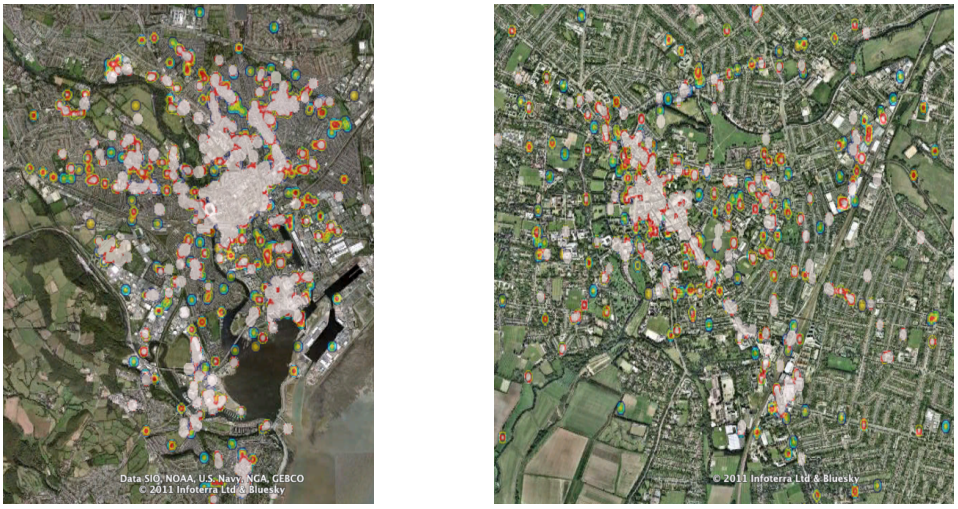
3.2.1 FOURSQUARE: visits to Foursquare venues

Many existing Foursquare datasets were collected by monitoring Twitter for users publicly tweeting their Foursquare checkins [NSLM12]. This indirect method relies on a Foursquare user having associated their Foursquare account with their Twitter account and then choosing to have their checkins tweeted (this can be done on a per-checkin basis or automatically). This approach has the advantage of collecting many checkins at large scales (e.g., country-wide and larger), but does not capture all of an individual's checkins or achieve comprehensive coverage of a particular geographic region. For the purpose of this thesis a dataset of complete visit chronologies is needed, and therefore we collected our own dataset consisting of all checkins in a number of small regions.

We collected our data over 54 consecutive days in 2011. At this time Foursquare allowed users to opt-in to having the venue they are currently visiting be displayed on that venue's *Here Now* list. Real-time monitoring of the Here Now lists for all venues in a particular region allowed us to construct a complete snapshot of participating individuals' movements in that region. Using the Foursquare API we collected a list of all venues in a given region, and then implemented a real-time crawler to query each venue's Here Now list every five minutes, recording users' visits in a database. We also collected each venue's category and geographic coordinates for use in our analyses.

The geographic regions we selected were three cities in the UK; namely, Bristol, Cardiff, and Cambridge. Data was collected in 2011, from 13th May to 6th July. Heatmaps showing the distribution of the checkins observed in the monitoring period over two of the three cities are given in Figure 3.1. As can be seen, the heavy concentrations of checkins tend to match up to features such as roads and dense areas. Checkins are concentrated within the centre of both cities as may be expected.

A summary of the data collected over these three cities is presented in Table 3.1. The combined dataset consists of 810 checkins per day. This equates to an average of 7.7 checkins per user over the 54 days; however, we note that the distribution of users and



(a) Cardiff. (b) Cambridge.

Figure 3.1: Heatmaps showing the geographic distribution and intensity of checkins in two cities.

	Bristol	Cardiff	Cambridge	Combined
Urban population	587,400	292,150	122,700	1,002,250
Collection area (km^2)	1,066	1,395	1,584	4,045
Active users	2,683	2,104	1,376	5,654
Venues	5,161	5,637	3,153	13,951
Active venues	1,774	2,118	882	4,774
Checkins	15,904	19,841	8,022	43,767

Table 3.1: Summary of the base Foursquare dataset. Checkins in Bristol, Cardiff, and Cambridge over 54 days. Population estimates taken from 2011 UK census data. Active users and venues are those with at least one checkin during the collection period.

checkins is not uniform. A small proportion of users are responsible for a large amount of the checkins in this dataset. The overall number of chronologies (i.e., user-venue pairs with at least one visit) is 16,155.

3.2.2 UNDERGROUND: visits of London Underground passengers

The London Underground is a metropolitan rapid-transit rail system serving most of Greater London. An average four-week period in 2010 consisted of 84 million journeys across the Underground's 270 different stations[1]. The *Oyster* AFC system is used

[1]UK Government, Department for Transport. London Underground statistics annual reports. `http://www.dft.gov.uk`. Table reference: LRT9901.

by many passengers, requiring each user to touch his/her personal Oyster RFID card at the station of entry and station of exit. Each touch represents a visit by the Oyster user to a particular Underground station, thus providing a partial record of the user's movement in the transit system. Although not all passengers use Oyster card fare payment, it is estimated that roughly five million cards are used per month[1]. We obtained an anonymised dataset of all Oyster card journeys over 28 days in March 2010 from *Transport for London* (TfL), the government body responsible for the service, to use in this thesis.

The dataset is provided as passenger journeys, consisting of an entry and exit station for each journey. As a basic sanitisation step, we filtered out journeys missing an entry or exit location (e.g., due to a passenger neglecting to touch in or touch out). To obtain visit chronologies each Oyster card touch is treated as a visit to a station, regardless of whether it is a touch in or touch out, producing visit-event three-tuples in the form: `(station_id, passenger_id, timestamp)`. A visit chronology $\mathcal{S}_{v,l}$ for a particular passenger v's visits to a given station l is then obtained by collecting and ordering all visit events corresponding to v and l.

The raw dataset consists of 5,177,134 individuals, each having visited one of the 270 stations at least once. In the four weeks, there were 159,111,519 visits to stations, distributed among 35,338,486 chronologies, giving a mean number of visits per chronology of 4.5.

3.2.3 REALITY: Bluetooth encounters in the MIT Reality Mining project

The 2004-2005 Reality Mining project carried out at the Massachusetts Institute of Technology (MIT) followed 100 subjects equipped with Bluetooth-enabled mobile phones and recorded information about their behaviour over a nine-month academic period [EPL09]. These subjects were staff and students at MIT. 68% of the subjects were postgraduates and staff working in the same building and the remaining subjects are students beginning the same degree.

[1] Transport for London FOI request 0291 1011.

Among the data collected are Bluetooth sightings between subjects, with Bluetooth scanning carried out at five-minute intervals. The dataset also includes sightings with devices outside of the experiment. We ignore these encounter chronologies as we cannot guarantee whether the external device is a device that is reliably carried by another individual.

531,703 encounter events were recorded during the dataset's nine-month duration, giving an average of 19.7 encounters per subject per day. These encounters are distributed over 2,675 chronologies.

3.2.4 DARTMOUTH: visits and encounters on Dartmouth College campus

Visits in the DARTMOUTH dataset are drawn from the use of wireless access points (APs) by staff and students at Dartmouth College campus in the United States [HKA08]. Over 450 APs placed across the 800km^2 of campus provide wireless coverage for most of the area, serving roughly 5,000 undergraduates and 1,200 faculty. When staff and students with wireless-enabled electronic devices (such as laptops and mobile phones) access the campus network the AP used to do so is logged at a central server, thus providing a partial record of the users' movements across the campus. The dataset providers estimate that at least 75% of undergraduates owned portable laptops in the collection period. The campus includes a variety of facilities, including residences, auditoriums, and social spaces. The type of building in which the AP is located is also included in the dataset.

From the four years of wireless traces available in the Dartmouth movement dataset we selected visits from a more-recent year (2003) for our experiments, as recent years are likely to feature more mobile devices (such as wireless-enabled smartphones or personal digital assistants), and thus provide a richer record of user mobility. Such devices did not become very common until recently, however.

To prepare this dataset for our experiments we carried out a number of sanitisation and filtering steps. In particular, we found many cases where a user repeatedly visited with the same AP at a short interval (typically less than 15 minutes). These are artefacts of

the WLAN AP protocol and are caused by the same device periodically re-associating with the same AP. When these re-associations are separated by less than 15 minutes we assume that the user has not moved and therefore we discard the repeat events. In addition to this, we also acknowledge that some devices may be stationary throughout their stay on campus. Although many devices are carried with the students and staff (such as laptops or smartphones), some individuals may have unportable devices (such as desktop computers) accessing the campus WLAN. Since our aim is to study human mobility, we wish to focus on the devices often carried by the user. To mitigate the effect of stationary devices we only included devices that visited at least five different APs. The resulting 365-day dataset consists of 4,724,255 visits, 7,187 users, 567 APs, and 275,843 visit chronologies.

This dataset also forms the basis for generating a dataset of encounters using the inference approach discussed in Section 3.1.3.2. To generate an encounter event we identify occurrences of pairs of individuals visiting the same AP within 10 minutes of each other. An occurrence of a pair of individuals visiting the same location within 10 minutes of each other translates to one encounter event between those two individuals, with the time of the encounter taken as the midpoint between the two individuals' respective visit times.

The resulting encounter dataset contains 6,800,755 encounters, 7,173 active individuals (down from the 7,187 devices in the visit dataset due to 14 devices with no encounters), and 897,996 chronologies. A user in this dataset is on average involved in 2.6 encounters per day.

3.2.5 Summary of datasets

A summary of the base datasets is given in Table 3.2 and Table 3.3. These datasets and their derivatives will be used in the experiments in this thesis.

As depicted in Figure 3.2, a small number of very-active individuals in each dataset are responsible for a large proportion of visits. The distribution for DARTMOUTH also suggests a power law, but has a slight bulge rather than strictly following a straight line. UNDERGROUND exhibits significantly different behaviour. We observe three trans-

	FOURSQUARE	DARTMOUTH	UNDERGROUND
Area(s)	Bristol, Cardiff, & Camb.	Dartmouth	London
Scale	Urban	Campus	Metropolitan
Duration	54 days	365 days	28 days
Location type	Venue	Access point	Metro station
Visit type	Checkin	Association	Card swipe
Individuals	5,654	7,187	5,177,134
Locations	4,774	567	270
Visits	43,767	4,724,255	159,111,519
Visits per day	811	12,943	5,682,554
Visits per individual per day	0.14	1.80	1.10
M	16,155	275,843	35,338,486
$\langle L \rangle$	2.7	17.0	4.5

Table 3.2: Summary of base visit datasets. M denotes the number of chronologies and $\langle L \rangle$ denotes the mean number of visits per chronology. A chronology $\mathcal{S}_{v,l}$ is only included in a dataset if v visited l at least once in the duration of the dataset. Only active individuals and locations are counted; that is, locations and individuals are only counted if they were involved in at least one chronology.

	DARTMOUTH	REALITY
Area(s)	Dartmouth	MIT
Duration	365 days	270 days
Encounter type	Access point co-location	Bluetooth proximity
Encounter range	$\leq$ 40m	$\leq$ 10m
Individuals	7,173	100
Encounters	6,800,755	531,703
Encounters per day	18,632.2	1,969.3
Encounters per individual per day	2.6	19.7
M	897,996	2,675
$\langle L \rangle$	7.6	198.8

Table 3.3: Summary of base encounter datasets. M denotes the number of chronologies and $\langle L \rangle$ denotes the mean number of encounters per chronology. A chronology $\mathcal{S}_{v,u}$ is only included in a dataset if v encountered u at least once in the duration of the dataset. Only active individuals are counted; that is, an individual is only counted if he/she was involved in at least one chronology.

itions in behaviour, the first at 100 (3.6 visits per day), the second at 355 (9.1 visits per day), and the third at 1,250 (44.6 visits per day). This may be explained by the types of individual that use the Oyster card for payment. Some tourists choose to use the Oyster card over other travel cards, and these are likely to be responsible for many of the low-frequency visits. Commuters contribute to the middle of the distribution. Finally, the most-active users are likely to be London Underground stewards and individuals who rely on metro transport as part of their job, such as for messengering and newspaper delivery. An important factor is the financial incentive for using the Oyster card. Oyster cards can be used in conjunction with weekly, monthly, and annual travelcards, and individuals' travel frequency will influence their decision on whether to opt-in to using an Oyster card to receive these discounts. (A deeper analysis of Oyster card usage and fare discounts can be found in [LC11].)

We estimate an upper-bound for reasonable Underground travel frequency to be roughly 60 visits per day, accounting for very-active individuals that need to travel in and out of many Underground stations as part of their job. Only a very small proportion of individuals (0.003%) are more active than this. These are likely to be Underground stewards, as these individuals have access to privileged Oyster cards to allow them to manage the transport network. We filter these users out in subsequent analysis as they are not representative of typical travel behaviour.

Finally, we can see from Table 3.2 that measuring visits through WLAN AP associations results in more-frequent visits (1.80 per individual per day). This is due to the fine-grained localisation of AP data. Small movements of an individual (such as to another room in a building) can result in the individual's device associating with a different AP, and therefore a new visit being recorded. We can also observe the effect of filtering out devices that have not visited at least five locations by the absence of individuals with fewer than five visits in Figure 3.2. In terms of encounter behaviour, we note that REALITY exceeded the number of encounters per user per day than even DARTMOUTH (Table 3.3).

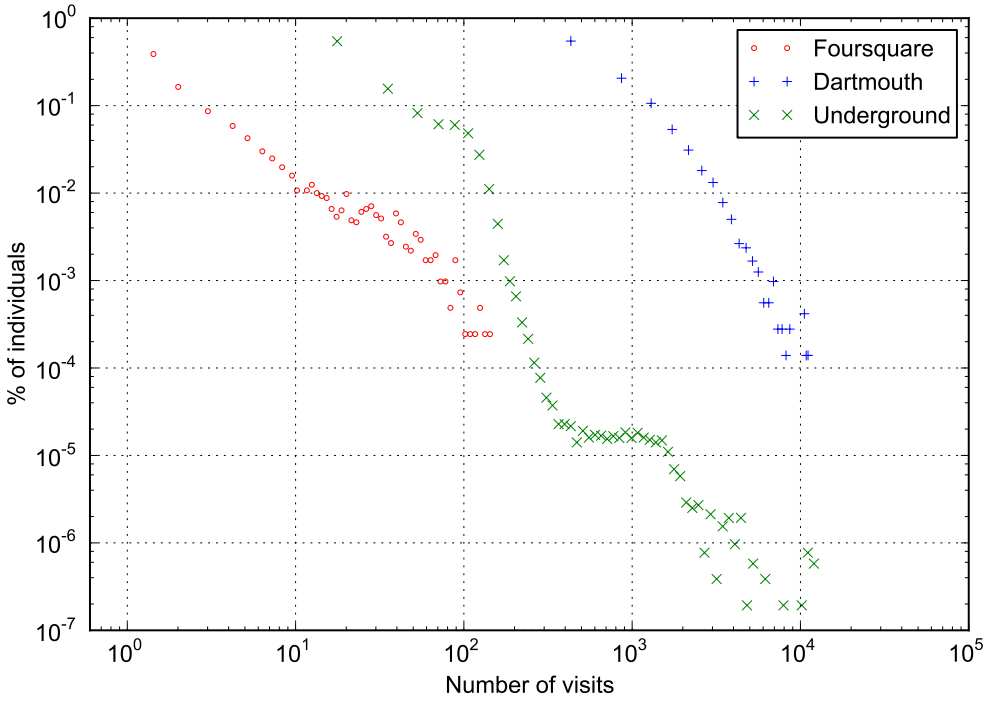


Figure 3.2: Distribution of the number of visits by an individual in each dataset on a log-log scale.

3.3 Conclusions

We have discussed the datasets and experimental methods used to investigate human mobility. The sources of data for human visits and encounters are wide-ranging and have been collected in many different contexts using a variety of methods. We have selected four datasets for study in the rest of this thesis. These datasets represent a range of scenarios on various geographic scales, allowing us to generalise our findings and test our methods in different conditions. In particular, we consider students and staff on university campuses (REALITY and DARTMOUTH), passengers travelling between stations on a metropolitan-scale transport system (UNDERGROUND), and users of a location-sharing service in urban areas (FOURSQUARE). The latter dataset was collected for use in this thesis. In all cases the data are inherently event based or can be reduced to an event representation. The methods in the following chapters are developed for this type of event visit and encounter data.

Chapter 4

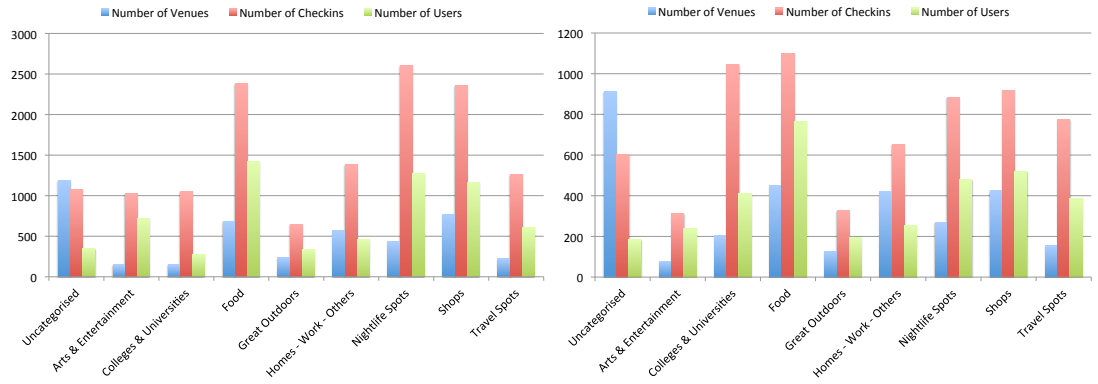
Regularity in human visiting patterns

Introduction

Identifying regular behaviour in individuals' visiting patterns has applications in a wide range of fields including context for digital assistants (such as Google's *Google Now* and Apple's *Siri*), customer profiling for shop owners, urban planning [CSHS12], and understanding the spread of biological viruses and electronic malware [WGHB09].

We note that in referring to visit regularity we do not mean visit frequency. Our goal is to identify consistency between time (e.g., the time of week) and the places an individual visits. In other words, we refer to regularity as the behaviour of an individual to consistently visit a particular location at similar times each day or week. Frequent visits to a location are not necessarily regular as the timing of those visits may be inconsistent. This notion of regularity is also held by individuals' own views on their regularity. This is discussed in [LC11], where it is found that transport network users tend to think of travel regularity as being related to destinations and time of travel, rather than amount of travel.

Many existing studies of periodic and regular behaviour have focused on collective-scale dynamics. However, for the applications mentioned above, regularity context is most valuable when it describes an individual's behaviour rather than aggregate behaviour of a population. Furthermore, not all places an individual visits are necessarily regular. There is likely a great deal of diversity in the visiting patterns between individuals and locations. Factors such as wealth, profession, lifestyle, and health affect an individual's routine, and therefore his or her mobility patterns. In fields outside human mobility, diversity has been found to be fundamental to human behaviour, both within



(a) Cardiff. (b) Cambridge.

Figure 4.1: Checkin statistics for venues in Cardiff and Cambridge grouped by top-level category. In particular, each plot shows the number of active venues in each category, the number of checkins in the venues in each category, and the number of unique visitors to venues in each category.

the same population and among different populations, even having an evolutionary component [BDSL11]. Diversity in visiting regularity may also exist among locations, with some places, such as workplaces, having a natural predisposition for routine.

In this chapter we carry out two tasks. We develop a method that quantifies the amount of regularity in an individual's visits to a location that is applicable to event stream data. This method is then used to investigate the prevalence and character of regularity in human visit patterns. To introduce this we also discuss collective-scale behaviour before moving to the individual scale.

Chapter outline

We briefly consider collective-scale visiting routine in Section 4.1 before focusing on individual-scale visit patterns in the rest of this chapter. In Section 4.2 we introduce IEI-irregularity (inter-event interval irregularity) which is then applied to empirical visit chronologies in Section 4.3. A discussion of the findings and their relationship to related work is given in Section 4.4. The chapter is concluded in Section 4.5.

The work in Section 4.1 has contributed to [CCW+11][1] and [CCW+12]. Work in Section 4.2 and Section 4.3 has appeared in [WWA12b].

[1] [CCW+11] is a refereed workshop paper whose proceedings were not published.

4.1 Collective visiting behaviour: a case study

In this section we investigate collective-scale visiting patterns in the FOURSQUARE dataset (Section 3.2.5), showing how routine behaviour appears when many individuals' visits are aggregated. Using Foursquare's category hierarchy we can investigate the types of venue that are visited and how their usage is influenced by time, and the extent to which patterns can be seen at the aggregate level.

4.1.1 Collective checkins by category

Foursquare's category hierarchy consists of eight top-level categories, each of which can have a number of sub-categories and sub-sub-categories. These sub-categories are too many and too specific to easily examine, so we have grouped venues by their top-level categories, as shown for Cambridge and Cardiff in Figure 4.1.

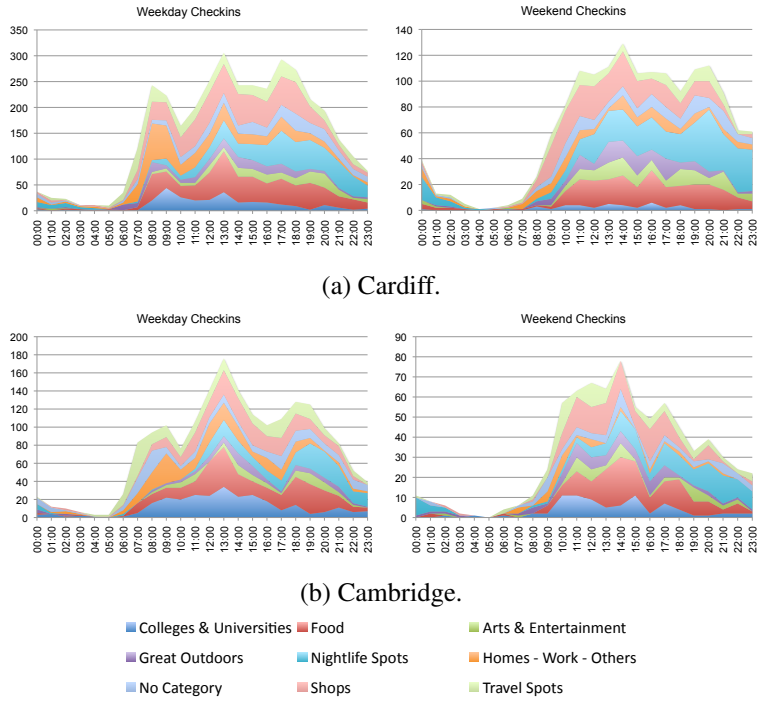
This figure reveals some interesting characteristics of individuals in the two cities. For example, Cambridge has almost as many checkins in the *College & Universities* category as the *Food* category, while in Cardiff checkins in the *Food* category outnumber those in the *College & Universities* category by almost 2 to 1. Looking at ratios between the number of venues and the number of checkins reveals further details, for instance in Cambridge the ratio of venues to checkins in the *Nightlife Spots* category is approximately $3:1$, while in Cardiff this ratio is approximately $6:1$.

Although this aggregate view tells us about general behaviour of individuals, it does not indicate when individuals tend to visit locations in these categories. To capture more of the temporal component, we can instead look at the venues visited by hour-of-day.

4.1.2 Daily check-in behaviour

The per-hour distribution of checkins offers insights into how the behaviour of Foursquare users is influenced by the time of day. Although this only describes collective behaviour, it does indicate when certain categories of venue are more popular.

The plots in Figure 4.2 show the total number of checkins during each hour of the



(a) Cardiff.

(b) Cambridge.

Figure 4.2: Number of checkins per hour of the day in Cardiff and Cambridge. The contribution of each venue category to the total checkins for each hour is indicated by colour

day, along with the categories of venues that were visited. To limit potential bias due to occasional interruptions in data collection, a 21-day period with the fewest outages was selected. The number of checkins recorded in this period were 5,226 and 2,609 for Cardiff and Cambridge respectively. Checkins occurring within the same hour of the day were aggregated to produce the figures. We also differentiate between checkins on weekdays (i.e., Monday to Friday) and checkins on weekends (i.e., Saturday and Sunday), as we expect work patterns to have a strong influence on weekday check-in behaviour.

We found that there was a slight increase in the number of checkins per day on weekends. Cardiff averaged $\mu_1 = 246.2$ checkins per day on weekdays (with $\sigma_1 = 57.7$) and $\mu_2 = 255.5$ checkins per day on weekends (with $\sigma_2 = 73.6$). A comparable increase was found in Cambridge, which averaged $\mu_1 = 123.5$ checkins per day on weekdays (with $\sigma_1 = 29.3$) and $\mu_2 = 126.0$ checkins per day on weekends (with $\sigma_2 = 51.6$). Within each city we tested whether the difference in the weekday and weekend means

were statistically significant by applying the two-sample T-test with the null hypothesis $H_0 : \mu_1 = \mu_2$. The resulting p-values lead us to accept H_0 for both cities ($p = 0.79$ for Cardiff and $p = 0.91$ for Cambridge) at the 0.05 significance level. Therefore we conclude that there is no statistically significant change in the number of checkins on weekdays versus checkins on weekends.

Both cities also appear to have the same ratio of weekday checkins to weekend checkins. The fraction of weekday checkins is $f_1 = 0.704$ for Cardiff and $f_2 = 0.710$ for Cambridge. We tested the statistical significance of this similarity between Cardiff and Cambridge using the two-proportion Z-test with the null hypothesis $H_0 : f_1 = f_2$. The test gave a p-value of 0.55, indicating that there is no statistically significant difference between the two ratios.

Despite the negligible difference in the number of checkins per day, the distribution of checkins on weekdays and weekends are noticeably different. In the case of weekdays, we find that both cities have three bursts of checkins that occur during the day. In Cardiff there is a morning burst from 07:00 to 10:00, an early-afternoon burst from 12:00 to 14:00, and an early-evening burst from 17:00 to 19:00. Cambridge checkins follow a similar pattern, but with the morning and early-evening peaks being smaller relative to the early-afternoon peak.

The morning burst on weekdays is likely due to users waking up and checking in to venues as they travel to work. Indeed, in the 07:00 to 10:00 burst the four most-popular venue categories for Cardiff are *Home - Work - Others* (29%), *Food* (15%), *Shops* (15%), and *Travel Spots* (15%). It is not surprising that the early afternoon (12:00 to 14:00) checkins are predominantly at *Food* (24%) and *Shops* (21%) venues. The percentage of *Home - Work - Others* checkins in the early afternoon (11%) is much smaller than in the morning burst (29%), which may indicate that either many individuals do not leave work for lunch or neglect to check-in on their return to work. A notable change in the early evening checkins is an increase in the percentage of *Nightlife Spots* checkins (21% for 17:00 to 19:00). This category becomes increasingly popular as the evening progresses.

Although we can observe a three-burst pattern for both Cardiff and Cambridge during

weekdays, there is no strong similarity between their weekend checkin patterns. The number of checkins per hour in Cardiff does not drastically change between 11:00 and 21:00. Cambridge has a high rate of checking in between 10:00 and 15:00; however, unlike Cardiff, this rate is not sustained into the evening. These results indicate that weekday user behaviour is predominantly driven by routine, whereas there is scope for more variation and less predictability in weekend patterns. Chen et al. [CCLS11] have found similar weekday structure in visit patterns within other cities, indicating that this may be a universal behaviour.

4.1.3 Summary

At the collective scale, we have found a strong relationship between the time of week and visits to Foursquare venues. The checkin rate across all venue categories follows a three-burst pattern during weekdays. On weekends, however, this three-burst pattern is much less pronounced. In this case study we also demonstrated how moving from AGGREGATE temporal context to PERIODIC revealed richer information about the categories visited by users over time. In particular, Section 4.1.1 shows that the aggregate number of checkins per category varies, and in Section 4.1.2 we see that the type of venue visits is related to the time of week. In the rest of this thesis we narrow our scale further, moving our focus to INDIVIDUAL-PERIODIC.

4.2 Measuring individual regularity

The investigation into collective-scale regularity discussed in Section 4.1, and found in other studies (e.g., [CCLS11]), shows that there is a periodic component in aggregate human visiting patterns. On an individual scale, however, there has been little work examining the extent to which individuals' regular (or irregular) patterns contribute to the population's aggregate behaviour. By focusing on the individual scale we can explore the patterns of individuals from which the collective properties emerge. In this section we develop a method to quantify the amount of regularity in the visits of an individual to a particular location.

We define regularity as a visiting pattern that is repeated with a reoccurring time-frame (for example, on a week-by-week or day-by-day basis); in other words, an individual that consistently visits a particular location at similar times each week is defined as having a highly regular visiting pattern with that location.

The event-based nature of the data we consider precludes the application of existing methods that require continuous, densely-sampled data. User visit data such as this is very sparse and consequently challenging to effectively model. Even the most active users typically visit the same place fewer than a dozen times per week. This sparsity makes it difficult to apply many established approaches for measuring regularity and periodicity, such as nonlinear time series analysis, harmonic analysis, and recurrence quantification analysis, as these are most effective for time series that are continuous and densely sampled. As an alternative to these approaches, we can instead draw on the large body of relevant work in the neurophysiology community dealing with the problem of finding regularity in event-based data. In particular, we adapt an efficient *neural synchrony* measure named *ISI-diversity* (inter-spike interval diversity [KCA$^+$09, KCGA11]) to develop a method of quantifying the regularity of a visit's chronology. To the best of my knowledge this is the first application of neural synchrony methods to deal with event-based human mobility data.

4.2.1 Neural synchrony methods

Neural coding is the branch of neurophysiology concerned with the coding of information among the neurons in the brain. In the context of neural coding, neurophysiologists deal with ensembles of *spike trains*, where each train represents the instantaneous electrical *spikes* (or pulses) of a particular neuron. A spike train is a model of neuron activity as zero-duration events, similar to how we model visit data in this thesis.

Identifying patterns in spike trains has become an essential part of analysing and understanding neuron activity. A large variety of methods and tools tackling this task have been presented in the neurophysiology literature. Among the methods applied, techniques detecting regular neuron firing patterns have borrowed from information theory [BD05], wavelet analysis [RBY$^+$01, PP06], and Fourier analysis [HR99]. Some of

these approaches require binning of data along the temporal axis, and therefore result in a loss of temporal resolution and the addition of an extra parameter (the bin width).

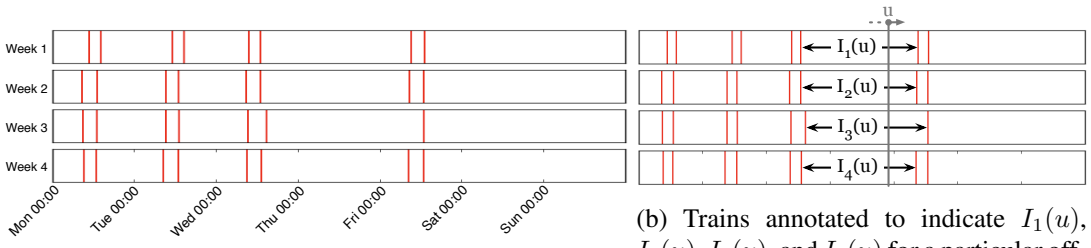
To quantify regularity we adapt a neural synchrony measure named *ISI-diversity* (inter-spike interval diversity) [KCA+09]. Neural synchrony is the behaviour of an ensemble of neurons to jointly fire with a similar pattern [BKM04]. An ensemble of spike trains is said to exhibit high synchrony if the spikes in the trains occur at similar times. ISI-diversity is one of a number of methods developed to detect this behaviour [KCGA11]. In the domain of neural coding this measure is preferred because it considers all trains concurrently, as opposed to other methods which evaluate pairwise similarity and produce an average. The measure also retains temporal resolution of events, rather than quantising the trains through binning, and has the additional benefit of being parameter free.

Spikes can be regarded as abstract, zero-duration events; for the purpose of measuring visit regularity, spikes correspond to visits to a particular location. In the rest of this thesis we will generalise the concept of a spike to an event and refer to an inter-spike interval (ISI) as an inter-event interval (IEI).

4.2.2 IEI-irregularity

Here we present the regularity measure used in this chapter, named *IEI-irregularity* (inter-event interval irregularity). Our approach is based on measuring the dissimilarity in the timing of visits on a per week (or other period such as per day) basis. We use ISI-diversity to quantify the level of dissimilarity in visits in different weeks; if visits in each week occur at very similar times, then dissimilarity is very low, and thus regularity is high. This models regularity as repeated routine over time. For example, an individual visiting a location at very similar times each week is considered to have a highly regular pattern for that location. On the other hand, if the individual visits the location at very different times each week it is considered to be a very irregular pattern.

Throughout this chapter we use week-by-week comparison to determine regularity; however, in the following formulation we generalise this to any **window size**, denoted by ω. Although human mobility follows cycles at multiple scales (e.g., daily, weekly,



(a) Example trains.

(b) Trains annotated to indicate $I_1(u)$, $I_2(u)$, $I_3(u)$, and $I_4(u)$ for a particular offset u.

Figure 4.3: Example visit trains for a particular user and access point in the DARTMOUTH dataset. Window width $\omega = 7$ days.

biweekly, and yearly), here we will focus on weekly regularity as this is one of the most prominent and captures a range of routine, including weekday and weekend patterns.

More formally, recall that $\mathcal{S}_{v,l}$ is a chronology of an individual v's visits to a location l, denoted by the ordered sequence of times

$$\mathcal{S}_{v,l} = \{t_i \,|\, i = 1, \ldots, L\}\,,$$

where L is the number of v's visits to l (Definition 3.1). These times are assumed to be offsets from some arbitrary origin, giving values $t_i \in (0, T_{max}]\ \forall\, i = 1, \ldots, L$. The chronology is segmented into disjoint windows of duration ω to build N **visit trains**. The absolute times of visits are translated to offsets from the start time of their corresponding window; thus, each train has visit times in the interval $(0, \omega]$. We assume T_{max} and ω are chosen such that $\omega\, N = T_{max}$. We denote the number of visits in the nth train with L_n and the sequence of visit time offsets in train n with

$$U^n = \{u_i^n \,|\, i = 1, \ldots, L_n\}\,.$$

An example of the visit trains for a four-week chronology in the DARTMOUTH dataset are shown in Figure 4.3a.

Irregularity is quantified by applying the **ISI-diversity** [KCA+09] measure to the ensemble of N visit trains. The measure is computationally efficient, scaling linearly in both the number of visit trains N and number of visits L [KCA+09, KCGA11].

We begin by defining the inter-event interval (IEI) as the time between two consecutive visits.

Definition 4.1

The **instantaneous inter-event interval function** $I^n(u)$ gives the IEI for the nth

train at time offset $u \in (0, \omega]$; formally, instantaneous IEI is defined for three cases:

$$I^n(u) = u_1^n \quad \text{if} \quad 0 < u \leq u_1^n\,,$$
$$I^n(u) = \omega - u_{L_n}^n \quad \text{if} \quad u_{L_n}^n < u \leq \omega\,,$$

and

$$I^n(u) = \min(u_i^n \mid u_i^n \geq u) - \max(u_i^n \mid u_i^n < u) \quad \text{if} \quad u_1^n < u \leq u_{L_n}^n.$$

Figure 4.3b shows the example train from Figure 4.3a annotated with example instantaneous IEI values at a particular offset u.

We define two further instantaneous measures which are computed from instantaneous inter-event intervals. For time offset u, the instantaneous mean $\mu(u)$ is given by

$$\mu(u) = \frac{1}{N} \sum_{n=1}^{N} I^n(u)$$

and the instantaneous standard deviation $\sigma(u)$ is given by

$$\sigma(u) = \left(\frac{1}{N-1} \sum_{n=1}^{N} \left(I^n(u) - \mu(u)\right)^2 \right)^{1/2}.$$

Using these two instantaneous measures we can evaluate the dispersion in IEI values at a particular time offset, which represents the degree of dissimilarity in the timings of events across the N trains at that offset.

Definition 4.2

The **coefficient of variation** $c_{\text{var}}(u)$ provides a measure of dispersion in the IEI values at time offset u,

$$c_{\text{var}}(u) = \frac{\sigma(u)}{\mu(u)}\,.$$

$c_{\text{var}}(u)$ is a unitless measure and normalised against the mean, which enables comparison between the dispersion in collections of large IEI values and collections of small IEI values.

By integrating over time offset u we obtain a measure of overall dissimilarity $D(\mathcal{S}_{v,l})$ in the ensemble of visit trains for chronology $\mathcal{S}_{v,l}$.

Definition 4.3

The **IEI-irregularity** of a visit chronology $\mathcal{S}_{v,l}$, denoted $D(\mathcal{S}_{v,l})$, is given by

$$D(\mathcal{S}_{v,l}) = \frac{1}{\omega} \int_0^{\omega} c_{\text{var}}(u)\; du\,.$$

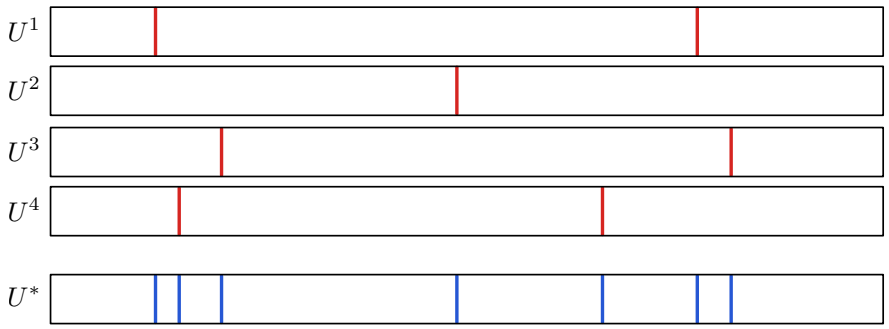


Figure 4.4: Four example visit trains (U^1, U^2, U^3, and U^4) and their corresponding master train U^*.

The resulting $D(\mathcal{S}_{v,l})$ is a non-negative value, with $D(\mathcal{S}_{v,l}) = 0$ indicating identical trains (i.e., perfect regularity), and higher values indicating more irregularity in the visiting patterns.

4.2.3 Computing IEI-irregularity

IEI values are computed by inserting dummy event times 0 and ω into each train and computing the difference between neighbouring events in each train. The arrival of an event in a train indicates a change in the instantaneous IEI for the period up to (but not including) the next event in that train. In other words, given two consecutive spikes u_i^n and u_{i+1}^n in train n, the instantaneous inter-event interval $I^n(u)$ during $u \in (u_i^n, u_{i+1}^n]$ is $u_{i+1}^n - u_i^n$.

As part of computation it is necessary to know offset intervals where there are no intervening events. These intervals represent durations where $c_{\text{var}}(\cdot)$ is constant. To support this task we create a master train to store the event times taken from all trains.

Definition 4.4

A **master train**, denoted U^*, is the sequence of events $U^* = U^1 \cup \ldots \cup U^N$ in ascending order. For convenience we write the master train and its ordered events as $U^* = \{u_1^*, \ldots, u_L^*\}$.

An example of a collection of visit trains and the corresponding master train is depicted in Figure 4.4.

We compute a new $c_{\text{var}}(\cdot)$ at each event in the master train. For each $i = 2, \ldots, L$

a coefficient of variation $c_{\mathrm{var}}(u_i^*)$ is computed which yields the constant coefficient of variation for the interval $(u_{i-1}^*, u_i^*]$. In addition, we also calculate the coefficient of variation at the first event (i.e., u_1^*) to give the dispersion over $(0, u_1^*]$, and at ω to give the dispersion over $(u_L^*, \omega]$. Finally, to calculate the IEI-irregularity $D(\mathcal{S}_{v,l})$, we compute the sum of the $c_{\mathrm{var}}(\cdot)$ values for each of these $L+1$ intervals, weighted by the duration of the interval; in other words,

$$D(\mathcal{S}_{v,l}) = \frac{1}{\omega}\left(c_{\mathrm{var}}(u_i^*)\,(u_i^*) \;+\; c_{\mathrm{var}}(\omega)\,(\omega - u_L^*) \;+\; \sum_{i=2}^{L} c_{\mathrm{var}}(u_i^*)\,(u_i^* - u_{i-1}^*)\right) .$$

Two variables relevant to the algorithm's time complexity are the number of events L and the number of trains N. Given that the input chronology is in ascending order, the task of translating it to N trains and a master train is trivial. Computing the coefficient of variation at a given time offset requires a look up of one instantaneous IEI value from each train, making the time complexity linear in the number of trains N for fixed L. To understand the reciprocal case (i.e., fixed N and varying L) we consider the effect of adding an event to a chronology. The result is one additional interval in the master train, requiring one additional calculation of the coefficient of variation and another value involved in the weighted sum calculation. Therefore, assuming fixed N, the algorithm is linear in the number of events L.

4.3 Character and prevalence of regularity in visiting patterns

In this section IEI-irregularity is applied to real-world data to explore regularity in human visit patterns. In particular, we compare four-week segments within each of our three visit datasets. The base datasets from which we derive our four-week datasets are discussed in Section 4.3.1. We divide our analysis into four areas of interest. We first consider the influence of the time of week on the inter-visit intervals of chronologies (Section 4.3.2). In Section 4.3.3 we compare the datasets in terms of their irregularity, followed by studying how the types of the locations within the datasets contribute to the overall irregularity (Section 4.3.4). Finally, we consider how prevalent regular visiting

Dataset	FOURSQUARE	DARTMOUTH	UNDERGROUND
Area(s)	Bristol, Cardiff, and Camb.	Dartmouth	London
Scale	Urban	Campus	Metropolitan
Month	June	April	March
Location type	Venue	Access point	Metro station
Visit type	Checkin	Association	Card swipe
Individuals	293	1,681	1,167,363
Locations	336	391	270
Visits	4,640	229,300	58,945,475
M	401	3,656	2,260,354
$\langle L \rangle$	11.6	62.7	26.1

Table 4.1: Summary of datasets used in the analysis of regularity. Each dataset corresponds to a four-week period. M denotes the number of chronologies and $\langle L \rangle$ denotes the mean number of visits per chronology. A chronology $\mathcal{S}_{v,l}$ is only included in a dataset if v visited l at least twice in each of the four weeks. Locations and individuals are only counted if they were involved in at least one chronology.

patterns are among the individuals in each dataset and the relationship between visit frequency and visit regularity (Section 4.3.5).

4.3.1 Four-week visit datasets

A four-week segment was selected from each dataset to obtain visit chronologies for study. We selected four weeks in June from FOURSQUARE and four weeks in March from UNDERGROUND (these are the only weeks available to us). The DARTMOUTH dataset is particularly sensitive to the time-of-year because much of the user behaviour is driven by teaching semesters (changing usage patterns during each year are discussed in [HKA08]). We therefore selected four weeks in April 2003 from DARTMOUTH as this corresponds to an uninterrupted period of teaching time at Dartmouth College.

The original datasets (as presented in Section 3.2) contained many individuals that visited certain locations very rarely or exclusively in only a few of the four weeks. Chronologies such as these are not suitable for studying regularity, as their activity is too rare and too transient. We restrict the four-week datasets to chronologies with at least two visits in each of the four weeks. The derived four-week datasets used in this chapter are summarised in Table 4.1. The filtering of infrequent chronologies culled

roughly 93% of the original four-week person-location pairs in both DARTMOUTH and UNDERGROUND, indicating that, although the set of places a person has visited at least once may be large, many of these places are only visited very occasionally. The number of chronologies for FOURSQUARE reduced to 3% of the original, leaving a small sample of 401. The remaining chronologies in FOURSQUARE involve 4% of the users, a small proportion compared to 67% in DARTMOUTH and 23% in UNDERGROUND. A substantial number of transient users (such as tourists or individuals passing through) may also contribute to this substantial reduction.

4.3.2 Inter-visit intervals and the time of week

As discussed in Section 4.2, our approach focuses on the weekly patterns of inter-event intervals (IEIs) for an individual's visits to a particular location. The IEIs themselves, along with their level of dispersion at a particular time-of-week, are an interesting property of human mobility and thus we consider them specifically in Figure 4.5. The figure shows how IEI dispersion (as quantified by the coefficient of variation c_{var} of a chronology at a given time-of-week) varies throughout the week.

The small standard deviations in visit rates indicate that the volume of visits is very similar in each week. This contrasts with the $\langle c_{\mathrm{var}} \rangle$ values which have very high standard deviation. This highlights the person-specific nature of an individual's visiting patterns with a location; in other words, the visiting patterns (and therefore IEIs) of two different individuals visiting the same location can be very different.

In the UNDERGROUND dataset we observe that, on average, chronologies' IEIs are most-dispersed between 10:00 and 16:00 on weekdays, and least-dispersed during nighttime. This is because the relative effect of a discrepancy in visit times that are close together is greater than when the visit times are further apart. For example, the morning and afternoon commute on the same day are separated by roughly nine hours, whereas the time between the afternoon commute and the following day's morning commute is roughly 15 hours. Therefore, minor discrepancies in the visits to a commuter's stations will have a greater influence on the dispersion of daytime IEIs than nighttime IEIs. The same behaviour is responsible for the dip in IEI dispersion dur-

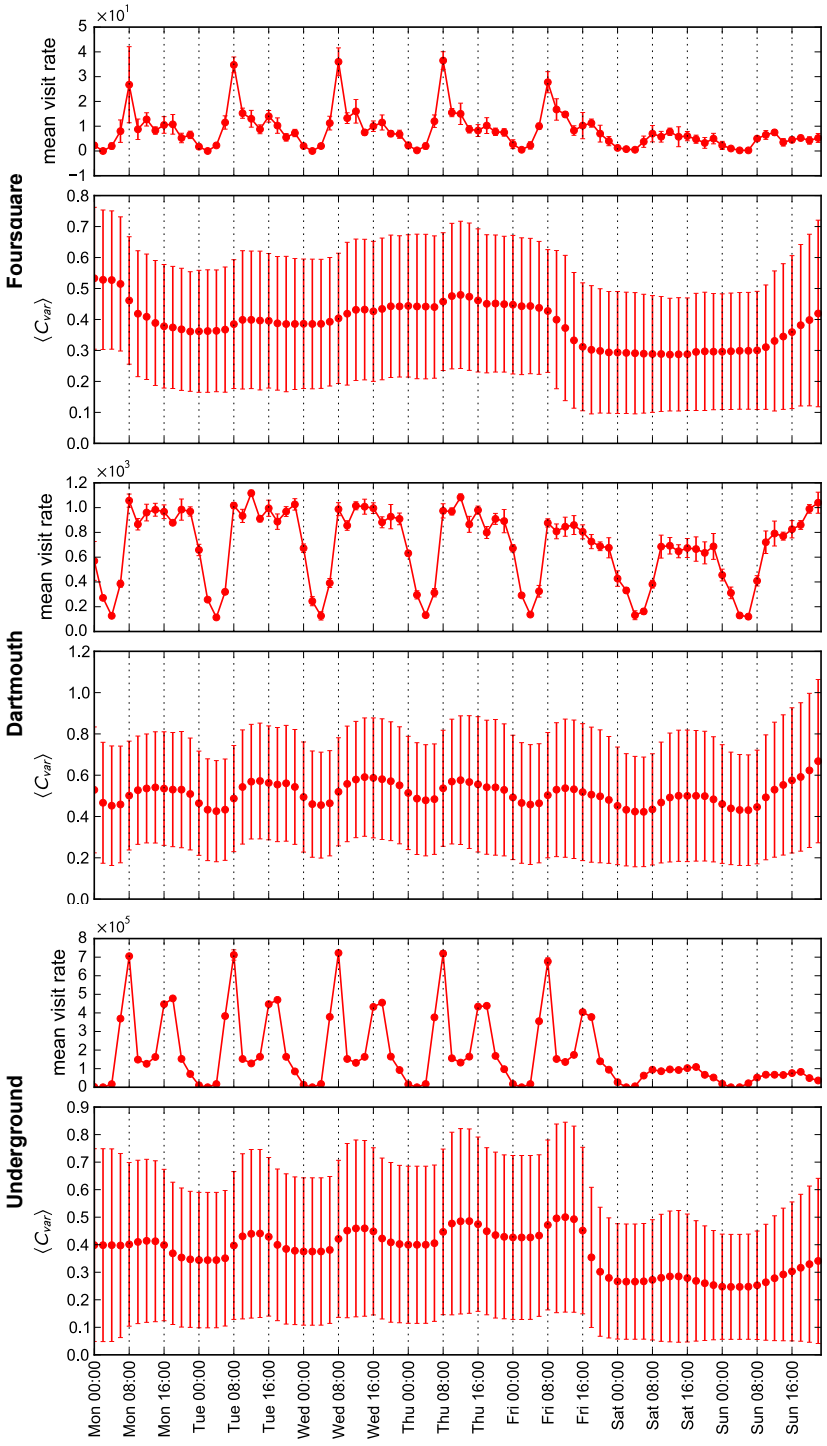


Figure 4.5: Time-of-week means of visit rates and coefficients of variation ($\langle c_{var} \rangle$) for each dataset. $\langle c_{var} \rangle$ is obtained by averaging over the c_{var} values in the corresponding two-hour time slot of all chronologies. A high $\langle c_{var} \rangle$ indicates that the instantaneous IEI values were, on average, more dispersed during that time of week.

ing the weekend. Many chronologies consist of predominantly weekday visits. The weekends for these chronologies will correspond to large IEIs spanning from Friday to Monday, and so the dispersion (c_{var}) will be less during this period.

When comparing DARTMOUTH and UNDERGROUND we note that DARTMOUTH's weekday visit activity is sustained throughout the day and lasts longer into the evening, rarely declining before midnight. This reflects the fact that the DARTMOUTH dataset includes many types of visit (including social, residential, and academic), whereas UNDERGROUND is restricted to transportation. This late-evening visit activity is also the reason for the delayed dip in IEI dispersion, which does not decrease until 22:00 (compared to 16:00 in UNDERGROUND). It is also worth noting that the DARTMOUTH decline in visit rate on the weekend is small. This is explained by a large number of students living on-campus, compared to a small proportion of students and staff who either live off-campus or spend the weekend elsewhere.

4.3.3 Comparison of regularity between datasets

Given that the three datasets differ in context, time of year, and geographic scale we would expect differing visiting behaviours in each. Indeed, we have already discussed how the three datasets' time-of-week visit rates exhibit different patterns. The same is also true of the level of regularity present in each dataset, as shown in Figure 4.6. DARTMOUTH is distinct from the other two datasets, with the weight of its distribution shifted towards higher irregularity. This is reflected in the mean irregularity $\langle D \rangle$ (which we take over the available user-location chronologies), which is higher for DARTMOUTH (0.510) than for FOURSQUARE and UNDERGROUND (0.381 and 0.373, respectively). This suggests that the patterns of individuals visiting locations on Dartmouth campus tend to be more irregular. This is unlikely to be due to a sudden change in routine, as the duration of the dataset (April 2003) is a continuous period of term-time teaching, uninterrupted by holidays or exams. The small deviations in visit rates (see visit rate plots in Figure 4.5) also indicate that there was no overall change in visiting patterns between the weeks. An alternative reason for the increased irregularity may be the highly dynamic and spontaneous nature of student behaviour. This con-

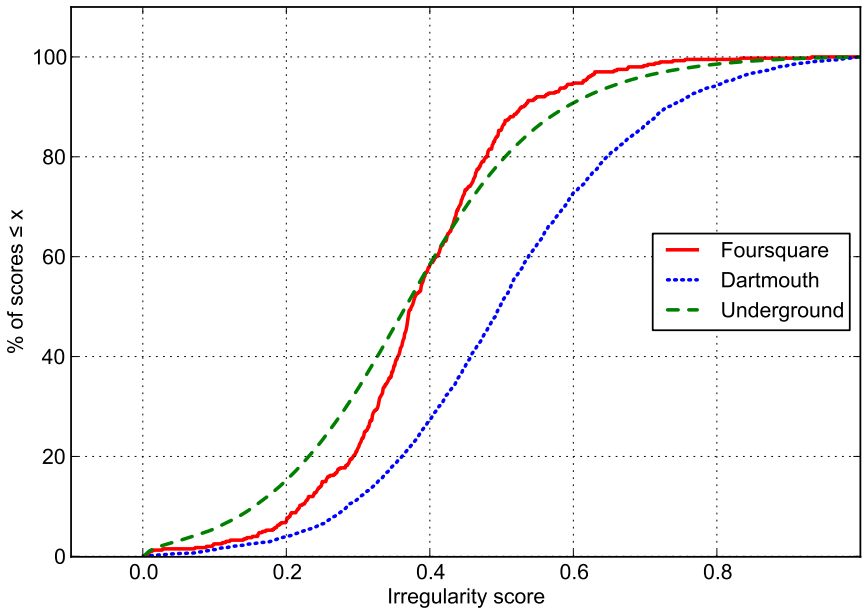


Figure 4.6: Cumulative distributions of IEI-irregularity scores (i.e., $D(\cdot)$ values) in each dataset. High $D(\cdot)$ indicates high irregularity. The mean IEI-irregularity value $\langle D \rangle$ is 0.381 ($\pm$ 0.131) for FOURSQUARE, 0.510 ($\pm$ 0.185) for DARTMOUTH, and 0.373 ($\pm$ 0.173) for UNDERGROUND.

trasts with Underground passengers and Foursquare users, whose student proportion is likely to be much smaller, consisting instead of a large population of individuals following less-flexible routines (for example, commuters).

The finer-grained localisation of the DARTMOUTH dataset may also contribute to the increased irregularity. The Dartmouth APs had an indoor range of up to 40m, so most buildings required multiple APs to achieve good WLAN coverage. This means that users moving as little as a few tens of metres can register as having visited a new location. These short-distance movements are likely to be more unpredictable and driven less by routine than larger-distance movements, and thus result in higher irregularity in AP visits.

We also note the similar mean irregularities of FOURQSUARE and UNDERGROUND chronologies, which may be attributed to both datasets being at a city-wide scale and consisting of a broad cross-section of people, as opposed to Dartmouth campus's predominantly student population.

FOURSQUARE			
Venue category	M	$\langle L \rangle$	$\langle D \rangle$
Arts & Entertainment	6	8.5	0.335 ($\pm$ 0.215)
Food	35	8.5	0.354 ($\pm$ 0.103)
Nightlife Spots	32	8.6	0.356 ($\pm$ 0.130)
Shops	63	9.0	0.373 ($\pm$ 0.112)
Homes/Work/Others	139	14.0	0.374 ($\pm$ 0.123)
Travel Spots	36	11.6	0.380 ($\pm$ 0.120)
Colleges & Universities	37	13.7	0.432 ($\pm$ 0.153)
Great Outdoors	20	11.7	0.438 ($\pm$ 0.166)
DARTMOUTH			
Building type	M	$\langle L \rangle$	$\langle D \rangle$
Academic	965	30.1	0.375 ($\pm$ 0.165)
Library	135	26.6	0.445 ($\pm$ 0.169)
Social	119	42.3	0.446 ($\pm$ 0.168)
Admin	81	59.7	0.531 ($\pm$ 0.185)
Residence	2,276	79.5	0.573 ($\pm$ 0.161)
Athletic	65	78.9	0.579 ($\pm$ 0.185)

Table 4.2: Comparison of irregularity by type of location. For each subpopulation of chronologies we show the number of traces M, the mean IEI-irregularity value $\langle D \rangle$ along with its standard deviation, and the mean number of visits per chronology $\langle L \rangle$. Uncategorised Foursquare venues and Dartmouth APs are not included.

4.3.4 Influence of location type

To further study regularity in these datasets we separate the chronologies into subpopulations by the type of location they are involved in. Table 4.2 lists the location types in the FOURSQUARE and DARTMOUTH datasets, along with their subpopulation irregularity means and other relevant properties.

APs in the Dartmouth dataset are categorised by the type of building they were placed in. For Foursquare venues we used the venue's top-level category. We also show the distributions of irregularity values in the Dartmouth subpopulations in Figure 4.7. We do not plot the FOURSQUARE subpopulations due to the small sample sizes.

The results for DARTMOUTH (Table 4.2 and Figure 4.7) reveal that the source of the dataset's high overall irregularity (discussed in Section 4.3.3) is the large *Residence* subpopulation (consisting of $M = 2{,}276$ chronologies) with high mean irregularity ($\langle D \rangle = 0.573$). It is intuitive that residential locations are the least regular. First, since

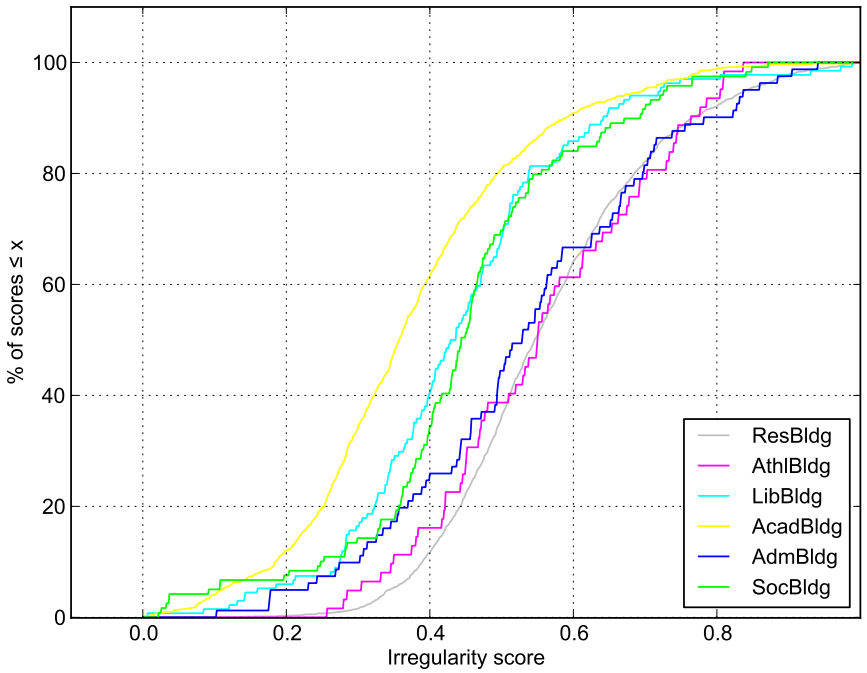


Figure 4.7: Cumulative distributions showing the distribution of IEI-irregularity scores (i.e., $D(\cdot)$ values) by type of location in the DARTMOUTH dataset.

these are where students spend most of their time, there are likely to be more fine-grained movements to different APs in the same building, which are likely to increase irregularity. Second, the comparatively small area of the campus means it is convenient for students to return to their residence between making visits elsewhere; therefore, many visits to residences depend on other events (such as the end of a lecture), allowing for a lot more variation in the week-by-week visits. Residences also have no curfew, unlike other types of building that have restricted periods of use (libraries have limited opening times, for example).

Another intuitive result is that the subpopulation with the lowest mean irregularity is the *Academic* subpopulation ($\langle D \rangle = 0.375$). This is unsurprising since most visits to academic locations are caused by timetabled events such as weekly lectures and seminars.

In the case of FOURSQUARE, we find that the *Great Outdoors* subpopulation of chronologies had the highest mean irregularity ($\langle D \rangle = 0.438$). This can be explained by the nature of these venues. As they are outdoor venues, they are highly subject to weather conditions. Many outdoor activities will therefore be rescheduled to coincide

with fairer weather, rather than occurring at a specific time of week.

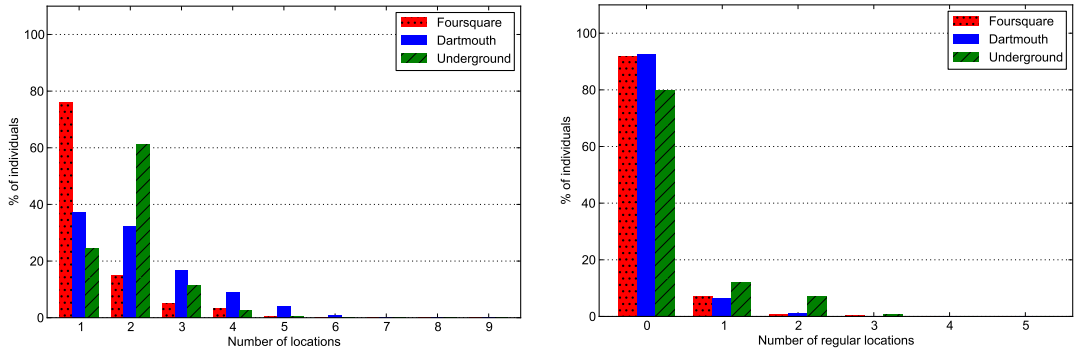
It is surprising that *Colleges & Universities* is among the least-regular subpopulations. We would expect its visits to be driven by a timetabled routine, as with DARTMOUTH's academic buildings. We note that this may be because Foursquare's *Colleges & Universities* top-level category includes a wide variety of subcategories, such as cafeterias and accommodation. Further sub-dividing the subpopulation of chronologies may reveal subcategories of high regularity; however, the sample size prevents any strong conclusions from being drawn.

4.3.5 Prevalence of regularity among individuals

We now study the extent to which an individual has regular relationships with the locations he or she visits. We begin by considering the overall number of locations individuals tend to visit, as shown in Figure 4.8a. In FOURSQUARE and DARTMOUTH the percentage of individuals decreases with the number of different locations, with DARTMOUTH users typically visiting a wider variety of locations. UNDERGROUND follows a similar pattern, except its peak is at two locations rather than one, which is explained by the nature of Underground journeys. An individual with only one location can occur in the rare case where a passenger bypasses the exit turnstile or exits from the entry station, or where a passenger has had one of their stations discarded in the minimum-visits filtering we discussed in Section 4.3.1.

Using the IEI-irregularity $D(\mathcal{S}_{v,l})$ of an individual v's visits to location l we can evaluate whether v's visits to l are regular or irregular. We set a threshold for irregularity, below which we will regard v's visits to l as regular. In Figure 4.8b we plot the distribution of individuals and how many of the locations they visited were deemed regular in this way. We set a strict threshold of 0.2, as we wish to find the chronologies with near-perfect regularity. As shown in Figure 4.6, a minority of chronologies in each dataset are within this threshold (8.2% in FOURSQUARE, 4.4% in DARTMOUTH, and 17.4% in UNDERGROUND).

Figure 4.8b shows how the set of highly regular chronologies is distributed among the individuals. 8% of Foursquare users and Dartmouth WLAN users had at least one



(a) Distributions showing the number of locations for individuals used in our analyses. Individuals exceeding nine locations are not plotted.

(b) Distributions showing the number of *regular* locations (i.e., where $D(\mathcal{S}_{v,l}) \leq 0.2$) for individuals in each dataset.

Figure 4.8: Number of regular locations per individual compared to the overall number of locations per individual.

location that they visited with high regularity. The percentage increases in the case of Underground passengers, with 21% of individuals having at least one regular location, likely due to the more-routine nature of travel. At stricter thresholds (i.e., thresholds closer to 0), the size of the core group of users with at least one regular venue decreases. The threshold at which the size of this group dropped to 1% of individuals was 0.009 for FOURSQUARE, 0.050 for DARTMOUTH, and 0.007 for UNDERGROUND.

We also consider whether there is any relationship between an Underground passenger's most-visited station and his or her most-irregular station. Most-visited stations are likely to be 'home' stations, which we expect to have irregular visiting patterns, since they represent a convolution of many different routines throughout the week. We consider the probability $p(m)$ that, given an individual v who has visited m stations, the individual's most irregular station l (i.e., l such that $D(\mathcal{S}_{v,l})$ is maximised) is also the station that v visited the most. We find that $p(2) = 0.55$, $p(3) = 0.37$, $p(4) = 0.29$, $p(5) = 0.28$, and $p(6) = 0.28$, indicating that the probability of these stations matching is slightly higher than chance. The deviation from chance becomes greater when individuals have four or more frequently visited stations. This deviation is more significant in DARTMOUTH, which has probabilities $p(2) = 0.57$, $p(3) = 0.47$, and $p(4) = 0.43$.

Finally, a relevant insight to this thesis we wish to consider is the finding in [LC11] that a transport network user's perception of their regularity is not necessarily related to the amount of travel. Through application of our regularity measure (Figure 4.9) we

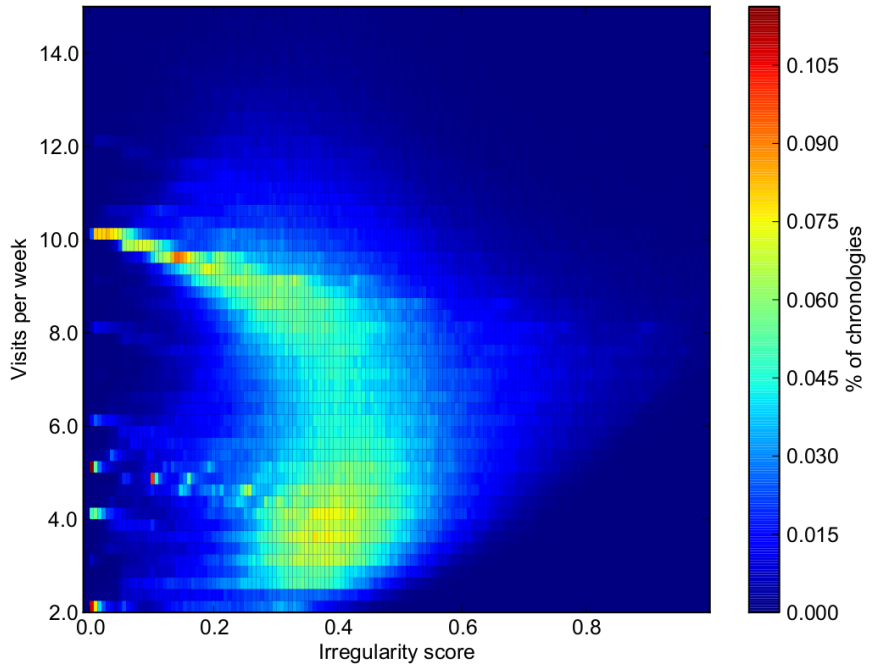


Figure 4.9: Joint frequency distribution of visit frequencies and irregularity scores of chronologies in the UNDERGROUND dataset. Non-zero probabilities occur outside the plotted area but are omitted due to their rarity.

can see that this is indeed the case. The joint distribution shows that there is no linear correlation between chronology visit rate and IEI-irregularity.

4.4 Discussion and related work

At the collective scale, the three-burst pattern we noted in aggregate Foursquare visiting behaviour has also been observed in the hourly variation in the rate of cell transitions of mobile phone users [CGW+08]. This is not surprising since morning, lunchtime, and evening correspond to periods of movement activity. One difference between the two datasets, however, is that the largest peak in cell transitions is in the morning, whereas the Foursquare morning burst is smaller than both the lunchtime and the evening bursts. This is explained by the differing definitions of location in the two datasets. In [CGW+08], the many cell tower transitions during long-distance commutes are recorded as location visits. On the other hand, Foursquare checkins are self-reported locations that have some degree of significance for the user.

At the individual scale, Eagle and Pentland [EP06] have also presented an approach to quantifying patterns in human visits. Information entropy is used to measure the predictability of mobile phone users' patterns of transition between home and work. The work we have presented attempts to go beyond only home and work, considering the many other locations a person visits. An interesting observation in [EP06] is that university students, especially those in their first year of study, have the highest entropy, and therefore are the least predictable. This agrees with our finding that DARTMOUTH individuals have higher irregularity. In addition, unlike the entropy measure presented in [EP06], our measure is time-of-week resolved, allowing us to further investigate how IEI dispersion within a particular chronology varies throughout the week, as discussed in Section 4.3.2.

While both the results of this chapter and those in [EP06] agree that student behaviour is less predictable, it is interesting that when collective-scale time-varying encounter statistics are analysed in [SMML10], DARTMOUTH is found to be the most regular of the datasets considered. This may be due to differing scales (individual and collective) and the differing nature of encounter and visit behaviour.

Song et al. [SQBB10] have made two key contributions relevant to our findings. First, the authors investigate a different but related concept of regularity, which is defined by them as the probability that an individual is found at his or her most-visited location. They find that this property is tied to the time-of-week, as we also observed with the mean coefficient of variation (Section 4.3.2). As previously mentioned, we have gone beyond the individual's most-visited location and consider their relationships with other places. We note that although [SQBB10] considers mobile phone record data, periodicity in return probability is also found in Foursquare checkins [CCLS11]. Second, Song et al. find that a significant amount of predictive information is encoded in the sequence and ordering of visits. With our regularity measure we have focused on IEIs and their variation by time of week; patterns in the sequences of IEIs is an interesting direction for future work.

4.5 Conclusions

In this chapter we detailed our definition of visit regularity and introduced a method for its measurement that is designed for event-based data. The method is adapted from the neural coding concept of synchrony and is computationally efficient. Our approach retains the temporal resolution of event timings and can be applied in cases of relatively low visit rates (e.g., as few as two visits per week).

Using this measure we have investigated the visiting patterns of individuals in our three datasets. Most existing analyses of regularity in visit patterns have either been at a collective scale or focused on an individual's relationship with their home location. Our work in this chapter goes further into visit periodicity than previous studies and specifically considers individuals' regular visit patterns with the many different places they visit. Our analyses have revealed features of individual behaviour beyond the weekday daily three-burst aggregate pattern we observed at the collective scale. We find that campus visits are the most irregular, likely due to the flexible nature of student behaviour, and transport visits are most regular, likely due to the significant commuter population. In all three datasets we find a core group of individuals that visit at least one location with near-perfect regularity. We also note a correlation between an individual's most-visited location (likely to be a associated with their home) and irregularity. We have observed that the type of location, which is typically associated with a particular activity, has strong influence over individuals' visiting patterns. There are location types whose usage is predominantly driven by inflexible constraints (such as lectures in academic buildings) whereas others, such as outdoor venues, are less constrained or subject to external random effects. We have found that there is no strong correlation between frequency of visit and regularity. Using IEI-irregularity we can identify irregular and regular visit behaviour, which we may wish to treat differently in some applications. For example, shop owners may wish to identify which of their customers visit routinely, but not necessarily frequently.

One limitation of our measure is that a very-regular pattern (e.g., a visit each Tuesday afternoon) embedded within other non-regular visits receives a high irregularity score, despite the presence of consistent recurrent behaviour among the visits. These embed-

ded patterns are particularly important in the case of encounters, where events are more frequent (as shown in the number of events per day in Section 3.2.5) and we need to identify repeated patterns embedded within many incidental occurrences. In the next chapter we will focus on encounter patterns specifically. One cause of regularity in encounters is through pairs and groups of individuals regularly visiting the same location; however, regular encounter patterns can also emerge from individuals visiting different locations together, such as friends who each lunch together at different locations.

Chapter 5

Periodicity in human encounter patterns

Introduction

The method introduced in Chapter 4 measured regularity in visit patterns, where regularity was defined as similarity in the weekly timing of visits. In this chapter we move on to consider encounter (i.e., person-meets-person) patterns, rather than person-visits-place patterns. Although both are functions of human mobility, encounter periodicity differs from visit periodicity in that the places where two individuals periodically meet may be different.

The effect of multiple incidental encounters mixing with periodic encounters makes the task of identifying periodic behaviour more challenging. As discussed in the previous chapter, our IEI-irregularity method for measuring periodic visit patterns treats additional non-periodic events as irregularities. In this chapter we instead explore a data mining perspective to develop a periodic encounter pattern detection method. Data mining provides an existing body of work dealing with the extraction of periodic events embedded among incidental events. Furthermore, these approaches allow the period of repetition to be treated as a feature and therefore are not restricted to only patterns of an a priori period (unlike the window parameter ω in Chapter 4). Collective-scale analysis of encounter networks [SMML10] has highlighted strong periodic components at different temporal scales (in particular, at one, seven, and 14-day periods) and therefore we begin our exploration of individual-scale encounter patterns by allowing more than one periodicity. This also generalises our approach to settings where we

are unable to anticipate the nodes' encounter periodicities beforehand, such as in wildlife networks. The challenging sparse and event-based nature of the encounter data requires us to quantise the original time-resolved events into bins to make data mining approaches applicable.

As with the method we developed in Chapter 4, our method for detecting periodic encounter patterns is designed to be computable on individual devices without the need for centralised infrastructure. To identify periodic encounters with direct neighbours this requires that each node analyse its pairwise local encounter histories. Furthermore, a pair of individuals who are involved in a periodic encounter relationship may also be periodically encountering other individuals, giving rise to broader periodic communities consisting of multiple individuals. Not all individuals in a community such as this are necessarily able to directly observe all other individuals in the community, and therefore we require a method for individuals to communicate and discover the broader periodic communities they belong to. We refer to such communities as *periodic encounter communities* (PECs). Devices can directly share, gain, and convey information and knowledge within their PECs with some degree of reliability due to the periodic re-occurrence of encounters.

Enabling nodes to detect the PECs they belong to provides them with useful context about the network they operate in, especially in the field of human encounter networks. The existence of PECs in a network has a substantial impact on the diffusion of information among mobile nodes and can be used to inform content forwarding decisions. Although many existing encounter-aware protocols focus on routing over pairwise (i.e., non-community) relationships, recent innovations, such as Habit [MMC09] and BUBBLE [HCY11], have shown that nodes can exploit multi-hop relationships (i.e., beyond neighbours) to improve the efficiency of disseminating media [MMC09] and to reduce routing delivery cost [HCY11]. PECs represent multi-hop encounter relationships that have stability over time, and therefore can be exploited in extensions to protocols such as Habit and BUBBLE. We note that our definition of PEC treats a periodic encounter pattern between a pair of individuals as a subcase of a periodic encounter community; in other words, our approach captures both pairwise patterns as well as community patterns.

Apart from human-based opportunistic networks (i.e., *human encounter networks*), other examples of application domains that would benefit from decentralised periodic encounter community detection include wildlife monitoring networks [JOW+02] and vehicular ad-hoc networks (VANETs) [LW07].

In this chapter we introduce, formalise, and model the concept of periodic encounter communities. This forms, to the best of my knowledge, the first treatment of the concept of a periodic encounter community and the first decentralised algorithm for their detection. The algorithm we present automatically detects the periodicities with which communities occur. Our approach combines data mining for the extraction of periodic encounter information at individual nodes with opportunistic sharing of this information when nodes are in communication range. Through opportunistic communication all nodes are able to discover the complete periodic encounter communities they appear in, including those parts of the community that a node cannot directly observe. We evaluate our approach using the real-world REALITY dataset and explore its behaviour with a number of metrics.

Chapter outline

The rest of the chapter is organised as follows. In Section 5.1 we formulate the PEC detection problem along with its local-knowledge variant and discuss the relation between PEC detection and the periodic subgraph mining problem from the literature. Our decentralised PEC detection algorithm is presented in Section 5.2. In Section 5.3 we introduce a model for investigating the information diffusion characteristics of PECs and the limits of decentralised PEC detection. This model is applied to the REALITY Bluetooth encounter dataset in Section 5.4 to analyse the decentralised PEC detection algorithm. Section 5.5 discusses related work in the area. Finally, we conclude this chapter in Section 5.6.

The work in this chapter appears in [WWA12a].

5.1 The PEC detection problem

A PEC (periodic encounter community) can be thought of as a group of nodes that encounter one another periodically. All pairs of nodes in the community do not necessarily have to directly encounter one another, but may instead have an acquaintance in common with whom they are both encountering with the same periodicity.

More formally, the structure of a PEC is defined in graph theoretic terms as a connected graph representing the nodes and their encounters. The temporal information of the PEC specifies the period with which the encounters (as represented by the graph) repeat in time and how long the pattern repeats for. We note that the same nodes may have encounter patterns at more than just one periodicity (for example, a group of nodes may meet daily during the week, and fortnightly on weekends), and thus the same set of nodes may belong to multiple PECs.

The formal definition of PECs and the language we use to discuss them are presented in detail in Section 5.1.1, along with a formulation of the general PEC detection problem. We present the local-knowledge variant of the PEC detection problem in Section 5.1.2. It is this local-knowledge PEC detection problem that we must solve in the context of opportunistic networks, since the limited connectivity and decentralised nature of these networks make it unfeasible to maintain a single source of complete knowledge of the network. Furthermore, it would be very inefficient to have nodes flood their whole (unprocessed) local encounter histories through the network to emulate a global knowledge scenario.

In Section 5.1.3 we show that the global PECs (the result of the general PEC detection problem) can be decomposed into multiple locally detectable PECs, and thus there is a viable solution to the local-knowledge PEC detection problem. The relationship between PEC detection and the existing problem of periodic subgraph mining is discussed in Section 5.1.4, along with reasons why periodic subgraph mining is not directly applicable to the local-knowledge problem.

5.1.1 General PEC detection formulation

PEC detection and periodic subgraph mining [LB09] are closely related and we adopt consistent terminology in our formulation. The representation of time in our formulation is as a series of discrete *timesteps*. The duration Q that each timestep spans is referred to as the *granularity*. In particular, for some arbitrary start time c, a timestep t spans the interval $[\, c + (t-1)Q,\ c + tQ\,)$.

Definition 5.1

A **simple encounter graph** $G_t = (V_t, E_t)$ is a snapshot of all encounters and nodes appearing within the time window corresponding to timestep t. That is, $\{v, u\} \in E_t$ if and only if nodes $v \in V_t$ and $u \in V_t$, where $v \neq u$, were in range at least once during the time interval represented by t. A simple encounter graph $G_t = (V_t, E_t)$ is **proper** if and only if $\forall v \in V_t$ there exists $u \in V_t$ such that $\{v, u\} \in E_t$; in other words, a proper simple encounter graph is one where every node is involved in at least one encounter.

Definition 5.2

A **dynamic encounter graph** $\mathcal{D} = \langle G_1, \ldots, G_T \rangle$ is a time-ordered sequence of proper simple encounter graphs.

We note that our definition of dynamic encounter graph implies that a node exists in V_t if and only if it is involved in an encounter with another node at timestep t. The definition of dynamic graph in [LB09] is less strict as it permits nodes with degree zero.

The encounter events between individuals v and u in an encounter chronology $\mathcal{S}_{v,u} = \{x_i \,|\, i = 1, \ldots, L\}$ (Definition 3.1) that covers an overall duration T_{max} have a corresponding representation in a dynamic encounter graph that consists of T_{max}/Q timesteps. We assume T_{max} and Q are chosen such that Q divides T_{max} and that $\forall\, x \in \mathcal{S}_{v,u}$, $x < T_{max}$. A given encounter event at time $x \in \mathcal{S}_{v,u}$ is represented in the dynamic encounter graph as an edge between v and u in the simple encounter graph at timestep $\lfloor\, x/Q \,\rfloor + 1$.

Definition 5.3

The subgraph $C = (V, E)$ of a proper simple encounter graph G_t is an **encounter community** if C is connected and $|V| > 1$.

We write $F_2 \subseteq F_1$ to denote that F_2 is a subgraph of, or equal to, F_1. We say that an encounter community C *exists* in the dynamic encounter graph $\mathcal{D}$ at timestep t if $C \subseteq G_t$. An encounter community C may exist in $\mathcal{D}$ at periodic timesteps, leading to the following definition.

Definition 5.4

A **periodic support set**, denoted by S_λ where $\lambda = (i, p, n)$, for an encounter community C in a dynamic encounter graph $\mathcal{D} = \langle G_1, \ldots, G_T \rangle$ is a subsequence of $n > 1$ timesteps

$$S_\lambda = \langle i, i + p, i + 2p, \ldots, i + (n - 1)p \rangle$$

for which C exists, where $i \geq 1$ and $i + (n - 1)p \leq T$. The kth timestep specified by a periodic support set, where $1 \leq k \leq n$, is given by $i + (k - 1)p$ and denoted by $S_\lambda(k)$.

Given a periodic support set $S_\lambda, \lambda = (i, p, n)$ for encounter community C, we refer to n as the number of *periodic occurrences* of C specified by S_λ. We write $|S_\lambda|$ to denote periodic support set size, noting that $|S_\lambda| = n$.

Definition 5.5

A periodic support set S_λ, where $\lambda = (i, p, n)$, for an encounter community C in a dynamic encounter graph $\mathcal{D} = \langle G_1, \ldots, G_T \rangle$ is a **maximum periodic support set** if both $C \not\subseteq G_{i-p}$ and $C \not\subseteq G_{i+pn}$.

Definition 5.6

Denoted as the pair $\langle C, S_\lambda \rangle$, a **periodic encounter community** (or **PEC**) is an encounter community C along with a maximum periodic support set S_λ for which C exists.

Note that an encounter community may exist in a dynamic encounter graph for more than one maximum periodic support set. A maximum periodic support set may be wholly contained within, intersect, or be disjoint from another. If a maximum periodic support set is contained within another, the contained periodic support set is redundant and the containing periodic support set subsumes all of the temporal information conveyed by the contained periodic support set.

Definition 5.7

A periodic support set S_{λ_2} is a **subset** of S_{λ_1} if and only if all timesteps contained in S_{λ_2} are contained in S_{λ_1}. Letting $\lambda_2 = (i_2, p_2, n_2)$ and $\lambda_1 = (i_1, p_1, n_1)$, an equivalent definition is that S_{λ_2} is a subset of S_{λ_1} if and only if all of the following conditions hold:

1. $i_2 \geq i_1$ and $i_2 + p_2(n_2 - 1) \leq i_1 + p_1(n_1 - 1)$
   (i.e., S_{λ_2} is temporally bounded by S_{λ_1});
2. $p_2 = kp_1$ for some integer $k > 0$
   (i.e., the period p_1 is a factor of p_2);
3. $i_2 = i_1 + lp_1$ for some integer $0 \leq l < n_1$
   (i.e., the first timestep in S_{λ_2} must be equal to a timestep in S_{λ_1}).

We denote by $S_{\lambda_2} \subseteq S_{\lambda_1}$ the relationship of S_{λ_2} being a subset of S_{λ_1}. We call S_{λ_2} a **proper subset** of S_{λ_1} if and only if $S_{\lambda_2} \subseteq S_{\lambda_1}$ and $S_{\lambda_2} \neq S_{\lambda_1}$. This relation is denoted $S_{\lambda_2} \subset S_{\lambda_1}$.

The definition of the subset relation for periodic support sets formalises the concept of temporal subsumption. If we have an encounter community C which exists in periodic support sets S_{λ_1} and S_{λ_2} such that $S_{\lambda_2} \subset S_{\lambda_1}$, then S_{λ_1} conveys more information than S_{λ_2} about the periodic occurrences of C.

Definition 5.8

An encounter community C' is a **subcommunity** of encounter community C if and only if $C' \subseteq C$.

Subsumption can also occur between the structural components of PECs. For example, given a PEC $\langle C, S_\lambda \rangle$, any subcommunity C' of C also exists for S_λ. If S_λ is maximum for C' then $\langle C', S_\lambda \rangle$ forms a PEC; however, in the case that $C' \subset C$ the PEC $\langle C', S_\lambda \rangle$ contains only some of the structural information conveyed by $\langle C, S_\lambda \rangle$.

Definition 5.9

Let $\mathcal{P}_1 = \langle C_1, S_{\lambda_1} \rangle$ and $\mathcal{P}_2 = \langle C_2, S_{\lambda_2} \rangle$ be two PECs. We say that $\mathcal{P}_1$ is **subsumed** by $\mathcal{P}_2$ if and only if $S_{\lambda_1} \subseteq S_{\lambda_2}$ and $C_1 \subseteq C_2$. We denote this relationship by $\mathcal{P}_1 \subseteq \mathcal{P}_2$.

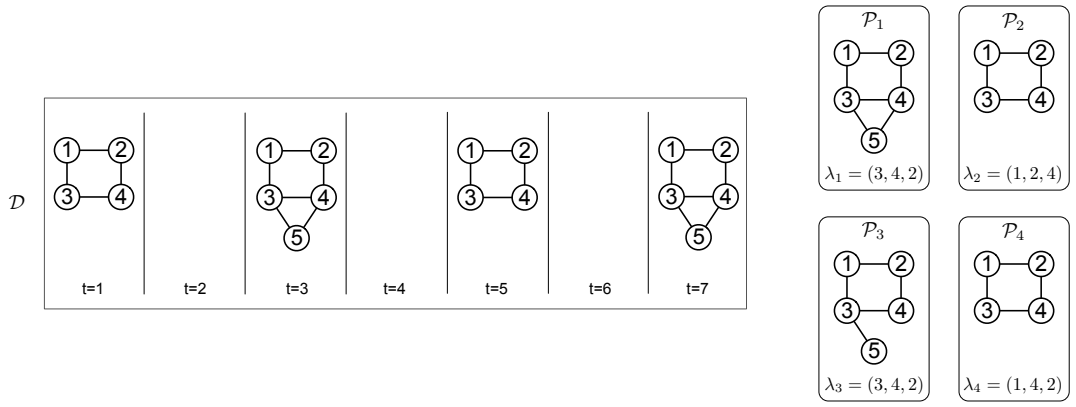


Figure 5.1: A dynamic encounter graph $\mathcal{D}$ and a selection of PECs in $\mathcal{D}$.

Definition 5.10

A PEC $\mathcal{P}_1$ is **maximal** if and only if there does not exist another PEC $\mathcal{P}_2$, where $\mathcal{P}_1 \neq \mathcal{P}_2$, such that $\mathcal{P}_1$ is subsumed by $\mathcal{P}_2$.

Figure 5.1 demonstrates the subsumption and maximality criteria on an example dynamic encounter graph $\mathcal{D}$. PECs $\mathcal{P}_1$ and $\mathcal{P}_2$ are the only maximal PECs in $\mathcal{D}$ because they are each not subsumed by any other PEC. PECs $\mathcal{P}_3$ and $\mathcal{P}_4$ are examples of sub-maximal PECs. In particular, $\mathcal{P}_3$ is structurally subsumed by $\mathcal{P}_1$ due to the lack of edge $\{4, 5\}$ and $\mathcal{P}_4$ is temporally subsumed by $\mathcal{P}_2$ because the period of $\mathcal{P}_2$ divides that of $\mathcal{P}_4$ causing S_{λ_4} to be a subset of S_{λ_2}.

Maximal PECs are the fundamental PECs that we wish to extract from a dynamic encounter graph. A node may appear in multiple maximal PECs, which reflects the real-world property that an individual may belong to more than one community. Communities can form due different types relationship, such as work, social, and familial. Our extension of static communities into the temporal domain allows us to capture the pattern with which the community reoccurs in addition to the connections within it. An implication of our definition of PEC is that a node's membership in a community is a binary property. This strict definition is a necessary trade-off to make the automatic detection of community-specific periodicities tractable. The data mining algorithm we use to extract communities and their periods is introduced later in this chapter.

With knowledge of all maximal PECs, all other PECs are redundant. The collection of all maximal PECs represents the most compact and complete description of the

periodic encounter communities present in a dynamic encounter graph.

Definition 5.11

The **periodic encounter community detection problem** is the problem of finding all maximal periodic encounter communities that exist in a dynamic encounter graph.

5.1.2 Local-knowledge PEC detection formulation

The problem as introduced in Section 5.1.1 is presented as a *global-knowledge* problem, where mining of PECs could be carried out with the full graphs in the dynamic encounter graph available to a mining algorithm, as is the case with the PSE-Miner in [LB09]. Alternative to this is the node-centric perspective where the entire graph G_t is not available to any single entity. In particular, each node has only the knowledge of encounters that directly involve it. We formalise the concept of local knowledge in the following definitions.

Definition 5.12

For an encounter graph $G_t = (V_t, E_t)$, the **intrinsic encounter graph** $G_t^v = (V_t^v, E_t^v)$ is the subgraph of G_t induced by selecting only the edges $E_t^v = \{e \,|\, e \in E_t \wedge v \in e\}$ and their incident vertices.

Definition 5.13

Consider the dynamic encounter graph $\mathcal{D} = \langle G_1, \ldots, G_T \rangle$. The **intrinsic dynamic encounter graph** of a node v is the sequence of graphs $\mathcal{D}_v = \langle G_1^v, \ldots, G_T^v \rangle$.

Figure 5.2 shows a set of intrinsic dynamic encounter graphs and the corresponding global dynamic encounter graph. A node v's intrinsic dynamic encounter graph represents the encounter information that is directly observable by v. We note that the global encounter graph at timestep t is the aggregation of all intrinsic graphs at t; in other words, if we have dynamic encounter graph $\mathcal{D} = \langle G_1, \ldots, G_T \rangle$ and denote the set of all nodes by $\mathcal{V} = V_1 \cup \ldots \cup V_T$, then

$$G_t = \bigcup_{v \in \mathcal{V}} G_t^v \ .$$

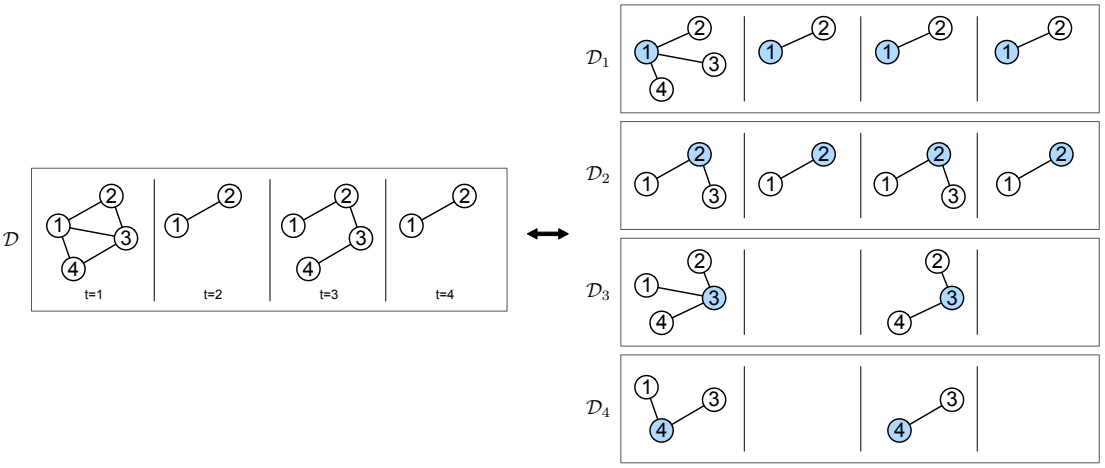


Figure 5.2: A global dynamic encounter graph $\mathcal{D}$ and its intrinsic dynamic encounter graphs. $\mathcal{D}_v$ denotes the intrinsic dynamic encounter graph for node v.

Knowledge of the (global) dynamic encounter graph is effectively distributed among the nodes in the network.

We distinguish PECs that are maximal in the global dynamic encounter graph by referring to them as *globally maximal PECs*. An *intrinsic PEC* is a PEC (be it maximal or submaximal) that exists in an intrinsic dynamic encounter graph.

Definition 5.14

Local-knowledge periodic encounter community detection is the problem of identifying all globally maximal periodic encounter communities from local knowledge. This is a special case of the periodic encounter community detection problem (Definition 5.11) where no global view of the dynamic encounter graph exists. In particular, the following restrictions apply:

- **Local knowledge:** knowledge of encounters is expressed only as intrinsic dynamic encounter graphs, all of which are distributed among the corresponding nodes in the network;
- **Local exchange:** information may be exchanged between a pair of nodes only when they encounter each other.

The decentralised detection scenario corresponds to the local-knowledge problem.

5.1.3 Decomposition of PECs

Here we show that all globally maximal PECs decompose into intrinsic PECs. This is an important property as it means that if individual nodes extract their intrinsic PECs from their intrinsic dynamic encounter graphs, they can combine these intrinsic PECs with those of other nodes to find globally maximal PECs. Therefore, all globally maximal PECs can be detected in the local-knowledge problem.

Definition 5.15

A set of encounter communities $\Gamma = \{C_1, C_2, \ldots, C_m\}$ is a **community cover** of encounter community C if

$$\bigcup_{C' \in \Gamma} C' = C \ .$$

Consider the PEC $\langle C, S_\lambda \rangle$ in dynamic encounter graph $\mathcal{D}$ and a community cover Γ of C. From the definition of a PEC (Definition 5.6) and the subgraph property of a subcommunity, it follows that any encounter community C' in Γ exists for periodic support set S_λ. Although S_λ may not be maximum for the subcommunity C', there must exist a maximum periodic support set $S_{\lambda'}$ for C' that contains S_λ, and therefore there exists a PEC $\langle C', S_{\lambda'} \rangle$.

Definition 5.16

The **intrinsic cover** of encounter community $C = (V, E)$ is the set of communities $\{C_v \,|\, v \in V\}$, where $C_v = (V_v, E_v)$ is the subcommunity of C induced by selecting only the edges $E_v = \{e \,|\, e \in E \wedge v \in e\}$ and their incident vertices.

A subcommunity C' in the intrinsic cover of C corresponds to a particular node's intrinsic (i.e., local) view of C. We note that the intrinsic cover of C is also a community cover of C.

Consider a PEC $\mathcal{P} = \langle C, S_\lambda \rangle$ and a subcommunity C' in the intrinsic cover of C. It follows that there must be an intrinsic PEC that subsumes $\langle C', S_\lambda \rangle$. Therefore, $\mathcal{P}$ decomposes into multiple intrinsic PECs. The same applies if $\mathcal{P}$ is a globally maximal PEC, and so any globally maximal PEC can be reconstructed from a local-knowledge representation.

5.1.4 Relation to periodic subgraph mining

The periodic subgraph mining problem introduced by Lahiri and Berger-Wolf in [LB09] is related to the PEC detection problem that we present in this chapter. Rather than extracting periodic encounter communities as we do in our work, the periodic subgraph mining problem seeks to extract *periodic subgraph embeddings* (PSEs). A PSE in a dynamic encounter graph $\mathcal{D}$ is defined as a pair $\langle F, S_\lambda \rangle$ where F is a subgraph that exists in $\mathcal{D}$ for the periodic support set S_λ. Subsumption and maximality rules apply to PSEs as they do to PECs. The key distinction between a PEC and a PSE is the encounter community property of PECs. In particular, the definition of a PSE is more general as it allows subgraphs that are disconnected and subgraphs consisting of only one node.

If we assume global knowledge of the dynamic encounter graph, the PEC detection problem becomes a special case of the PSE mining problem. By extracting connected subgraphs consisting of at least two nodes from the graphs of maximal PSEs in a dynamic encounter graph, we obtain the maximal PECs. Lahiri and Berger-Wolf also show that the time and space complexity of the problem is polynomial in the size of the input dynamic encounter graph. The PSE-Miner algorithm presented as a solution to the PSE mining problem requires global knowledge, making it unsuitable for directly extracting all maximal PECs in the local-knowledge PEC detection problem. For the local-knowledge problem we instead follow a local mining and local sharing approach.

5.2 Decentralised PEC detection algorithm

In this section we describe our decentralised PEC detection algorithm. From the decomposition in Section 5.1.3 we know that global maximality of PECs can be reached from a local-knowledge representation. Therefore, the aim of the detection algorithm is to build globally maximal PECs from the local-knowledge distributed across all the nodes in the system.

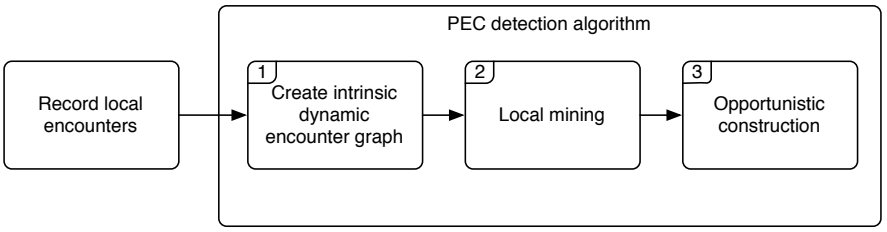


Figure 5.3: An overview of the stages of the PEC detection algorithm from the perspective of a node.

5.2.1 Algorithm overview and parameters

Figure 5.3 provides an overview of the stages that a node goes through during the operation of the detection algorithm. Here we provide a brief introduction to the detection algorithm. The individual stages are described in detail later in this section. Note that the task of a node finding its local periodic communities and the periods with which these communities repeat (i.e., the task of extracting local PECs) is carried out in Stage 2. These local PECs are subsequently combined with the local PECs found by other nodes in Stage 3.

Three parameters are required during the detection algorithm. In Stage 1 the granularity Q is used. In Stage 2 p_{max} (the maximum PEC period) and n_{min} (the minimum number of periodic occurrences) are used.

Before the detection algorithm is initiated, all nodes record the times of their encounters as *encounter chronologies* (this corresponds to the initial state in Figure 5.3). On initiation, the first stage of the detection algorithm is for each node to build its intrinsic dynamic encounter graph (Definition 5.13) from its encounter chronology. It is this stage where the granularity parameter (denoted by Q) is applied. As described in Section 5.1.1, the granularity Q is the duration of each timestep in the node's intrinsic dynamic encounter graph. The choice of granularity Q depends on the domain and application. Choosing a fine granularity results in more timesteps in the intrinsic dynamic encounter graph and so increases the computational overhead of the mining algorithm (which occurs in Stage 2). Fine granularities also have the disadvantage that the effect of small-scale randomness in the times of encounters is greater. However, in some cases we may still wish to use a fine granularity for the purpose of identifying repeating behaviour with a fine degree of temporal resolution (e.g., identifying regular

encounters to within a specific hour of the day).

Details of the initiation of the PEC detection algorithm (including the building of the intrinsic dynamic encounter graph in Stage 1) and its data structures are given in Section 5.2.2. Note that subsequent stages of the detection algorithm only consider time in terms of timesteps in the intrinsic dynamic encounter graph. Resulting PECs are described in terms of timestep indexes rather than real-time units. (For example, the period of a PEC $\langle C, S_\lambda \rangle, \lambda = (i, p, n)$ is p timesteps.) It is trivial to convert from timesteps back to real-time units.

The node's intrinsic dynamic encounter graph built in Stage 1 is used as input to the *local mining* stage (Stage 2), which is detailed in Section 5.2.3. In brief, during the local mining stage individual nodes mine their intrinsic dynamic encounter graphs to obtain their intrinsic PECs. Each node implements the PSE-Miner algorithm (detailed in [LB09]) which extracts all (maximal intrinsic) PECs found in the node's intrinsic dynamic encounter graph. The correctness of the PSE-Miner is shown in [LB09] and thus we know that all PECs present in the node's intrinsic dynamic encounter graph will be identified (the criteria that define a PEC are given in Definition 5.6). All the attributes that constitute each PEC are automatically found by the PSE-Miner. For a PEC $\mathcal{P} = \langle C, S_\lambda \rangle, \lambda = (i, p, n)$ these attributes are the community C, the start timestep i, the period p, and the number of periodic occurrences n. Importantly, it is the PSE-Miner that identifies the one or more periods that a community repeats with in the node's intrinsic dynamic encounter graph, resulting in one or more PECs for the community. The two parameters, $p_{max} \geq 1$ and $n_{min} \geq 2$, specified in this stage control the maximum period and minimum number of periodic occurrences, respectively. Formally, only PECs that meet the conditions $p \leq p_{max}$ and $n \geq n_{min}$ are identified. Although PECs with larger periods may exist in the intrinsic dynamic encounter graph, these are ignored.

The intrinsic dynamic encounter graph is only a local subset of the global dynamic encounter graph, and so the PECs resulting from the local mining stage (Stage 2) are not necessarily globally maximal. It is through knowledge exchange during the *opportunistic construction* stage (Stage 3) (detailed in Section 5.2.4) that nodes learn the globally maximal PECs they belong to. Encounters between pairs of nodes offer the

opportunity for those nodes to share and expand the PECs they have discovered so far. The process of combining PECs results in PECs that have a larger community, and possibly a new period derived from the source PECs. Note that nodes will only seek to learn the PECs they are a member of.

5.2.2 Algorithm setup and initiation

We denote by $\mathcal{V}$ the set of all nodes in dynamic encounter graph $\mathcal{D}$. Each node $v \in \mathcal{V}$ maintains its local history of encounters with other nodes. When the detection algorithm is initiated, each node $v \in \mathcal{V}$ first builds its intrinsic dynamic encounter graph $\mathcal{D}_v = \langle G_1^v, \ldots, G_T^v \rangle$ from its encounter chronology. Building $\mathcal{D}_v$ is done by segmenting time into T timesteps, where each timestep represents a duration of time Q. Given some arbitrary start time c, the encounter graph G_t^v at timestep t represents v's encounters in the time interval $\big[\ c + (t-1)Q,\ c + tQ\ \big)$. The granularity Q is only used for the purpose of building the intrinsic dynamic encounter graph and is not used at any future point in the algorithm.

Each node also maintains a *knowledge base* (Definition 5.17), which is a data structure that holds the PECs discovered by a node so far.

Definition 5.17

> The **knowledge base** for a node v, denoted by K_v, is a set that consists of the PECs known by v.

Knowledge bases are updated over time as locally stored PECs are combined with PECs received from other nodes. During each update, the algorithm ensures that a knowledge base K_v meets the following conditions:

1. Relevance to v: $\forall \langle C, S_\lambda \rangle \in K_v$ node v is a member of encounter community C.
2. Maximality among K_v: $\forall\, \mathcal{P}_1 \neq \mathcal{P}_2 \in K_v,\ \mathcal{P}_1$ does not subsume $\mathcal{P}_2$.

By Condition 1, a node only stores PECs that are relevant to it, and Condition 2 ensures that no redundant PECs are stored.

Once the intrinsic dynamic encounter graph $\mathcal{D}_v$ for a node v is formed and the know-

ledge base K_v is initialised, v then mines the intrinsic PECs from $\mathcal{D}_v$ and places them in K_v. This mining algorithm is detailed in Section 5.2.3. From timestep $T + 1$ onwards, nodes share and update PEC information whenever they encounter each other, as detailed in Section 5.2.4.

The point in time to initiate mining depends on domain and application. Most applications would benefit from obtaining PEC information early; however, mining too early may result in there being too few timesteps for periodic patterns to be present.

5.2.3 Local mining: extraction of intrinsic PECs

In the decentralised PEC detection algorithm each node v executes the PSE-Miner algorithm [LB09] on its intrinsic dynamic encounter graph $\mathcal{D}_v$ to extract its (locally maximal) intrinsic PECs. As mentioned in Section 5.1.4, the PSE-Miner algorithm is capable of extracting all maximal PECs in a dynamic encounter graph. Therefore, if a node implements PSE-Miner, it can extract its maximal intrinsic PECs from its intrinsic dynamic encounter graph. Note that period detection is part of the mining process itself, and therefore periods do not need to be specified beforehand.

For the purpose of the PSE-Miner algorithm a dynamic encounter graph is represented as a sequence of sets of integers. To establish an invertible mapping between graphs and sets, all nodes and all edges in a dynamic encounter graph are each mapped to a unique integer label. The set representation for a particular graph $G_t = (V_t, E_t)$ is the set A_t of size $|V_t| + |E_t|$ where the integer label of each element in $V_t \cup E_t$ appears in A_t. This set representation allows fundamental operations such as graph hashing and maximal common subgraph finding to be carried out efficiently by the PSE-Miner [LB09].

The PSE-Miner is a single-pass algorithm. During execution the miner maintains two core data structures: a *pattern tree* and a *subgraph hash map*. As soon as a PSE ceases to be periodic it is flushed to the output stream. Those PSEs that do not have a sufficient number of periodic occurrences (n_{min}) are filtered out. A full description of the operation of the PSE-Miner algorithm, including how subgraphs and their periods are automatically identified, is provided in [LB09].

After a node v executes PSE-Miner on its intrinsic dynamic encounter graph, the node discards any PSEs that consist only of v (these are valid PSEs but not valid PECs). All other PSEs extracted by PSE-Miner are (locally maximal) intrinsic PECs and are therefore added to v's knowledge base.

5.2.4 Opportunistic construction

Opportunistic construction is the process whereby pairs of nodes share and combine their locally stored PECs when in communication range. Through repeated opportunistic construction, nodes obtain more information on the structure of the globally maximal PECs they belong to. As mentioned in Section 5.1.3, any non-intrinsic PEC can be obtained from its intrinsic PECs. Thus, if a construction strategy is correct and there are sufficient exchange opportunities, nodes will eventually obtain their globally maximal PECs.

When a node v encounters a node u, it receives knowledge base K_u. It is the task of v to update its own knowledge base K_v by pairwise combining the PECs in K_v with those in K_u. This update mechanism is described in Section 5.2.4.2. As part of knowledge base updating, node v must check if a pair of PECs are compatible to be combined to derive a new PEC. Compatibility and combination are explained in Section 5.2.4.1.

Note that although the local mining step returns the intrinsic PECs for a node, over time these may be subsumed by PECs generated during opportunistic construction. An intrinsic PEC subsumed by another PEC is removed since the subsuming PEC contains all the information conveyed by the intrinsic PEC. This reduces the size of knowledge bases without affecting the ability of the algorithm to build globally maximal PECs.

5.2.4.1 PEC compatibility and combination

Upon node v receiving a PEC $\mathcal{P}$ from another node, v must check which PECs in its knowledge base K_v can be combined with $\mathcal{P}$ to derive new PECs. A derived PEC must be connected, exist in its periodic support set, and be relevant to v.

Definition 5.18

Two PECs $\langle C_1, S_{\lambda_1} \rangle$ and $\langle C_2, S_{\lambda_2} \rangle$ with encounter communities $C_1 = (V_1, E_1)$ and

$C_2 = (V_2, E_2)$ are **compatible** for node v if all of the following hold:

1. Relevance to v: $v \in V_1$ and $v \in V_2$;
2. Structural overlap: $E_1 \cap E_2 \neq \emptyset$;
3. Temporal containment: either $S_{\lambda_1} = S_{\lambda_2}$, $S_{\lambda_1} \subset S_{\lambda_2}$, or $S_{\lambda_2} \subset S_{\lambda_1}$.

The method of combination for two compatible PECs $\langle C_1, S_{\lambda_1} \rangle$ and $\langle C_2, S_{\lambda_2} \rangle$ depends on the direction of periodic support set containment. The case $S_{\lambda_1} = S_{\lambda_2}$ is a simple case because both C_1 and C_2 exist for the same periodic support set. In the case $S_{\lambda_1} \subset S_{\lambda_2}$, we know that C_2 exists for S_{λ_1}, but C_1 does not exist for all timesteps in S_{λ_2}. Therefore, when combining two PECs where $S_{\lambda_1} \subset S_{\lambda_2}$, the contained periodic support set (i.e., S_{λ_1}) is chosen to ensure that the resulting community exists in its support.

Formally, two compatible PECs $\langle C_1, S_{\lambda_1} \rangle$ and $\langle C_2, S_{\lambda_2} \rangle$ are combined to derive PEC $\langle C', S_{\lambda'} \rangle$ as follows:

- if $S_{\lambda_1} = S_{\lambda_2}$ then $C' = C_1 \cup C_2$ and $S_{\lambda'} = S_{\lambda_1} = S_{\lambda_2}$;
- if $S_{\lambda_1} \subset S_{\lambda_2}$ then $C' = C_1 \cup C_2$ and $S_{\lambda'} = S_{\lambda_1}$;
- if $S_{\lambda_2} \subset S_{\lambda_1}$ then $C' = C_1 \cup C_2$ and $S_{\lambda'} = S_{\lambda_2}$.

5.2.4.2 Knowledge base updating

During an encounter between two nodes v and u their knowledge bases K_v and K_u are exchanged. Although we assume for simplicity that whole knowledge bases are exchanged, in practice a sender node can identify PECs in its knowledge base that will not be relevant to the recipient node. Withholding these PECs reduces communication overhead without affecting PEC construction. Examples include withholding a PEC that the recipient node is not a member of and withholding a PEC that has already been sent to the same node during a previous encounter. Even without filtering, the size of a knowledge base is typically small, which minimises the storage and bandwidth requirements of the opportunistic construction stage. PECs themselves have a compact representation, consisting of a graph and three integers to describe the periodic support set, and the algorithm only maintains locally maximal PECs, therefore avoiding storage

and transfer of redundant data.

Algorithm 5.1: KB-Update

Input: Node v whose knowledge base K_v is to be updated
Input: External knowledge base K_u
Create empty list L to hold candidate PECs
Generate candidates:
foreach $\mathcal{P}_a \in K_v$ and $\mathcal{P}_b \in K_u$ **do**
  if $\mathcal{P}_a$ and $\mathcal{P}_b$ are *compatible* for v and $\mathcal{P}_b \not\subseteq \mathcal{P}_a$ **then**
    Combine $\mathcal{P}_a$ and $\mathcal{P}_b$ to generate candidate PEC $\mathcal{P}_c$
    Add $\mathcal{P}_c$ to L
  end
end
Prune candidates list:
foreach $\mathcal{P}_c \in L$ and $\mathcal{P}_a \in K_v$ **do**
  if $\mathcal{P}_c \subseteq \mathcal{P}_a$ **then**
    Remove $\mathcal{P}_c$ from L
  end
end
Prune knowledge base:
foreach $\mathcal{P}_a \in K_v$ and $\mathcal{P}_c \in L$ **do**
  if $\mathcal{P}_a \subseteq \mathcal{P}_c$ **then**
    Remove $\mathcal{P}_a$ from K_v
  end
end
Insert candidates:
Add all PECs in L to K_v

For a node v receiving knowledge base K_u from node u, node v updates its own knowledge base K_v according to Algorithm 5.1. Candidate pruning is carried out to ensure that redundant PECs are not added to the knowledge base K_v. A candidate that passes the pruning step is one that is not subsumed by any PEC already in K_v and should therefore be added to K_v. Such candidates may subsume a number of PECs already in the knowledge base. To ensure maximality among PECs in K_v, knowledge base pruning is carried out to remove any pre-existing PECs made redundant by the addition of the candidate.

5.3 Analysis of PEC construction

The time required for a global PEC to be discovered by all its constituent nodes is of primary interest for the analysis of PEC construction. It is the encounters between individual nodes that enable the information of a PEC to be shared throughout the network, and thus the patterns of these encounters have a substantial impact on the time required for a node to discover the globally maximal PECs it belongs to.

To study the spread of information in the construction of global PECs we define the equivalent scenario of *token broadcast*. Informally, token broadcast is where each node of a PEC being studied attempts to flood a unique token to all other nodes in the PEC. The route taken by a token from node u to reach node v represents the spread of u's local PEC information to v during the opportunistic construction phase of the decentralised PEC detection algorithm. The event of the token sent from u reaching v corresponds to the event of v receiving a knowledge base including some of u's PECs for the first time. The token can also represent a general packet of information, and thus the token broadcast scenario provides insight into the flow of information within a PEC.

We formally define the token broadcast scenario as follows. Consider the (global) dynamic encounter graph $\mathcal{D} = \langle G_1, G_2, \ldots, G_T \rangle$ and an arbitrary PEC $\langle C, S_\lambda \rangle$ in $\mathcal{D}$ where $\lambda = (i, p, n)$ and $C = (V, E)$. Each node v in C stores a token set T_v of received copies of tokens. We denote node v's token set after t timesteps by $T_v(t)$. Initially, each token set T_v only consists of the single token τ_v. In other words, $\forall\, v \in V,\ T_v(0) = \{\tau_v\}$. Token sharing then progresses for each timestep $i, i+1, i+2, ..., i+(n-1)p$. To carry out token sharing at timestep t, all of the encounters in the time interval for t are applied in the order they occurred. When two nodes encounter each other, each copies all of its tokens to the other. We say that full coverage has been reached in timestep t if all nodes in V have received all tokens; that is, every node v in V has $T_v(t) = \{\tau_v \mid v \in V\}$.

To characterise the broadcast of a specific PEC, only those encounters that support the PEC are used as token sharing opportunities. More specifically, only encounters corresponding to edges in E during timesteps in S_λ are used as token sharing oppor-

tunities.

5.3.1 Token broadcast metrics

We define the following metrics for evaluating broadcast within a PEC.

First, to quantify the extent of token spread over time we introduce metrics for token coverage. The **coverage fraction** $f_c(v,t)$ for a node v in encounter community graph $C = (V, E)$ at the end of timestep t is given by

$$f_c(v,t) = \frac{|T_v(t)| - 1}{|V| - 1} \ .$$

This measures the relative number of tokens v has obtained by the end of timestep t, excepting its own token τ_v. For a PEC $\mathcal{P} = \langle C, S_\lambda \rangle$ where $\lambda = (i, p, n)$, we quantify the **PEC coverage** $\bar{f}_c(\mathcal{P}, t)$ as the average coverage of nodes in C at timestep t,

$$\bar{f}_c(\mathcal{P}, t) = \frac{1}{|V|} \sum_{v \in V} f_c(v,t) \ .$$

It is more convenient to talk in terms of the number of periodic occurrences of a PEC rather than the number of timesteps. The timestep for the kth periodic occurrence of $\mathcal{P}$ is given by $S_\lambda(k)$ and so we refer to $\bar{f}_c(\mathcal{P}, S_\lambda(k))$ for the coverage fraction after k periodic occurrences.

The **broadcast time**, denoted $\Lambda(\mathcal{P})$, measures the number of periodic occurrences of a PEC $\mathcal{P}$ that were required for $\mathcal{P}$ to reach full coverage. $\Lambda(\mathcal{P})$ is equal to the smallest positive integer k such that $\bar{f}_c(\mathcal{P}, S_\lambda(k)) = 1$. In the case that there were insufficient encounters to reach full coverage, $\Lambda(\mathcal{P}) = \infty$.

5.3.2 Worst-case token broadcast time

The worst-case token broadcast time, denoted by $\Lambda_{max}(\mathcal{P})$, is the theoretical maximum number of periodic occurrences that an arbitrary PEC $\mathcal{P}$ requires to reach full coverage, under the assumption that $\mathcal{P}$ continues recurring indefinitely. Knowledge of the existence of a PEC $\mathcal{P} = \langle C, S_\lambda \rangle$, where $C = (V, E)$ and $\lambda = (i, p, n)$, in dynamic encounter graph $\mathcal{D} = \langle G_1, \ldots, G_T \rangle$ implies some minimum conditions on the occurrences of encounters in $\mathcal{D}$; in particular, for each edge $\{v, u\}$ in E, there must be

at least one encounter between nodes v and u in each timestep $S_\lambda(1), S_\lambda(2), ..., S_\lambda(n)$. The worst-case analysis of the broadcast time for $\mathcal{P}$ considers the largest possible number of periodic occurrences that $\mathcal{P}$ would require to reach full coverage.

We note that if a PEC repeats indefinitely, the worst-case broadcast time is always finite. Since the encounters for each edge in E must occur at least once in each timestep in S_λ, if a token τ_u has not reached every node at the end of timestep $S_\lambda(k)$ then it will spread to at least one additional node in timestep $S_\lambda(k+1)$.

Definition 5.19

> The **broadcast front**, denoted by $B_v(t)$, of token τ_v at timestep t is the set of nodes that received τ_v in timestep t but did not have it in timestep $t-1$.

A worst case for the travel of token τ_u from node u to node v is presented as follows. Consider the case where, in timestep $S_\lambda(1)$, the encounters corresponding to edges incident to u occur after all other encounters in that timestep. The effect of this is that τ_u moves one node closer to v along the shortest paths between u and v, and the broadcast front $B_u(S_\lambda(1))$ consists only of u's neighbours. If in timestep $S_\lambda(2)$ the encounters corresponding to edges incident to the nodes in $B_u(S_\lambda(1))$ occur after all other encounters, τ_u will again only move one node closer to v. If the encounters corresponding to edges incident to nodes in $B_u(S_\lambda(k))$ are always the last to occur in each timestep $S_\lambda(k+1), k = 1, \ldots, |S_\lambda|$, then the number of periodic occurrences of C required for τ_u to reach v from u is equal to the shortest path distance between v and u.

A worst-case time for a PEC to reach full coverage results when v and u are peripheral nodes, requiring a number of periodic occurrences equal to the diameter of C, denoted by $d(C)$. Thus, for a PEC $\mathcal{P} = \langle C, S_\lambda \rangle$ we have $\Lambda_{max}(\mathcal{P}) = d(C)$.

5.4 Experiments and results

In this section we evaluate decentralised PEC detection through the study of token broadcast in PECs found in a real-world encounter network. In particular, we use the REALITY encounter trace. The long duration (nine months), frequent sampling (5

minutes), and direct (rather than inferred) nature of this dataset make it the best option for detecting both short-duration and long-duration PECs. However, we note that Bluetooth sampling is unreliable, resulting in some missed encounters. For PEC detection, a missed encounter may result in the true PEC being temporally or structurally partitioned.

5.4.1 Simulating token broadcast

Simulating token broadcast on the encounter trace follows from the framework established in Section 5.3. When extracting the dynamic encounter graph $\mathcal{D} = \langle G_1, \ldots, G_T \rangle$ with granularity Q from the encounter trace, the orderings of actual encounters (including any repeat encounters) within each timestep $1, 2, \ldots, T$ are retained for the purpose of simulating token exchange. To simulate token broadcast for a particular globally maximal PEC $\langle C, S_\lambda \rangle$ with $C = (V, E)$ and $\lambda = (i, p, n)$, the trace is filtered so that only the encounters corresponding to edges in E and occurring during timesteps in S_λ are retained. Unique tokens are placed on the nodes and then broadcast is simulated for each timestep $i, i+1, i+2, \ldots, i+(n-1)p$. In a timestep t, each encounter from the underlying encounter trace is used as a token sharing opportunity in the order it appears during t.

5.4.2 Experimental setup

We set the maximum period parameter (p_{max}) to be 30 days and the minimum periodic occurrences[1] parameter (n_{min}) to be four. Other PSE-Miner parameters were left as the defaults specified in [LB09]; i.e., the minimum period was set to one and no timestep smoothing was carried out.

Experiments were run with granularities (denoted by Q) of 4, 6, 12, and 24 hours. Choosing a fine granularity allows the identification of periodic behaviour with greater temporal precision, but at the cost of an increase in computational overhead. Furthermore, at very fine granularities the effect of small-scale randomness in human en-

[1] In [LB09] the minimum number of periodic occurrences is denoted by σ rather than n_{min}.

counter times becomes great, typically resulting in fewer PECs. Indeed, in experiments with granularity $Q = 1$ hour we found that very few PECs had periods longer than one day. The majority of PECs at this granularity were short-lived communities that repeated in consecutive timesteps for part of a day.

We note that the combination of noise in the trace dataset, the uncertain nature of human behaviour, and the crispness of our PEC definition means that PECs can become temporally fragmented. A break in encounter regularity in an encounter trace, be it due to inadequate sampling or true individual behaviour, results in a PEC either becoming temporally partitioned, structurally smaller, or not existing at all. Two or more PECs having the same encounter community, period, and phase, but spanning different durations in the trace, are assumed to be the same PEC and such duplicates were discarded from the experiments.

Finally, PECs whose communities had a diameter equal to one were not included in the analysis as these are a trivial case for PEC construction.

5.4.3 Results

Information on PECs obtained in the dataset is summarised in Table 5.1. The table shows that average diameter and average periodic support set size increase at coarser granularities. This is due to encounters being aggregated into wider snapshots, resulting in some encounter communities becoming merged. We note that in all experiments every PEC reached full coverage within the duration of time it existed.

Figure 5.4 shows the period and diameter of each PEC detected for granularities of 6 hours and 24 hours. We can clearly observe periodicities at one day, seven days, and 14 days, demonstrating the multiscale characteristic of human encounter behaviour. The figure also shows that many of the PECs occur at periods of one day and seven days. We suggest that this occurs because many of the PECs are students visiting the same campus on each weekday, resulting in one day PECs between Monday and Friday. Although these PECs end at the weekend, students following this weekday behaviour also exist in PECs at a period of seven days.

In Figure 5.5 we plotted the cumulative distribution of the normalised broadcast times

	Granularity (Q)			
	4hr	**6hr**	**12hr**	**24hr**
Number of maximal PECs	509	561	897	900
Average $d(C)$	2.17	2.21	2.34	2.42
Average $\|S_\lambda\|$	4.20	4.24	4.34	4.37
Average $\Lambda(\mathcal{P})$	1.32	1.33	1.51	1.54
Average $\Lambda(\mathcal{P})/\Lambda_{max}(\mathcal{P})$	0.298	0.306	0.449	0.465

Table 5.1: Summary of PECs in the REALITY dataset. Four experiments were run, each with a different granularity (denoted by Q). $d(C)$ denotes community diameter, $|S_\lambda|$ denotes periodic support set size (i.e., total number of periodic occurrences of a PEC), $\Lambda(\mathcal{P})$ denotes broadcast time (measured in number of periodic occurrences), and $\Lambda(\mathcal{P})/\Lambda_{max}(\mathcal{P})$ gives the normalised broadcast time. PECs with $d(C) = 1$ are not included in the experiments.

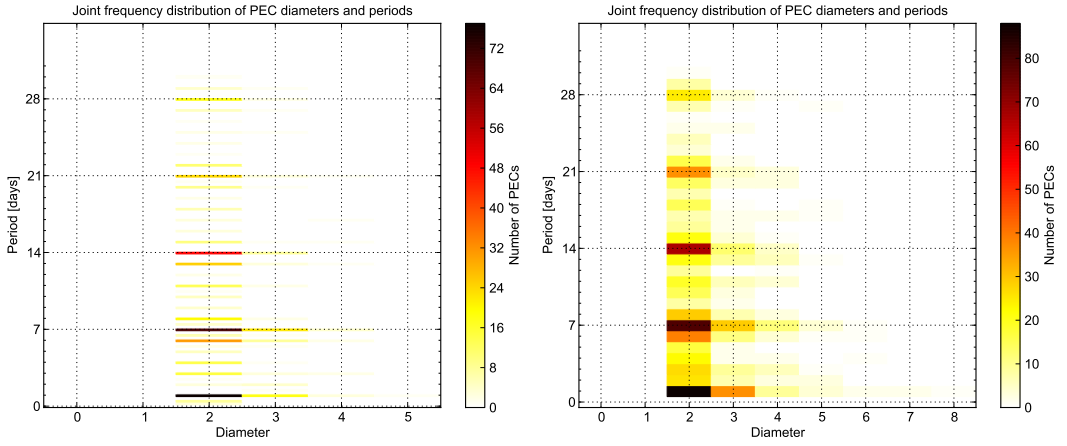


Figure 5.4: Joint frequency distribution of diameters and periods for PECs in the REALITY dataset. Left: $Q = 6$ hrs. Right: $Q = 24$ hrs.

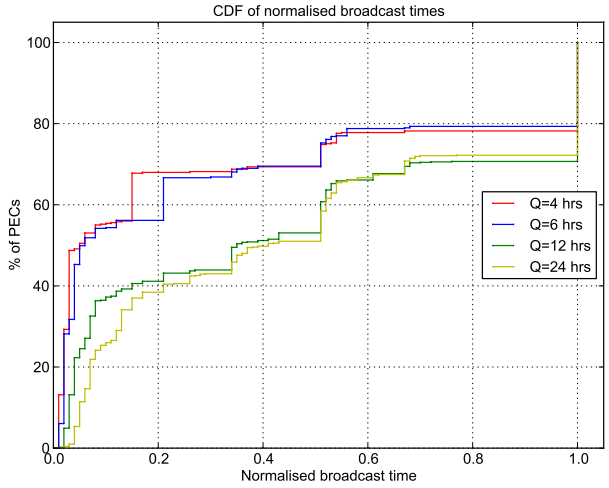


Figure 5.5: Cumulative distribution of normalised broadcast times for PECs in the REALITY dataset. The normalised broadcast time for a PEC $\mathcal{P}$ is given by $\Lambda(\mathcal{P})/\Lambda_{max}(\mathcal{P})$.

of PECs. The normalised broadcast time of a PEC $\mathcal{P}$ is its actual broadcast time $\Lambda(\mathcal{P})$, normalised by its potential worst-case time $\Lambda_{max}(\mathcal{P})$. This quantity indicates how close a PEC's actual broadcast time is to its worst case. For granularities of 4 hours and 6 hours, 68% of PECs reached full coverage in less than 0.22 of their potential worst-case times, and 78% of PECs reached full coverage in less than 0.55 of their potential worst-case broadcast times. For the same granularities, 21% of the PECs required worst-case broadcast time.

Figure 5.5 also shows that coarser granularities result in PECs with broadcast times closer to their worst cases. This is reflected in the plot of community coverage over time (Figure 5.6). The distribution of points shows that after the first periodic occurrence, coverage was typically higher for PECs with granularity $Q = 6$ than for PECs with granularity $Q = 24$. There were a number of PECs with $Q = 24$ hours that required a 4th occurrence to reach full coverage. It appears that, although coarser granularities result in more encounters per timestep, the broadcast time still increases. We suggest that this happens because coarser granularities result in many PECs having large diameter (Figure 5.7). In PECs with a large diameter, central nodes can have a greater negative effect on broadcast time by limiting the rate at which information spreads to the periphery of the community.

To further study the impact of diameter on broadcast time, we plotted the broadcast

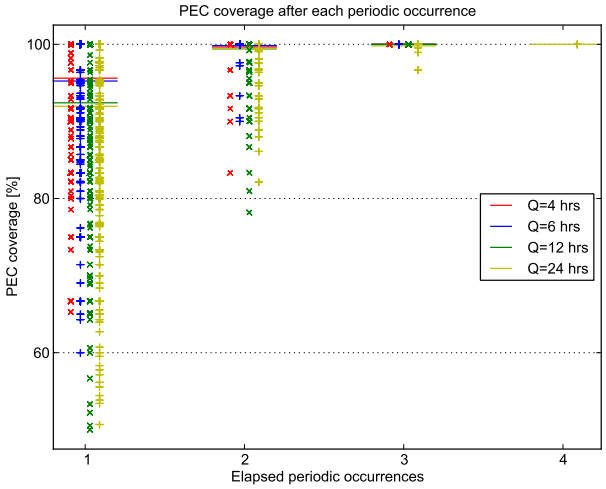


Figure 5.6: Coverage percentage after each periodic occurrence for PECs in the REALITY dataset. Points show the coverage $\bar{f}_c(\mathcal{P}, S_\lambda(x))$ of each PEC $\mathcal{P}$ after each of its $x = 1, 2, \ldots, \Lambda(\mathcal{P})$ periodic occurrences. Horizontal lines show the average coverages.

times for PECs grouped by diameter (Figure 5.8). We can see that the broadcast time increases for PECs with larger diameter. However, it is interesting that as diameter increases, PECs required worst-case broadcast time less frequently. For example, although PECs with diameter six have a potential worst-case broadcast time of six, none required more than four occurrences. Only at smaller diameters do broadcast times begin to approach worst-case times; for example, 8% of PECs with diameter three required worst-case time.

The implications of these results for decentralised PEC detection are that, for most PECs in the dataset, detection of maximal PECs by the nodes in a community occurs rapidly. On average, the coverage percentage reaches 92% after the first occurrence of the community. Furthermore, the patterns of encounters within the PECs are such that, for finer granularities, the majority of the PECs were detected within 0.22 of their potential worst-case time.

5.5 Discussion and related work

There is now an established literature for identifying sub-structures within static network topologies, both on a centralised and decentralised basis. However, existing static

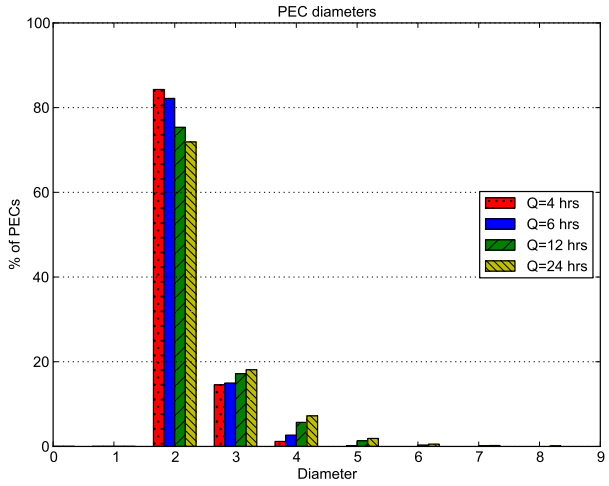


Figure 5.7: Diameters of PECs in the REALITY dataset at different granularities. PECs with diameter equal to one were not included in the experiments.

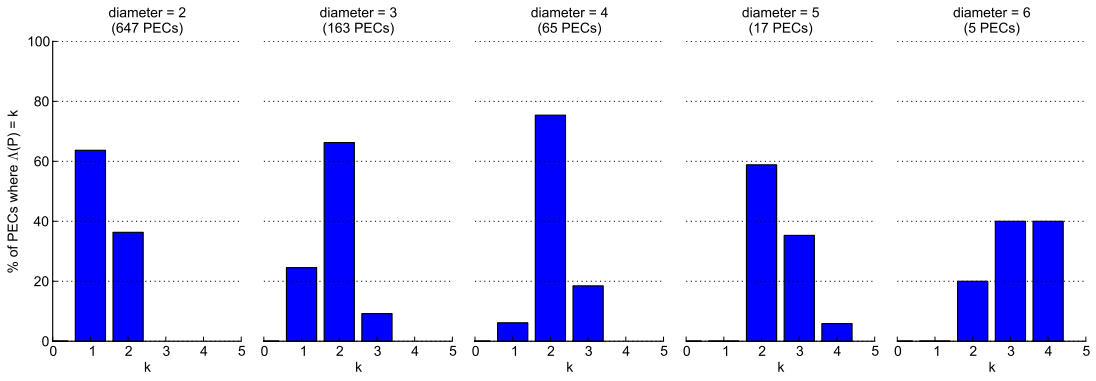


Figure 5.8: Comparison of PEC broadcast times (measured in number of periodic occurrences, k) by diameter. An individual plot shows the frequency distribution of broadcast times for PECs with the same diameter. $Q = 24$hrs.

methods (i.e., those in the category AGGREGATE) fail to capture periodicity in the encounters between individuals. In particular, we refer to the traditional community detection methods in the field of network science. Community detection seeks to identify highly clustered components in large real-world networks. Many community detection methods have been proposed, but most are intended for offline analysis of networks (see [For10] for a comprehensive survey of community detection methods). Furthermore, most methods analyse static networks; i.e., where interactions have been aggregated into a single graph regardless of their time and order.

Much of the existing network science literature has focused on communities that are defined only by node membership. This contrasts with our concept of a PEC, whose community structure is defined by both nodes and edges, making it similar to the static link communities studied by Ahn et al. in [ABL10]. This allows us to capture the configuration of the encounter relationships in a community as well as the individuals that belong to it. This also permits a node to belong to multiple communities, which can occur when the node's communities have different edge structure (as with link communities) or represent different periodic patterns.

The most relevant community detection algorithms to our work are those of Hui et al. [HYCC07]. These algorithms are notable as they offer a decentralised approach for nodes to detect the static encounter communities they belong to over time. Although this moves the method into the RECENT category in the classification scheme presented in Chapter 2, the algorithm does not consider periodic trend in the encounter patterns. Other recent research into the dynamics of community structure, such as that of Palla et al. [PBV07], has analysed the evolution of networks over time. So far there has been little work in this area that considers periodic communities.

Early analyses of human encounters focused on time-invariant characteristics, such as inter-encounter time and encounter duration [CHC$^+$07, HC08]. More recently, attention has been given to the analysis of temporal patterns in human encounters, such as the work discussed in Chapter 2. The work of Tang et al. in [TSM$^+$10] and [TMML09] is particularly relevant as it uses a dynamic graph representation to retain temporal information about encounters. The authors analyse the temporal dynamics of information diffusion in these graphs, but without specifically considering communities or period-

icity. Lahiri and Berger-Wolf [LB09] use a similar graph construction to formulate the problem of identifying subgraphs that appear periodically in real-world networks. We use the framework introduced by Lahiri and Berger-Wolf to define the periodic encounter community detection problem. However, the PSE-Miner algorithm proposed by the authors is intended for use in offline analysis and assumes global knowledge of the network, and is therefore not suitable in our decentralised setting. In addition, the formulation presented by the authors does not distinguish between communities and subgraphs.

A substantial amount of related work has been motivated by the study of opportunistic networks and, more specifically, human encounter networks. Such networks attempt to use encounters between wireless enabled devices to store, carry, and forward content for enabling a wide range of applications. Consequently, the temporal patterns of encounters allow content-sharing protocols in opportunistic networks to make better-informed forwarding decisions. Protocols such as those in [BCP08a, DFL01, LDS04] build an understanding of encounter familiarity between nodes. However, these protocols do not attempt to capture any regularity that may be present in encounter patterns. Some newer protocols, such as those in [MM09, MMC08], include statistical models that incorporate periodicity. These models require parameters regarding the periodicities of encounters to be known a priori. For example, in [MMC08] a single period must be specified, which precludes detection of repeating encounters at other periods.

The aforementioned protocols analyse only pairwise patterns. Broader relationships between nodes (e.g., acquaintances of acquaintances) are not considered. The Habit [MMC09] protocol attempts to merge both multi-node encounter behaviour and periodicity. Habit begins with node-centric pairwise analysis of regularity patterns between familiar strangers and, subsequently, nodes exchange their regularity patterns to build up a regularity graph. The model, however, requires a priori domain-specific period and memory parameters.

5.6 Conclusions

In this chapter we defined the concept of a periodic encounter community (PEC) and the problem of individuals self-detecting PECs in a decentralised network. Unlike our work measuring regularity in Chapter 4, our objective in this chapter has been to identify periodic encounters embedded within a mixture of incidental and routine encounters, without any restriction on the periods with which patterns repeat. To solve this problem we proposed a novel decentralised algorithm which is capable of automatically identifying community periodicities and is able to extract all globally maximal PECs, under the condition that there are sufficient exchange opportunities between nodes. Our analysis considered the diffusion of information within PECs, providing insight into the time required for PECs to be constructed.

The time required for individuals in a PEC to discover the whole community is of particular interest. This reveals the content-sharing dynamics of PECs in human encounter networks and is also a practical concern for protocols implementing PEC detection. Analytical study of diffusion in PECs shows that worst-case broadcast time for a PEC is given by its community diameter. The experimental results from the REALITY dataset show that PECs with large community diameter require a longer time to reach full coverage, further demonstrating the influence of community diameter on information diffusion. Our results also show that, in the dataset we studied, the time required for a PEC to reach full coverage was typically much shorter than the PEC's worst-case time.

Aside from insights concerning human periodic behaviour, this chapter also introduces the abstract notion of a community of nodes sharing the same temporal structure. In this chapter the temporal structure is a periodic support set; however, alternative temporal information may be used. In addition, we have introduced a temporal component to the idea of distributed nodes learning about the broader communities they belong to (presented by Hui et al. in [HYCC07]). After extracting their local (i.e., intrinsic) communities nodes discover their broader communities by incrementally extending their local perspective when they encounter other nodes, while at the same time avoiding the storage of redundant information. To use an alternative temporal structure in the same framework we simply need to define appropriate compatibility and construc-

tion rules, and develop an appropriate local mining algorithm that will extract locally maximal communities sharing this alternative definition of temporal structure. The extensibility of the framework introduced here will be leveraged in the next chapter where we use an alternative temporal structure which overcomes the limitations of our discrete-time representation of periodic encounter communities.

Chapter 6

Regularity in human encounter patterns

Introduction

In this chapter we introduce and explore an alternative concept of a community defined by its shared encounter patterns. Unlike the discrete-time representation used for the definition of a *periodic encounter community* in Chapter 5, we draw on the framework from Chapter 4 to define the concept of a *regular encounter community* (REC), which models periodicity in terms of IEI (inter-event interval) patterns. The strict discrete-time representation used for PEC detection resulted in some loss of temporal resolution. This occurs due to the binning of encounters into timesteps of size Q. The approach we introduce in this chapter overcomes the loss of temporal resolution by instead dealing with the underlying time-resolved events. Furthermore, this alternative approach is more tolerant to minor variations in event timing.

Informally, a REC is a community of individuals that all share the same time-of-week meeting pattern. As part of our definition of a REC, this chapter also introduces a novel IEI analysis method that allows nodes to distinguish between regular encounters and irregular encounters. This method allows us to determine whether different pairs of nodes share similar regular encounters. The concepts of periodic encounter community and regular encounter community are similar in that their respective definitions bring together a community of nodes and temporal information describing how those nodes meet over time. In the case of a PEC the temporal information is a periodic support set, whereas for a REC we use an alternative descriptor of periodic encounter behaviour,

called a *regularity mask*.

Although this chapter focuses on time-of-week encounter patterns, we note that our approach is not restricted specifically to a seven-day periodicity. Our motivation for selecting weekly patterns is our finding from Chapter 5 that the strongest periods identified in PEC detection were one, seven, and 14 days. Selecting a seven-day period for REC detection allows us also to incidentally detect any one-day periods that repeat long enough to appear as weekly patterns.

The REC detection algorithm presented in this chapter builds on the decentralised algorithm introduced earlier in the thesis. To adapt the algorithm to detect RECs we develop a new local miner and define REC compatibility and combination rules. Aside from these differences, the process of opportunistic sharing and construction is unchanged. As with Chapter 5, we explore RECs in empirical datasets by evaluating token broadcast performance. In particular, we explore RECs in the REALITY and DARTMOUTH encounter datasets. Our findings provide further insights into the role of periodic behaviour in human encounter networks and compares the characteristics of PECs and RECs.

Chapter outline

In Section 6.1 a method for distinguishing between regular and irregular encounters is presented. This method is fundamental to the definition of the REC detection problem, which we present formally in Section 6.2. We then adapt the decentralised PEC detection algorithm to detect RECs in Section 6.3. In Section 6.4 we explore the character and prevalence of RECs in real-world datasets and evaluate REC token broadcast dynamics. The chapter is concluded in Section 6.5.

6.1 Identifying regular encounters

In this section we build on the IEI analysis tools in Chapter 4 to identify which events in a chronology (if any) belong to a regular weekly pattern. These tools provide a way to identify repeat behaviour while retaining the time-resolution of events.

To distinguish events that consistently occur at the same time of week from irregular events we consider the dispersion in IEI values, as measured by the instantaneous coefficient of variation $c_{\text{var}}(\cdot)$ (Definition 4.2). More formally, let $\mathcal{S}_{v,w} = \{t_i \,|\, i = 1, \ldots, L\}$ be a chronology of encounters between two individuals v and w, where L is the number of encounters between v and w. As in Section 4.2.2, we translate the chronology to a set of N event trains, each covering a window of length ω. Given that we deal specifically with weekly similarity, ω is set to one week, and therefore each of the N trains represents one week in the chronology. We denote the event train corresponding to the nth week as

$$U^n = \{u_i^n \,|\, i = 1, \ldots, L_n\}$$

where L_n is the number of encounters in the nth window. A low $c_{\text{var}}(u)$ for some time-of-week offset u indicates that encounters occurring in that region of low coefficient of variation are consistently timed across all weeks. For an encounter at offset u we can consider whether it belongs to a time-of-week where events are consistently timed by consulting the $c_{\text{var}}(u)$ value, and marking the encounter as regular or not accordingly. More formally, we set a *dispersion threshold* θ to identify regular encounters. An event at offset u where $c_{\text{var}}(u) \leq \theta$ is marked as regular. We seek to extract the events across all N trains that are marked regular in this way, resulting in a subset of all time-of-week events, which we will refer to as the regular set $R(\mathcal{S}_{v,w})$. $R(\mathcal{S}_{v,w})$ is therefore the subset of events in the chronology's master train (Definition 4.4) that were marked as regular.

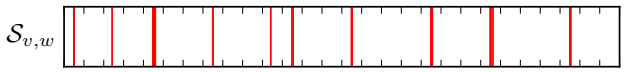
Definition 6.1

Given a chronology $\mathcal{S}_{v,w}$ and its corresponding master train U^* of duration ω, the **regular set** $R(\mathcal{S}_{v,w})$ is the set of regular events

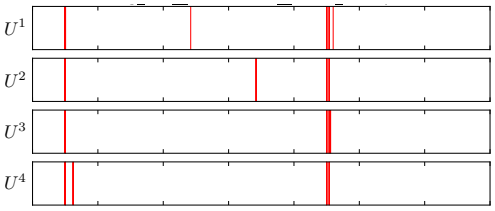
$$\{\, u \mid u \in U^* \;\wedge\; c_{\text{var}}(u) \leq \theta \,\}\,.$$

Figure 6.1 provides an example of each stage in obtaining the regular events for a chronology.

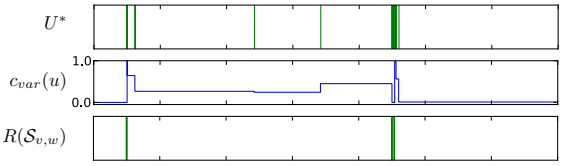
Obtaining the regular set forms the basis for the definition of a regular encounter community, which we will discuss in the following section. As noted in Section 4.2.3, computation of IEI measures such as coefficient of variation are linear in complexity.



(a) An example chronology $\mathcal{S}_{v,w}$ depicting encounters between individuals v and w over 28 days.



(b) Four event trains U^1, U^2, U^3, and U^4 each of duration $\omega = 7$ days, corresponding to chronology $\mathcal{S}_{v,w}$.



(c) Master train U^* and the instantaneous coefficient of variation $c_{\text{var}}(u)$ at each event $u \in U^*$. Filtering by dispersion threshold $\theta = 0.2$ yields regular events $R(\mathcal{S}_{v,w})$.

Figure 6.1: Pipeline for obtaining the regular events from an example chronology.

The algorithm for obtaining $R(S_{v,w})$ is therefore more computationally efficient than extracting local PECs using PSE-Miner, but this improvement comes at the additional requirement that a single period of regularity (controlled by the window size ω) must be set a priori.

6.2 The REC detection problem

A regularity set $R(\mathcal{S}_{v,w})$ represents the regular encounters between a given pair of individuals. At a broader scale, the encounters among a whole community of individuals may also share the same regularity, giving rise to a regular encounter community (REC). Before presenting the preliminaries necessary to define the concept of a REC we will informally introduce the concept with an introductory example.

6.2.1 Introductory example

In the simplest case, a given individual v may meet two friends w and x both with exactly the same regularity. For example, v always meets the two friends at 14:00 Tuesday and at 12:30 Friday. In this case, we have $R(\mathcal{S}_{v,w}) = R(\mathcal{S}_{v,x})$, and a REC exists consisting of v, w, and x on the basis of this shared regular encounter pattern.

For a more-complex scenario, let us add a third friend y who also meets with v at

14:00 Tuesday each week, but not at any other time. We may be tempted to include this third edge in the aforementioned REC; however, $R(\mathcal{S}_{v,y})$ does not share a regular 12:30 Friday encounter. Here we instead have a second REC, consisting of v, w, x, and y, on the basis of a 14:00 Tuesday regular encounter pattern. This property of a REC mirrors PECs and their rules regarding when one PEC subsumes another and when two PECs are distinct (Section 5.1.1).

These examples highlight that the definition of a REC unites a community not only in that each edge independently has a number of regular events, but also collectively all edges share the same regularity.

6.2.2 Problem formulation

When comparing the regular events $R(\mathcal{S}_{v,w})$ among the edges of a prospective community we must permit a small degree of uncertainty around the timing of regular events. To model this we allow for an amount of jitter around the regular events identified in $R(\cdot)$. The amount of jitter we permit is controlled by parameter ϕ which we use to map a regular event at time u to an interval $(u - \phi, u + \phi]$. Formally, we model a regularity mask $R_{v,w}$ that represents the regions of regularity corresponding to the regular events in $R(\mathcal{S}_{v,w})$.

Definition 6.2

Let $R(\mathcal{S}_{v,w})$ be the set of regular events for the chronology of encounters $\mathcal{S}_{v,w}$ between individuals v and w, constructed with window size ω. Given a **jitter tolerance** of ϕ, the **regularity mask** $R_{v,w}$ is defined as the set

$$\bigcup_{u \in R(\mathcal{S}_{v,w})} (u - \phi, u + \phi]$$

with values falling outside $(0, \omega]$ being wrapped around.

Although our definition shows a regularity mask for the encounters between a single pair of individuals, we can also obtain a regularity mask for two or more edges. To do this we obtain the intersection of the two or more regularity masks. The intersection of regularity masks of multiple edges represents the regions of regularity that are shared by the encounter patterns at all these edges. In other words, by intersecting multiple

regularity masks we obtain a mask that contains the regular regions common to all masks, if any exist. For example, given a path graph through individuals v, w, x, y, and z, the regularity mask $R_{v,w} \cap R_{w,x} \cap R_{x,y} \cap R_{y,z}$ represents the subset of regular regions common to all four regularity masks.

The commutative and associative properties of the intersection operator are beneficial when developing a decentralised solution to REC detection. These properties simplify the process of incrementally combining RECs through knowledge sharing. In particular, it means that a node does not need to know the original regularity masks from which an intersection was computed, and the node can apply additional intersection operations without the order in which they are applied affecting the result.

We should also note other notation we intend to use when presenting regularity masks. The empty regularity mask is denoted by $\emptyset$ and indicates that no regular regions were shared among the corresponding pair of individuals or collection of edges. We write $R_1 \subseteq R_2$ to denote the relationship of regularity mask R_1 being a subset or equal to regularity mask R_2. Finally, we use $|R|$ to denote the length of a regularity mask R. $|R|$ can be thought of as the overall duration of the regular regions in R and is defined as follows.

Definition 6.3

The **regularity length** $|R|$ of a regularity mask R of window size ω is

$$\int_0^{\omega} f(u)\, du$$

where $f(u) = 1$ if $u \in R$ and $f(u) = 0$ otherwise. $|R|$ is equivalent to the Lebesgue measure of R.

In practice, we implement a regularity mask as an ascending series of non-overlapping intervals, each interval representing a continuous regular region. Given regularity masks R_1 and R_2, with R_1 represented by n_1 intervals and R_2 represented by n_2 intervals, computing $R_1 \cap R_2$ is an efficient operation, requiring $\mathrm{O}(n_1 + n_2)$ comparisons.

Our formal definition of a regular encounter community (REC) is similar to that of a PEC (Definition 5.6). PECs and RECs are both represented by two components; namely, the encounter community (Definition 5.3) and a description of an encounter pattern that is shared by all edges in the community. The key difference in the case

of RECs is that the temporal information is defined in terms of a regularity mask, as captured in the following definition.

Definition 6.4

Denoted as the pair $\mathcal{R} = \langle C, R \rangle$, a **regular encounter community** (or **REC**) is an encounter community $C = (V, E)$ along with a non-empty regularity mask R where

$$R \subseteq \bigcap_{(v,w) \in E} R_{v,w} \ .$$

A REC's regularity mask represents the durations of the week where the community is regular. In many cases these regularity masks are intersections of two or more pairwise regularity masks. Given a REC $\langle C, R \rangle$ with community $C = (V, E)$ constructed with jitter threshold ϕ, the regularity mask R encodes information about the locations of regular spikes in the chronologies corresponding to edges in E. In particular, a time-of-week offset u, such that $u \in R$, indicates that for each edge $(v, w) \in E$ there exists at least one regular encounter x in the regular set $R(\mathcal{S}_{v,w})$ where $|u - x| \leq \phi$.

The concepts of REC subsumption and maximality follow from those of PEC subsumption and maximality, substituting the periodic support with a regularity mask in the original definition of PEC subsumption (Definition 5.9).

Definition 6.5

Let $\mathcal{R}_1 = \langle C_1, R_1 \rangle$ and $\mathcal{R}_2 = \langle C_2, R_2 \rangle$ be two RECs. We say that $\mathcal{R}_1$ is **subsumed** by $\mathcal{R}_2$ if and only if $R_1 \subseteq R_2$ and $C_1 \subseteq C_2$. We denote this relationship by $\mathcal{R}_1 \subseteq \mathcal{R}_2$.

Definition 6.6

A REC $\mathcal{R}_1$ is **maximal** if and only if there does not exist another REC $\mathcal{R}_2$, where $\mathcal{R}_1 \neq \mathcal{R}_2$, such that $\mathcal{R}_1$ is subsumed by $\mathcal{R}_2$.

Finally, we formulate the REC detection problem.

Definition 6.7

The **regular encounter community detection problem** is the problem of finding all maximal regular encounter communities that exist among the chronologies of a

population of individuals.

We have so far not discussed the minimum number of events required for a regular encounter pattern in a chronology to be meaningful. In the previous definition we also omit this requirement; however, in our experiments, we will assume that chronologies with too few events are are filtered out before REC detection. As with Chapter 4, we will use a minimum of two events per week.

6.3 Decentralised REC detection

In this section we present a decentralised algorithm to solve the regular encounter community detection problem. This REC detection algorithm follows the same opportunistic construction approach used to solve the PEC detection problem (presented in Section 5.2). To make the REC detection problem amenable to the decentralised PEC detection algorithm we must define a local REC miner and the rules of REC compatibility and combination. The task of the local REC miner is to extract all maximal local RECs at a given node. The locality of a node is the set of chronologies incident at the node and is relevant to our problem as it corresponds to the encounter data that a node can directly observe. Compatibility and combination rules are necessary during the opportunistic construction stage. These allow a node to compare a local REC to a REC held by a proximate node and, if they are compatible, to combine these two RECs to generate a new REC.

6.3.1 Compatibility and combination rules

To adapt compatibility to the case of RECs we simply modify the temporal containment rule to consider the intersection of regularity masks rather than the compatibility of periodic support sets.

Definition 6.8

Two RECs $\langle C_1, R_1 \rangle$ and $\langle C_2, R_2 \rangle$ with encounter communities $C_1 = (V_1, E_1)$ and $C_2 = (V_2, E_2)$ are **compatible** for node v if all of the following hold:

1. Relevance to v: $v \in V_1$ and $v \in V_2$;
2. Structural overlap: $E_1 \cap E_2 \neq \emptyset$;
3. Intersection of regular regions: $R_1 \cap R_2 \neq \emptyset$.

This leads to a similar modification to combination. Instead of taking the contained periodic support set, we take the intersection of the two RECs' regularity masks. This is the only regularity that both RECs share, and therefore the only maximal regularity mask valid for the combined encounter community. More formally, given two compatible RECs $\langle C_1, R_1 \rangle$ and $\langle C_2, R_2 \rangle$, we construct a new REC $\langle C', R' \rangle$ where $C' = C_1 \cup C_2$ and $R' = R_1 \cap R_2$. We note that in the case $R_1 = R_2$, both RECs will be subsumed by the new REC. Given the continuous-time nature of RECs we expect this case to be rare when compared to the crisp scenario of PEC detection.

6.3.2 Mining local RECs

The algorithm introduced here extracts maximal local RECs. Formally, we consider the task of mining all maximal local RECs for a node v. If we let $\mathcal{N}_v$ denote the set of nodes which v has encountered a minimum number of times, our task is to find all maximal RECs in the tree graph rooted at v and consisting of nodes $\{v\} \cup \mathcal{N}_v$.

There are $2^{|\mathcal{N}_v|} - 1$ connected subgraphs within this tree. If we were to use a brute-force approach to mining local RECs the algorithm would need to construct each of these communities and check if each is a valid REC. The task of checking whether a local connected subgraph at v is a valid REC is straightforward. Let W be a subset of neighbours $W \subseteq \mathcal{N}_v$. By inducing a subgraph from the set of nodes $\{v\} \cup W$ we obtain an encounter community $C = (V, E)$. The regularity mask intersection for C is given by

$$R = \bigcap_{(v,w) \in E} R_{v,w} \ .$$

We can therefore obtain a REC $\langle C, R \rangle$ if the condition $R \neq \emptyset$ is met.

This brute-force approach requires checking of each non-empty subset of $\mathcal{N}_v$ and is clearly computationally expensive. Furthermore, this approach requires an additional step to check whether each REC it generates is subsumed by RECs it has previously

generated. There are features of RECs that allow us to build a local miner that is more efficient than the brute-force approach.

Algorithm 6.1 presents our more-efficient approach, which exploits the properties of RECs to avoid unnecessary or redundant computation. First, we note that the algorithm avoids re-checking re-orderings of the same subset of neighbour nodes. Through the commutative property of regularity mask intersection we know that once a subset of neighbours has been checked, any re-orderings of that same subset will yield the same result.

A second improvement uses the property that if a particular neighbour subset $W \subset \mathcal{N}_v$ results in an empty regularity mask intersection, then any neighbour subset W' where $W \subset W' \subset \mathcal{N}_v$ will also result in an empty regularity mask intersection. A particular call to `REC-Generator` constructs $|S_0|$ neighbour sets, generates a local connected subgraph rooted at v from each one, and checks if each forms a valid REC. The function also recursively checks each neighbour set, with another neighbour node being introduced at each recursive call, until no more unused neighbour nodes remain. After the algorithm adds a node from S_0 to W_0 to produce W_1, if the neighbour set W_1 results in an empty regularity mask we know that any subsequent recursive calls adding another node to W_1 will also result in empty mask, and therefore no further recursion involving W_1 is necessary.

Finally, the algorithm prunes subsumed RECs during each recursive call to the function `REC-Generator`. Given that the algorithm begins with the largest possible regularity mask intersections (i.e., the regularity masks between v and each of its neighbours) and incrementally intersects these with other masks, the only subsumption the algorithm must check for is structural subsumption; that is, subsumption of REC $\langle C_1, R_1 \rangle$ by REC $\langle C_2, R_2 \rangle$ due to $C_1 \subset C_2$. To describe this situation further we consider a particular call to `REC-Generator`. Subsumption only occurs if a recursive call to `REC-Generator` by the current call returns a list containing a REC $\mathcal{R}_2$ that has the same regularity mask as a REC $\mathcal{R}_1$ stored in L_1. In this case we must check for subsumption of $\mathcal{R}_1$ by $\mathcal{R}_2$ or vice versa and cull the subsumed REC.

Algorithm 6.1: REC-Local-Miner

Input: Node v whose maximal local RECs will be extracted
Input: The set $\mathcal{N}_v$ of all nodes neighbouring v
foreach $w \in \mathcal{N}_v$ **do**
    Precompute $R_{v,w}$
end
$L \leftarrow$ `REC-Generator(` $v, \{\}, \mathcal{N}_v$`)`
Output: The set L of all maximal RECs local to v

Function `REC-Generator(`v, W_0, S_0`)`
    Input: Node v
    Input: A set W_0 of neighbour nodes
    Input: A set S_0 of neighbour nodes not yet added to W_0
    if $S_0 = \emptyset$ **then**
        return $\{\}$
    end
    Create empty list L_0 to hold candidate RECs
    Create set of neighbours S_1 as a copy of S_0
    foreach $s \in S_0$ **do**
        Remove s from S_1
        $W_1 \leftarrow W_0 \cup \{s\}$
        $R_1 \leftarrow \bigcap_{w \in W_1} R_{v,w}$
        if $R_1 \neq \emptyset$ **then**
            Generate a REC from neighbour set W_1**:**
            $E_1 \leftarrow \{(v, w) \mid w \in W_1\}$
            $V_1 \leftarrow W_1 \cup \{v\}$
            $C_1 \leftarrow (V_1, E_1)$
            $\mathcal{R}_1 \leftarrow \langle C_1, R_1 \rangle$
            Create list L_1 consisting of candidate REC $\mathcal{R}_1$
            Remove each REC in L_0 that is subsumed-by-or-equal-to a REC in L_1
            Remove each REC in L_1 that is subsumed-by-not-equal-to a REC in L_0
            Add all RECs in L_1 to L_0
            Generate RECs with nodes remaining in S_1**:**
            $L_2 \leftarrow$ `REC-Generator(`v, W_1, S_1`)`
            Remove each REC in L_0 that is subsumed-by-or-equal-to a REC in L_2
            Remove each REC in L_2 that is subsumed-by-not-equal-to a REC in L_0
            Add all RECs in L_2 to L_0
        end
    end
    return L_0
end

6.4 Experiments and results

In this section we study the presence and behaviour of RECs in the real-world. We begin our investigation by exploring the character of RECs in the REALITY dataset. This allows us to understand how common RECs are, their typical size, and at which time-of-week RECs are likely to appear. By focusing on REALITY RECs we are able to draw comparisons with the REALITY PECs discussed in Chapter 5.

Following this we evaluate the information sharing potential of RECs using the token broadcast scenario. As with our application of the token broadcast scenario in Section 5.3, these results inform us on the speed with which nodes of a REC can discover their REC, and the speed of content sharing within the REC. The datasets used for token broadcast analysis are REALITY and DARTMOUTH.

Given the REALITY dataset's academic context, the mobility of subjects is very likely to fluctuate throughout the year, leading to substantial variation in encounter rates and the consistency of patterns. For example, calendar events such as exam periods, recesses, and teaching semesters result in significant changes in behaviour among the participants. For our REC experiments we select a four-week period where encounter behaviour was most stable. Figure 6.2 shows the variation in the number of seven-day-period PECs in the dataset. Using this figure we can identify where PECs were most frequent, and therefore where RECs are also likely to be most common. Due to an anomalous lack of encounter activity on 31st October (possibly owing to data collection failure, a national holiday, or a local event) we avoid any durations intervened by this date. We select the period of 00:00 Monday 27th September to 00:00 Monday 25th October 2004 for our experiments, producing chronologies of duration $T_{max} = 28$ days. As with our analysis of visit patterns we select a window size $\omega = 7$ days and therefore chronologies are split into $N = 4$ encounter trains.

In the case of DARTMOUTH we select 7th April to 5th May 2003. This is the same 28 days used in the analysis of Dartmouth visits in Chapter 4.

Before running experiments we also discard any chronologies with fewer than two encounters each week. After carrying out this filtering, 33,484 REALITY encounters remain and $30,127$ DARTMOUTH encounters remain. In all following results we only

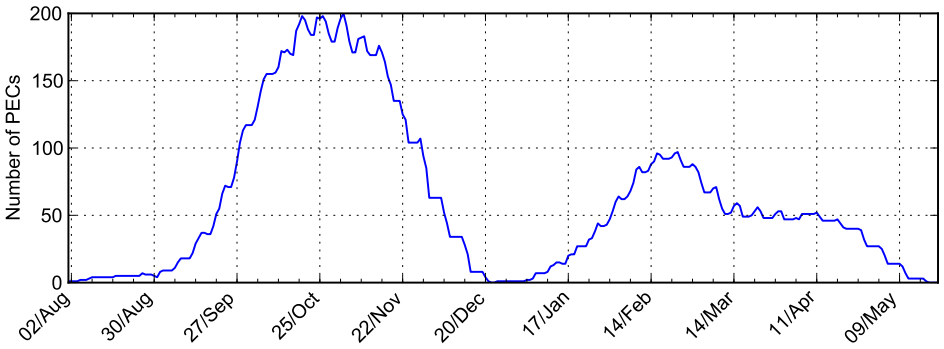


Figure 6.2: Occurrences of periodic encounter communities (PECs) found in the REALITY dataset that have a period of seven days. The height of the curve at date t corresponds to $|\{\, \langle C, S_{(i,p,n)} \rangle \in \mathcal{P}^* \mid i \leq t \leq i+(n-1)p \wedge p = 7 \text{ days} \,\}|$ where $\mathcal{P}^*$ is the set of all PECs extracted from REALITY with granularity $Q = 24$ hrs.

refer to these filtered datasets. For REC detection we set dispersion threshold $\theta = 0.2$ and jitter tolerance $\phi = 30$ minutes.

To illustrate the concept of a REC, five examples from the REALITY dataset are presented in Figure 6.3. $\mathcal{R}_1$ is a simple REC of two nodes that is regular at many points during the week. $\mathcal{R}_2$ consists of one node that regularly meets three other nodes on Mondays at 12:00 and 21:30. $\mathcal{R}_3$ is a Friday 20:00 REC whose community is a path graph and has diameter four. Finally, $\mathcal{R}_4$ and $\mathcal{R}_5$ are an interesting example of mutual non-subsumption. The triad of subjects in $\mathcal{R}_4$ is the same triad in $\mathcal{R}_5$ and both RECs have Wednesday 21:30 among their regular times. Given that $\mathcal{R}_5$ has an additional node we may incorrectly regard $\mathcal{R}_4$ as being subsumed by $\mathcal{R}_5$. However, $\mathcal{R}_4$ is regular at a second time of week that $\mathcal{R}_5$ is not (i.e., Thursday 11:00). The result is that $\mathcal{R}_4$ does not subsume $\mathcal{R}_5$ (due to $\mathcal{R}_5$ having an additional edge and node) and $\mathcal{R}_5$ does not subsume $\mathcal{R}_4$ (due to $\mathcal{R}_4$ having additional region of regularity).

6.4.1 Character of RECs in the REALITY dataset

210 REALITY RECs were detected in the four week period. Of the 50 nodes that remained in the dataset after removing those that did not have at least two encounters each week, 38 appear in one or more RECs. This indicates that RECs are prevalent in the REALITY dataset, with only 24% of nodes not exhibiting regular encounter beha-

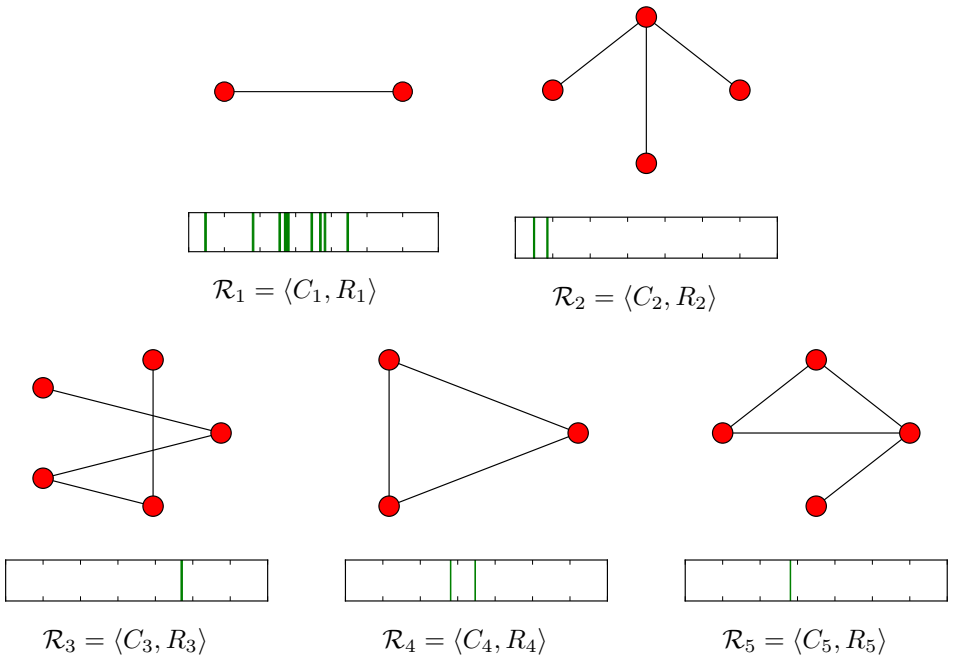


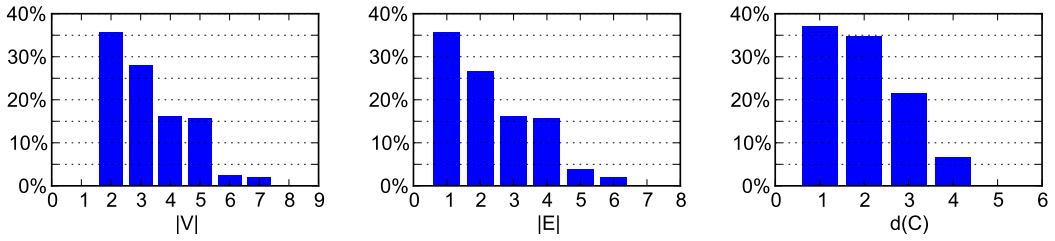
Figure 6.3: Example RECs from the REALITY dataset. A regularity mask is depicted as a rectangle containing green bars. Each rectangle is separated into seven chunks, each representing a day of week beginning with Monday and ending with Sunday. Ticks denote midnight. Green bars indicate the time-of-week during which the REC is regular.

viour.

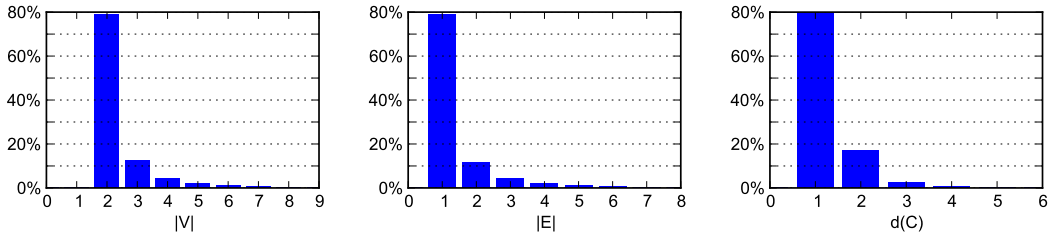
To investigate the structural size of RECs a number of community properties are presented in Figure 6.4. Unsurprisingly, RECs containing a large number of nodes and edges are less likely. Most-common are RECs containing two or three individuals, which accounts for 64% of the RECs.

The diameters of RECs are also relatively small, but typically larger than the diameters of the PECs extracted from the same dataset. 63% of RECs were found to have a diameter of 2 or greater, whereas only between 28% and 16% (depending on granularity Q) of PECs had diameters in this range. We suspect that the reason for the larger diameter is that the RECs tend to be shaped as path and tree graphs. Indeed, we found that only 4% of RECs contained one or more simple cycles, indicating that the large majority are tree graphs.

To investigate the structure of the communities further we consider the distribution of graph density in Figure 6.5. 76 of the RECs consist of two nodes. These two-node



(a) RECs.



(b) PECs. $Q = 4$ hrs.

Figure 6.4: Distributions of the number of nodes (denoted $|V|$), number of edges (denoted $|E|$), and diameters (denoted $d(C)$) of regular encounter communities (RECs) and periodic encounter communities (PECs) in the REALITY dataset.

communities are omitted from the distribution since they are by definition complete graphs and therefore always have a density of 1. We see that complete graphs among RECs with three nodes or more are rare. 44% of communities have density between 0.6 and 0.7, many of which are triad communities missing one edge. Intuitively, we would expect that if a node v meets two nodes w and p at roughly the same time each week, then w is also likely to have met p at the same time. We therefore expect triad RECs to tend to form cliques, a property referred to in social network analysis as transitivity [HL70]. However, our analysis finds that only three of the 59 triad RECs are transitive.

A number of factors may contribute to the intransitivity of RECs. The range of Bluetooth allows for an individual to detect two other devices that are not in range of one another. This may occur even in the case where all three individuals are stationary; however, intransitivity is more likely when one or more devices are moving. A highly mobile central node may encounter two or more individuals in succession without simultaneously being in contact with all at the same time. RECs such as these are interesting because they either represent an individual periodically repeating the same route that brings him/her into proximity with the same individuals, or an individual

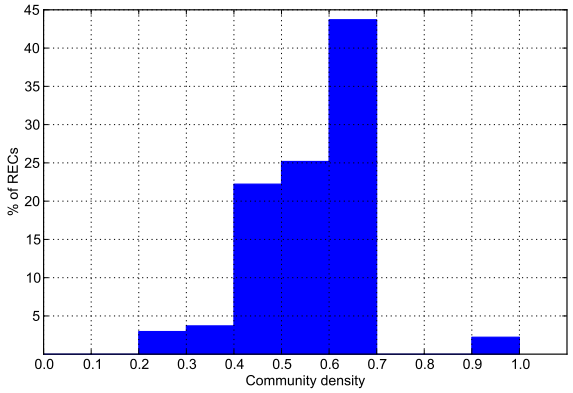


Figure 6.5: Distribution of community density for REALITY RECs consisting of at least three nodes. For a REC with community $C = (V, E)$, community density is given by $\frac{2|E|}{|V|^2-|V|}$.

who periodically acts as a bridge between two nearly proximate nodes. This finding also indicates that RECs are not necessarily cliques, highlighting a difference between our definition and that of a meeting group [YG10].

Regularity masks represent the temporal structure of RECs. As noted in Section 6.2.2, the regularity mask for a REC asserts that all edges in the community have at least one regular event within $\phi = 30$ minutes of all times covered by the regularity mask. The longer the length $|R|$ of a mask R, the more points during the week the community is regular for. Figure 6.6 shows the distribution of regularity mask lengths among the RECs extracted from REALITY. The average regularity mask length is 1.48 hours and the largest is 10.1 hours. Longer regularity mask lengths are associated with smaller diameters and, for the largest lengths, with two-node RECs. Regularity masks for RECs with more than three nodes are built from the intersection of the constituent edges, and are therefore almost always shorter in length than the two-node RECS they subsume.

Although the distribution of regularity mask lengths tells us the overall durations which a REC is regular during the week, it does not describe the times of week where RECs are regular. We explore this further in Figure 6.7 which plots the times of week that are typically covered by a REC's regularity mask. The tallest peak is at 16:45 Friday, indicating that many RECs were regular at this time (in addition to any other times they may be regular). Smaller peaks also appear on each weekday at 11:00 and 16:45. These

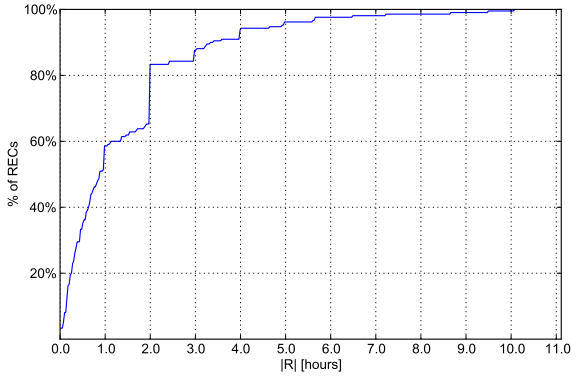


Figure 6.6: Cumulative distribution of regularity mask durations in the REALITY dataset. $|R|$ denotes the overall duration of a regularity mask R.

are significant times in the context of the REALITY dataset as they lie at the boundaries between MIT classes[1]. Exact class start and end times vary by day of week, but 11:00 is a common class start time and 16:30 is a common class end time. It is therefore likely that the 11:00 peaks correspond to subjects arriving at the same class each week, and that the 16:30 peaks correspond to subjects encountering one another during their commute from class or on arriving at their residences.

Figure 6.7 also shows that RECs are much less likely to form on weekends. This is likely due to a combination of two factors. First, many students choose to spend their weekends off campus, and therefore there is less encounter activity among participants. Second, when compared to weekdays, weekends have more erratic behaviour due to the lack of routine tasks such as timetabled lectures.

Although PECs and RECs are different definitions of community, it is interesting to consider whether there is any correspondence between the two. In particular, since the REALITY RECs we have extracted are based on weekly regularity (i.e., we set $\omega = 7$ days), we investigate whether these RECs resemble any REALITY PECs that have a period of seven days. To do this, we check each PEC to see if its nodes appear in one or more REC. If the nodes of a particular seven-day-period PEC are a subset of the nodes of a REC, then we regard these two communities as being similar. We base this analysis on the set of PECs with a period of seven days that appear during 27th

[1]MIT Fall 2004-2005 Class Schedule, via the Internet Archive: `http://web.archive.org/web/20040917051038/http://web.mit.edu/registrar/www/schedules/csbindex.shtml`

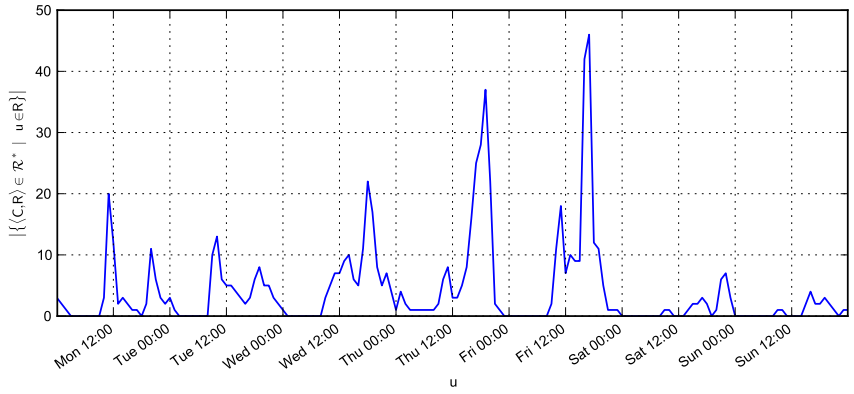


Figure 6.7: Distribution of regularity masks belonging to REALITY RECs throughout the week. Letting $\mathcal{R}^*$ denote the set of all RECs, the number of regularity masks that include the time of week u is given by $|\{\langle C, R\rangle \in \mathcal{R}^* \mid u \in R\}|$.

September to 25th October 2004 and were detected using granularity $Q = 24$ hours. This set consists of 82 PECs. Our results find that 58.5% of PECs had a node set that also appeared in at least one REC. Reciprocally, we found that only 14.2% of RECs had a node set that appeared in at least one PEC. From these results we draw two conclusions. First, there is not a one-to-one correspondence between the RECs and PECs in the REALITY dataset. Second, the majority of RECs (85.8%) have no corresponding PEC, and therefore RECs captured more weekly encounter behaviour than PECs.

6.4.2 Token broadcast in RECs

The principle of the token broadcast scenario in the context of RECs is the same as with PEC analysis (detailed in Section 5.3). Although PECs and RECs have different characteristics it is useful to evaluate them in a similar manner. Token broadcast is used to measure the time required for a global maximal REC to be discovered by all its constituent nodes through decentralised opportunistic sharing and construction. This scenario is analogous to each node in a REC attempting to broadcast a token to each other, and therefore also represents the speed of information propagation within the REC.

6.4.2.1 Selecting encounters relevant to a REC

Each encounter between two nodes represents a token-sharing opportunity. When evaluating token broadcast for a particular REC, only encounters described by that REC's information are used for token sharing. Assuming the window size is one week, these are the encounters whose time-of-week offset lies within the REC's regularity mask (or close enough, according to the jitter tolerance parameter) and correspond to one of the edges in the REC's community.

More formally, consider a REC $\mathcal{R} = \langle C, R \rangle$, where $C = (V, E)$, constructed from chronologies of duration T_{max}, with window size ω and jitter tolerance ϕ. To determine the set of edges where token exchanges will occur at time $t \in (0, T_{max}]$ we consider whether the window offset $t \bmod \omega$ is within ϕ of any value in R. If so, an encounter at t that corresponds to an edge in E can be used for token exchange. The following function $f(t)$ formalises this concept and represents the mapping of the time $t \in (0, T_{max}]$ to the set of exchanges occurring at time t:

$$f(t) = \Big\{ (v, w) \mid (v, w) \in E \wedge \exists\, t \in \mathcal{S}_{v,w},\, u \in R \text{ s.t. } |u - (t \bmod \omega)| \leq \phi \Big\}.$$

To evaluate $\mathcal{R}$ we apply token exchanges described by $f(t)$ for each $t \in (0, T_{max}]$ in ascending order.

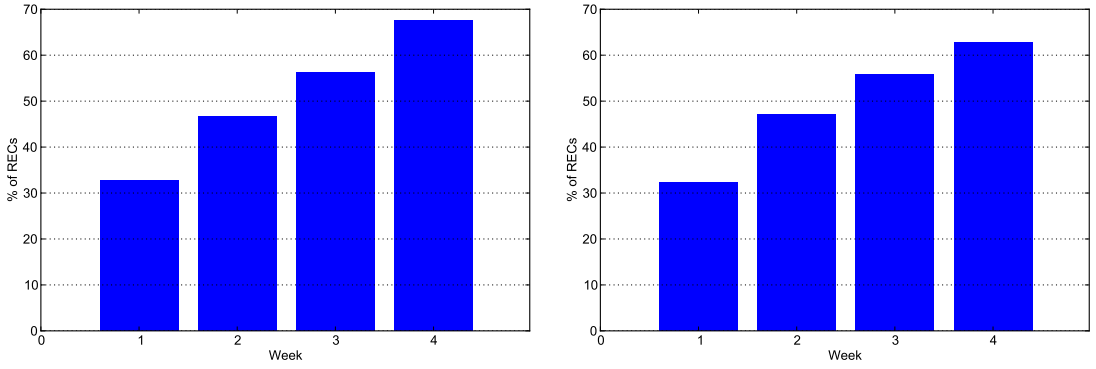
6.4.2.2 Broadcast time in REALITY and DARTMOUTH

Table 6.1 summarises the REALITY and DARTMOUTH datasets used in the following token broadcast experiments. We observe that RECs were less prevalent among nodes in DARTMOUTH than REALITY; in particular, 76% of REALITY nodes belonged to at least one REC, compared to 58% of DARTMOUTH nodes. This is likely due to the Reality Mining dataset being a closer representation of human encounters than the inferred Dartmouth WLAN encounters. REALITY's superior fidelity is due to it being direct-sensed data from subjects who are consciously participating in a study.

There are a number of RECs which failed to reach full token coverage after four weeks. This contrasts with the token broadcast analysis of REALITY PECs, where each PEC

	REALITY	DARTMOUTH
Duration	27/Sept to 25/Oct 2004	7/Apr to 5/May 2003
Total nodes	50	428
Total edges	166	863
Total encounters	33,484	30,127
RECs	210	773
Nodes appearing in one or more REC	38	247
Successful broadcasts	142	485
Average diameter	1.98	1.91

Table 6.1: Summary of datasets used in REC token broadcast experiments. Only nodes and edges that met the minimum number of encounters are included.



(a) REALITY. (b) DARTMOUTH.

Figure 6.8: Percentage of RECs that have reached full coverage (i.e., successful broadcast) at the end of each week. 32% of REALITY RECs did not reach full coverage. 37% of DARTMOUTH RECs did not reach full coverage.

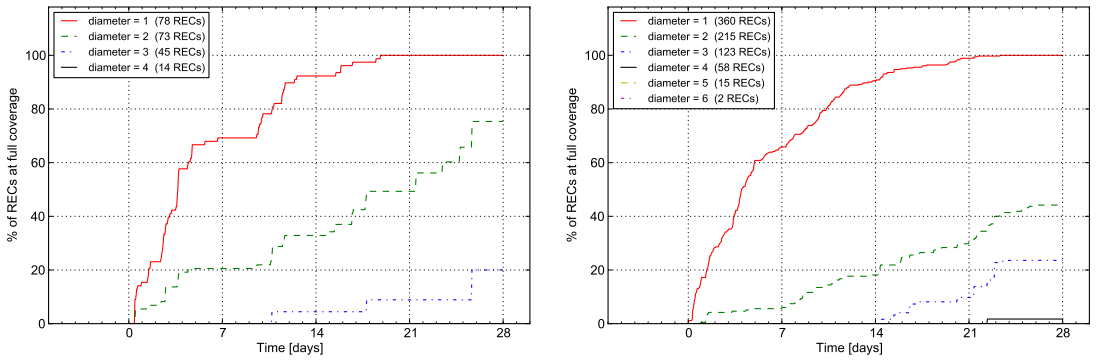
successfully broadcast all its tokens by the last timestep of its periodic support set. Figure 6.8 depicts the number of RECs that reached full coverage by the end of each week. This reveals how many RECs are able to successfully broadcast after applying each week of exchanges corresponding to a regularity mask. Over the first three weeks the two datasets are almost identical in the increases in successful broadcast. By the end of week four, however, broadcast in 68% of REALITY RECs was successful, compared to 63% of DARTMOUTH RECs.

Having found that community diameter has an influence on the speed of broadcast within PECs in Section 5.3.2, we investigate the extent to which diameter accounts for the higher failed broadcast rate in DARTMOUTH. A key difference between the two

types of community is that a PEC only exists if an encounter occurs in each of the timesteps described by its periodic support set, and therefore it is guaranteed that a token exchange will occur for each edge in the community in each periodic timestep. This is a necessary condition for a PEC's worst-case broadcast time to be its diameter. The definition of a REC is less strict than this, and permits a degree of variation in the timing of the regular encounters and is tolerant to occasional missing encounters. Although this means that the diameter no longer governs the worst-case broadcast time for RECs, community diameter does influence the speed of token propagation. In particular, a community of large diameter indicates the presence of two nodes separated by a large number of hops. Rapid propagation across multi-hop paths such as these requires frequent and interleaved encounters among the intermediate nodes, which is rare among the RECs we have extracted. On the other hand, low-diameter RECs allow for rapid token broadcast. For example, if we consider a REC $\mathcal{R} = \langle C, R \rangle$ that is a clique, it has diameter $d(C) = 1$ and in the worst-case only requires one encounter at each edge in C before reaching full coverage.

This behaviour is demonstrated in Figure 6.9, where we observe that all RECs with diameter one reached full coverage within 28 days and were the quickest to do so. The figure shows that 68% of one-diameter REALITY RECs reached full coverage within seven days and all were at full coverage within 19 days. One-diameter RECs in the DARTMOUTH dataset have a similar full-coverage rate, with 65% reaching full coverage within seven days. These RECs are cliques and, as mentioned earlier, only require one exchange per edge to reach full coverage. This statistic also indicates that in 31% (REALITY) and 35% (DARTMOUTH) of the one-diameter RECs there was at least one pair of nodes that did not have an encounter within the REC's regularity mask in the first week.

At larger diameters we see a significant reduction in both the number of successful broadcasts and the rate at which RECs reach full coverage. This confirms the significant influence of diameter on token broadcast within a REC.



(a) REALITY. (b) DARTMOUTH.

Figure 6.9: Percentage of RECs that have reached full coverage over time. RECs have been grouped to allow comparison of broadcast by diameter.

6.5 Conclusions

The strict discrete-time representation used for periodic encounter community (PEC) detection (detailed in Chapter 5) resulted in some loss of temporal resolution and was sensitive to minor variations in the timing of periodic encounters. In this chapter we solve these limitations by defining communities in terms of inter-event interval patterns, building on the tools we developed in Chapter 4 and the decentralised algorithm developed in Chapter 5. The concept of a regular encounter community (REC) introduced in this chapter is more tolerant to small variations in periodic encounter patterns and retains the time-resolution of encounters. We have re-used the same decentralised construction approach we introduced for PEC detection to allow nodes to self-detect their RECs, proving the extensibility of this algorithm to other types of temporal information.

Our results show that many individuals belong to one or more REC, making these an interesting feature for use in encounter-aware opportunistic forwarding protocols. Token broadcast analysis shows that diameter is again an important factor in the propagation of information, an observation we also made with periodic encounter communities (PECs) in Chapter 5. The requirement of a PEC that the community's encounters must strictly repeat according to the identified period means that diameter acts as a hard limit on the PEC's broadcast time. This contrasts with RECs, whose tolerance to minor variations in weekly patterns permits occasional missing encounters. We found that due

to this, and also due to the limited (four week) duration we allowed for propagation, a number of RECs failed to reach full token coverage. In practice, REC construction would be faster and have a higher success rate by allowing communities to also use their irregular encounters for opportunistic REC construction; however, for our experiments we restricted our evaluation to encounters intrinsic to each REC so that we could investigate the propagation characteristics specific to the community.

When directly comparing PECs and RECs of the same periodicity (i.e., weekly), we found that REC detection identified over 2.5 times the number of encounter communities than PEC detection. Given that human mobility is not a strictly timed behaviour, it is not surprising that permitting an amount of uncertainty in encounter patterns allows us to capture more periodic communities, and result in more RECs being detected. Indeed, RECs were able to account for 58.5% of the communities extracted by PEC detection, and also were able to identify an additional 180 communities (of 210 overall RECs) that did not appear as PECs.

Chapter 7

Conclusions

This thesis has proposed and developed methods for detecting periodic patterns in the visits and encounters of human individuals. As per the scope set out in Chapter 1, these methods are amenable to decentralised scenarios and operate on an event stream representation of data. We have used these methods to explore the presence and character of periodicity in human mobility, considering visit patterns in Chapter 4 and encounter patterns in Chapter 5 and Chapter 6.

7.1 Thesis summary and contributions

Our survey of models, methods, and analyses in Chapter 2 found that there is much interest in human mobility patterns, but limited work dealing specifically with periodic patterns concerning individual mobility. To help navigate the related approaches a classification scheme based on *scale* and *temporal context* was proposed. This classification scheme highlighted that our contributions are in the detection of PERIODIC temporal context at an INDIVIDUAL scale. The most-related work in the INDIVIDUAL-PERIODIC category lies in location prediction and mobile communication networks, where a number of recent approaches have modelled periodicity in visits and encounters in a variety of ways.

Throughout this thesis we use a number of real-world datasets to investigate periodic visit and encounter patterns. These are detailed in Chapter 3, along with a discussion of the datasets and experimental methods commonly used to investigate human visit and encounter behaviour. We noted that the empirical data were either naturally represented as event streams, or could be readily reduced to an event stream. This allows the meth-

ods developed in this thesis to be applied across a wide range of scenarios, a number of which we have empirically investigated in this thesis. In particular, the real-world datasets we explored include Foursquare checkins in three urban areas, WLAN AP visits on Dartmouth College campus, public transport users' visits to London Underground stations, and Bluetooth traces of encounters between Massachusetts Institute of Technology (MIT) students. This has allowed us to draw conclusions regarding periodic patterns in human mobility in a variety of scenarios.

The first method we developed in this thesis, named *IEI-irregularity*, was used to measure the amount of weekly periodicity in visit patterns. To deal with the visit event stream data in a way that retained the temporal resolution of events we adapted a neural coding technique which is based on the analysis of IEIs (inter-event intervals). This method is particularly suited to these data as it can be applied even when visits are sparse, which is often the case when dealing with individual mobility such as a particular individual's visits to a particular location. We presented IEI-irregularity and used it to explore real-world visit patterns in Chapter 4. IEI analysis was not only useful for the task of quantifying weekly periodicity in visit patterns, but also provides powerful and computationally efficient tools for handling patterns in event streams. Indeed, these IEI analysis tools were used in Chapter 6 to develop a method that identifies periodic encounters between two individuals. To my knowledge, this is the first application of neuroscience techniques to visit and encounter event stream data.

The exploration of real-world visit patterns using IEI-irregularity (results presented in Chapter 4) uncovered some interesting features of individual periodic mobility. In a variety of settings we found a core group of individuals that visit at least one location with near-perfect regularity. We also observed that the type of location, which is typically associated with a particular activity, has strong influence over individuals' visiting patterns. There are location types whose usage is predominantly driven by inflexible constraints (such as lectures in academic buildings) whereas others, such as outdoor venues, are less constrained or subject to external random effects. The IEI-irregularity measure allows features such as these to be automatically extracted by an individual's mobile phone, providing context to other mobile phone applications.

Chapters 5 and 6 explore the concept of a community of individuals that share the

same encounter pattern. Chapter 5 proposes a *periodic encounter community* (PEC), which is a community of nodes that meet according to a particular period (e.g., every 24 hours). The period at which a PEC encounters is itself a feature of the PEC and may be any value such that the community strictly encounters according to that period. On the other hand, a *regular encounter community* (REC), defined in Chapter 6, is a community of nodes that all share the similar regular encounters; specifically, we focused on RECs formed around having the same weekly encounters, although other choices of period may be used.

RECs and PECs are related concepts. Both capture a community that is periodically encountering in some way. Each has limitations and advantages for extracting periodic encounter patterns. The formulation of a PEC uses a discrete-time representation, resulting in reduced temporal precision, and making the existence of a PEC sensitive to variations in encounter times. This formulation is necessary to enable a periodic subgraph mining approach, which allows automated detection of a particular community's period. After investigating real-world PECs (Chapter 5) we found that the strongest PEC periods were one, seven, and 14 days. We therefore chose to select weekly patterns, as we also did in Chapter 4 for our visit patterns analysis, and build on the IEI analysis framework to propose the concept of a REC. Through the assumption of a single periodicity, RECs overcome the aforementioned limitations of PEC detection. Results presented in Chapter 6 that compared RECs and PECs showed that RECs did indeed capture more weekly patterns than PECs did.

The requirement of decentralisation makes developing REC and PEC detection methods particularly challenging; however, this is necessary for their deployment in human encounter networks. To design decentralised algorithms for the detection of PECs and RECs we proposed a framework consisting of local mining and opportunistic construction phases. This framework is sufficiently general to be applied in the PEC and REC detection algorithms proposed in this thesis, with only the components that deal with the differing temporal properties (regularity masks in the case of RECs and periodic support sets in the case of PECs) requiring modification. Our PEC detection algorithm adapts an existing centralised data mining approach, which allows the automatic detection of PECs' periods. For REC detection, a new REC local mining algorithm was

introduced. As far as I am aware, these are the first decentralised algorithms for the detection of encounter communities that periodically encounter.

Due to our interest in RECs and PECs as context in human encounter networks we gave particular attention to the content sharing dynamics within these communities. The token broadcast scenario (first presented in Chapter 5) let us analyse both content sharing dynamics and the time required for PECs and RECs to be discovered. In both cases community diameter acts as a limiting factor to broadcast within the community; communities with large diameters tend to take longer to complete broadcast than communities with short diameters. The strict, discrete-time formulation of PECs meant that diameter was a hard limit on broadcast time; on the other hand, since RECs allow for tolerance in the timing of regular patterns, their failed broadcast rate is greater.

7.2 Future directions

This thesis has made a substantial step in addressing the presence and detectability of periodicity in individual human mobility. We have validated the intuitive assumption that human routine has a profound influence on visit and encounter patterns, and provided methods for detecting periodic patterns. Our work opens a variety of interesting possibilities for future work. In particular, there are three broad directions: exploiting periodic mobility patterns in services and networks, further empirical study to understand the role of routine and periodicity in human behaviour, and extension of our methods.

A clear application area lies in context-aware mobile computing. Our methods can be used to augment existing context-aware software, such as digital personal assistants, to allow them to gain a deeper understanding of their owners' behaviours. By identifying periodic patterns in their users' visits, online location-based services can also benefit in a similar way. Regarding visit behaviour, we note that although our experiments primarily focused on the user's perspective, considering the prevalence of regular visitors at a particular location is useful for context for venue managers such as shop owners. This thesis also lays the foundation for investigation into the relationship

between periodic encounters and periodic visits and, in particular, exploring the extent to which joint regular visiting patterns lead to regular encounter patterns.

Opportunistic network research often relies on synthetic data and, as we noted in Chapter 3, there is limited work developing and validating synthetic periodic mobility models. The methods and findings in this thesis can inform mobility model design and be used to evaluate the extent to which periodic encounters and visits are expressed by a mobility model.

Recent opportunistic network routing protocols (and, specifically, those applied to human encounter networks) have been designed with the assumption of periodic encounter behaviour; however, prior to this thesis there was limited work empirically verifying and exploring the characteristics of these patterns. The findings of this thesis inform future protocol design and our decentralised PEC and REC detection methods can be exploited in encounter-aware routing. For example, communities of nodes that periodically encounter one another can act as a reliable backbone for routing. A future protocol can differentiate between these regular patterns and other irregular patterns and leverage them in different ways.

While the decentralised PEC and REC algorithms presented in this thesis are able to mine and then construct maximal communities over time, there is scope to extend them for application in a fully dynamic setting. In this setting, nodes would only detect encounter communities that currently exist, rather than also retaining those that have expired. A comprehensive solution to this problem would require a protocol that handles periodic re-mining and propagation of community updates, including community destruction as well as construction.

The decentralised methods we have introduced also provide a basis for a privacy-aware periodic community detection system. Our work in Chapter 5 and Chapter 6 already details how temporal community structures can be discovered by their constituent members without the need for members to provide information to a central authority or nodes outside their own communities. This is especially important for our methods because they reveal previously hidden mobility patterns. These patterns provide rich context to the variety of applications and services we have discussed in

this thesis, but in some cases there may be patterns that a user only wishes to share with particular friends and acquaintances. The opportunistic construction process of the PEC and REC algorithms relies on mobile peer-to-peer exchanges and it is this exchange mechanism that provides a convenient means to control how patterns are shared. Further work in this area would extend our construction process with mechanisms to incentivise cooperation and participation in the system while also preserving privacy of sensitive patterns and protecting against untrusted or malicious nodes.

Bibliography

[AB02] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[ABL10] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[AS95] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. 11th International Conference on Data Engineering*, pages 3–14, 1995.

[AS02] D. Ashbrook and T. Starner. Learning significant locations and predicting user movement with GPS. In *Proc. 6th International Symposium on Wearable Computers*, pages 101–108, 2002.

[ASC+09] H. Alani, M. Szomszor, C. Cattuto, W. Van den Broeck, G. Correndo, and A. Barrat. Live social semantics. In *Proc. 2009 International Semantic Web Conference (ISWC)*, Lecture Notes in Computer Science, pages 698–714. Springer Berlin Heidelberg, 2009.

[Bar05] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[BBG12] D. Bouneffouf, A. Bouzeghoub, and A. L. Gancarski. Following the user's interests in mobile context-aware recommender systems: the hybrid-e-greedy algorithm. In *26th International Conference on Advanced Information Networking and Applications (WAINA) Workshops*, pages 657–662, 2012.

[BCP08a] C. Boldrini, M. Conti, and A. Passarella. Exploiting users' social relations to forward data in opportunistic networks: the HiBOp solution.

Pervasive and Mobile Computing, 4(5):633–657, 2008.

[BCP08b] C. Boldrini, M. Conti, and A. Passarella. User-centric mobility models for opportunistic networking. In *Bio-Inspired Computing and Communication*, pages 255–267. 2008.

[BD02] P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*. Springer, 2002.

[BD05] G. S. Bhumbra and R. E. J. Dyball. Spike coding from the perspective of a neurone. *Cognitive Processing*, 6(3):157–176, 2005.

[BDSL11] G. R. Brown, T. E. Dickins, R. Sear, and K. N. Laland. Evolutionary accounts of human behavioural diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1563):313–324, 2011.

[BKM04] E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7(5):456–461, 2004.

[CCLS11] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. In *Proc. 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[CCW+11] M. J. Chorley, G. B. Colombo, M. J. Williams, S. M. Allen, and R. M. Whitaker. Checking out checking in: observations on foursquare usage patterns. In *Proc. International Workshop on Finding Patterns of Human Behaviors in Network and Mobility Data (NEMO) (ECML-PKDD Workshops)*, 2011.

[CCW+12] G. B. Colombo, M. J. Chorley, M. J. Williams, S. M. Allen, and R. M. Whitaker. You are where you eat: Foursquare checkins as indicators of human mobility and behaviour. In *Proc. 2012 IEEE Pervasive Computing and Communications (PERCOM) Workshops*, 2012.

[CE07] A. Clauset and N. Eagle. Persistence and periodicity in a dynamic proximity network. In *Proc. DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, 2007.

[CGW+08] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[CHC+07] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, 2007.

[CLP04] F. Chinchilla, M. Lindsey, and M. Papadopouli. Analysis of wireless information locality and association patterns in a campus. In *Proc. 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, volume 2, pages 906–917 vol.2, 2004.

[CML11] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, page 1082–1090. ACM, 2011.

[CMMD07] A. Chaintreau, A. Mtibaa, L. Massoulie, and C. Diot. The diameter of opportunistic mobile networks. In *Proc. 2007 ACM CoNEXT conference*, CoNEXT '07, page 12:1–12:12. ACM, 2007.

[CMRM07] R. Calegari, M. Musolesi, F. Raimondi, and C. Mascolo. CTG: a connectivity trace generator for testing the performance of opportunistic mobile systems. In *Proc. 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on the foundations of software engineering*, pages 415–424. ACM, 2007.

[CSHS12] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proc. 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[CVdBB+10] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7):e11596, 2010.

[DFL01] J. A. Davis, A. H. Fagg, and B. N. Levine. Wearable computers as packet transport mechanisms in highly-partitioned ad-hoc networks. In *Proc. 5th International Symposium on Wearable Computers*, pages 141–148, 2001.

[DH07] E. M. Daly and M. Haahr. Social network analysis for routing in disconnected delay-tolerant MANETs. In *Proc. 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, page 32–40. ACM, 2007.

[EKKO08] F. Ekman, A. Keränen, J. Karvo, and J. Ott. Working day movement model. In *Proc. 1st ACM SIGMOBILE workshop on Mobility models*, MobilityModels '08, page 33–40. ACM, 2008.

[EP06] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[EP09] N. Eagle and A. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[EPL09] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15274–15278, 2009.

[For10] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[FWYC10] M. P. Freeman, N. W. Watkins, E. Yoneki, and J. Crowcroft. Rhythm and randomness in human contact. In *Proc. International Conference on Advances in Social Networks Analysis and Mining*, 2010.

[GHB08] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[GKdPC12] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Next place prediction using mobility markov chains. In *Proc. 1st Workshop on Measurement, Privacy, and Mobility (MPM)*, page 3:1–3:6. ACM, 2012.

[GWH07] S. A. Golder, D. Wilkinson, and B. Huberman. Rhythms of social interaction: messaging within a massive online network. In *Proc. 3rd Int. Conf. on Communities and Technologies*, 2007.

[Har90] A. C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, 1990.

[HC08] P. Hui and J. Crowcroft. Human mobility models and opportunistic communications system design. *Philosophical Transactions of the Royal Society A*, 366(1872):2005–2016, 2008.

[HCS+05] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and Christophe Diot. Pocket switched networks and human mobility in conference environments. In *Proc. 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, WDTN '05, page 244–251. ACM, 2005.

[HCY11] P. Hui, J. Crowcroft, and E. Yoneki. BUBBLE rap: social-based forwarding in delay-tolerant networks. *IEEE Transactions on Mobile Computing*, 10(11):1576–1589, 2011.

[HKA08] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. *Computer Networks*, 52(14):2690–2712, 2008.

[HL70] P. W. Holland and S. Leinhardt. A method for detecting structure in sociometric data. *American Journal of Sociology*, 76(3):492–513, 1970.

[HR99] D. M. Halliday and J. R. Rosenberg. Time and frequency domain analysis of spike train and time series data. In *Modern Techniques in Neuroscience Research*, pages 503–543. Springer Berlin Heidelberg, 1999.

[HYCC07] P. Hui, E. Yoneki, S.-Y. Chan, and J. Crowcroft. Distributed community detection in delay tolerant networks. In *Proc. 2nd ACM/IEEE international workshop on mobility in the evolving internet architecture*, pages 1–8. ACM, 2007.

[HYCC09] P. Hui, E. Yoneki, J. Crowcroft, and S.-Y. Chan. Identifying social communities in complex communications for network efficiency. In *Proc. 1st international conference on complex sciences: theory and applica-*

tions (COMPLEX), pages 351–363, 2009.

[ISB+11] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of Theoretical Biology*, 271(1):166–180, 2011.

[Jan10] W. Jang. Travel time and transfer analysis using transit smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, 2144:142–149, 2010.

[JFG12] S. Jiang, J. Ferreira, and M. González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, pages 1–33, 2012.

[JLSZ08] H. Jeung, Q. Liu, H. T. Shen, and X. Zhou. A hybrid prediction model for moving objects. In *Proc. IEEE 24th International Conference on Data Engineering (ICDE)*, pages 70–79, 2008.

[JOW+02] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. S. Peh, and D. Rubenstein. Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with ZebraNet. *SIGARCH Computer Architecture News*, 30(5):96–107, 2002.

[Kab10] K. Kabassi. Personalizing recommendations for tourists. *Telematics and Informatics*, 27(1):51–66, 2010.

[KCA+09] T. Kreuz, D. Chicharro, R. G. Andrzejak, J. S. Haas, and H. D. I. Abarbanel. Measuring multiple spike train synchrony. *Journal of Neuroscience Methods*, 183(2):287–299, 2009.

[KCGA11] T. Kreuz, D. Chicharro, M. Greschner, and R. G. Andrzejak. Time-resolved and time-scale adaptive measures of spike train synchrony. *Journal of Neuroscience Methods*, 195(1):92–106, 2011.

[KS99] H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge University Press, 1999.

[LB08] M. Lahiri and T. Y. Berger-Wolf. Mining periodic behavior in dynamic social networks. In *Proc. 8th IEEE International Conference on Data*

Mining (ICDM), pages 373–382, 2008.

[LB09] M. Lahiri and T. Berger-Wolf. Periodic subgraph mining in dynamic networks. *Knowledge and Information Systems*, 24(3):467–497, 2009.

[LC11] N. Lathia and L. Capra. How smart is your smartcard?: measuring travel behaviours, perceptions, and incentives. In *Proc. 13th international conference on Ubiquitous computing*, UbiComp '11, page 291–300. ACM, 2011.

[LDH+10] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. In *Proc. 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, page 1099, 2010.

[LDS04] A. Lindgren, A. Doria, and O. Schelén. Probabilistic routing in intermittently connected networks. In *Proc. Service Assurance with Partial and Intermittent Resources (SAPIR)*, pages 239–254, 2004.

[LFC10] N. Lathia, J. Froehlich, and L. Capra. Mining public transport usage for personalised intelligent transport systems. In *Proc. 10th IEEE International Conference on Data Mining (ICDM)*, ICDM '10, page 887–892. IEEE Computer Society, 2010.

[LGA+12] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Proc. Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing*, 2012.

[LHK+09] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. SLAW: a new mobility model for human walks. In *Proc. 28th IEEE Conference on Computer Communications (INFOCOM)*, pages 855–863, 2009.

[LLS+06] J. Leguay, A. Lindgren, J. Scott, T. Friedman, and J. Crowcroft. Opportunistic content distribution in an urban setting. In *Proc. 2006 SIGCOMM workshop on Challenged networks*, CHANTS '06, page 205–212. ACM, 2006.

[LM01] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001.

[LPdR06] J. Lawrence, T. R. Payne, and D. de Roure. Co-presence communities: using pervasive computing to support weak social networks. In *Proc. 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 149–156, 2006.

[LRT04] K. Laasonen, M. Raento, and H. Toivonen. Adaptive on-device location recognition. In *Pervasive Computing*, pages 287–304. 2004.

[LW07] F. Li and Y. Wang. Routing in vehicular ad hoc networks: A survey. *Vehicular Technology Magazine, IEEE*, 2(2):12–22, 2007.

[MCK+04] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W. Cheung. Mining, indexing, and querying historical spatiotemporal data. In *Proc. 11th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, page 236–245. ACM, 2004.

[MGC+07] A. Miklas, K. Gollu, K. Chan, S. Saroiu, K. Gummadi, and E. de Lara. Exploiting social interactions in mobile systems. In *UbiComp 2007: Ubiquitous Computing*, pages 409–428. 2007.

[MHM05] M. Musolesi, S. Hailes, and C. Mascolo. Adaptive routing for intermittently connected mobile ad hoc networks. In *Proc. 6th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 183–189, 2005.

[MHN07] B. Mehta, T. Hofmann, and W. Nejdl. Robust collaborative filtering. In *Proc. 2007 ACM conference on Recommender systems*, RecSys '07, page 49–56. ACM, 2007.

[MM06a] M. Musolesi and C. Mascolo. A community based mobility model for ad hoc network research. In *Proc. 2nd international workshop on Multi-hop ad hoc networks: from theory to reality (REALMAN)*, page 31–38. ACM, 2006.

[MM06b] M. Musolesi and C. Mascolo. Evaluating context information predictability for autonomic communication. In *Proc. 2006 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 495–499. IEEE Computer Society, 2006.

[MM07] M. Musolesi and C. Mascolo. Designing mobility models based on social network theory. *SIGMOBILE Mob. Comput. Commun. Rev.*, 11(3):59–70, 2007.

[MM09] M. Musolesi and C. Mascolo. CAR: context-aware adaptive routing for delay-tolerant mobile networks. *IEEE Transactions on Mobile Computing*, 8(2):246–260, 2009.

[MMC08] L. McNamara, C. Mascolo, and L. Capra. Media sharing based on colocation prediction in urban transport. In *Proc. 14th ACM International Conference on Mobile Computing and Networking*, pages 58–69. ACM, 2008.

[MMC09] A. J. Mashhadi, S. B Mokhtar, and L. Capra. Habit: leveraging human mobility and social network for efficient content dissemination in MANETs. In *Proc. 10th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 214–219, 2009.

[MRM+10] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[MRM12] W. Mathew, R. Raposo, and B. Martins. Predicting future locations with hidden markov models. In *Proc. 2012 ACM Conference on Ubiquitous Computing (UbiComp)*, UbiComp '12, page 911–918. ACM, 2012.

[MSRJ12] J. McInerney, S. Stein, A. Rogers, and N. R. Jennings. Exploring periods of low predictability in daily life mobility. In *Proc. Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing*, 2012.

[MV05] M. McNett and G. M. Voelker. Access and mobility of wireless PDA

users. *SIGMOBILE Mob. Comput. Commun. Rev.*, 9(2):40–55, 2005.

[NKM03] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos. A data mining algorithm for generalized web prefetching. *IEEE Transactions on Knowledge and Data Engineering*, 15(5):1155–1169, 2003.

[NSLM12] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: new venue recommendation in location-based social networks. In *Proc. 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (SOCIALCOM-PASSAT)*, SOCIALCOM-PASSAT '12, page 144–153. IEEE Computer Society, 2012.

[NTM+12] V. Nicosia, J. Tang, M. Musolesi, G. Russo, C. Mascolo, and V. Latora. Components in time-varying graphs. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(2):023101–023101–11, April 2012.

[PBV07] Gergely Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[PD09] A.-K. Pietilainen and C. Diot. Experimenting with opportunistic networking. In *Proc. International Workshop on Mobility in the Evolving Internet Architecture (MobiArch)*, 2009.

[Pen07] A. Pentland. Automatic mapping and modeling of human networks. *Physica A: Statistical Mechanics and its Applications*, 378(1):59 – 67, 2007.

[PKK+12] V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. M. Dunbar. Sex differences in intimate relationships. *Nature Scientific Reports*, 2, 2012.

[PP06] A. N. Pavlov and O. N. Pavlova. Wavelet analysis of the structure of point processes. *Springer Technical Physics Letters*, 32(11):918–921, 2006.

[PPC06] L. Pelusi, A. Passarella, and M. Conti. Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. *Communications Magazine, IEEE*, 44(11):134–141, 2006.

[QC09] D. Quercia and L. Capra. FriendSensing: recommending friends using mobile phones. In *Proc. 3rd ACM conference on Recommender systems*, RecSys '09, page 273–276. ACM, 2009.

[QLC+10] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *Proc. 10th IEEE International Conference on Data Mining (ICDM)*, pages 971 –976, 2010.

[RBY+01] O. A. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann, and E. Başar. Wavelet entropy: a new tool for analysis of short duration brain electrical signals. *Journal of Neuroscience Methods*, 105(1):65–75, 2001.

[SCM+06] J. Su, K. K. W. Chan, A. G. Miklas, K. Po, A. Akhavan, S. Saroiu, E. de Lara, and A. Goel. A preliminary investigation of worm infections in a bluetooth environment. In *Proc. 4th ACM Workshop on Recurring Malcode (WORM)*, page 9–16. ACM, 2006.

[SGC+06] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. *Haggle Cambridge dataset at CRAWDAD*. Available at http://crawdad.cs.dartmouth.edu/cambridge/haggle, 2006.

[SK05] N. Samaan and A. Karmouch. A mobility prediction architecture based on contextual knowledge and spatial conceptual maps. *Mobile Computing, IEEE Transactions on*, 4(6):537–551, 2005.

[SKJH06] L. Song, D. Kotz, R. Jain, and X. He. Evaluating next-cell predictors with extensive Wi-Fi mobility data. *IEEE Transactions on Mobile Computing*, 5(12):1633–1649, 2006.

[SMM+11] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell. NextPlace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, volume 6696, pages 152–169. 2011.

[SMML10] S. Scellato, M. Musolesi, C. Mascolo, and V. Latora. On nonstationarity of human contact networks. In *Proc. 2nd Workshop on Simplifying Complex Networks for Practitioners (SIMPLEX 2010)*, 2010.

[SNLM11] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *Proc. 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[SPG06] N. Sarafijanovic-Djukic, M. Piorkowski, and M. Grossglauser. Island hopping: efficient mobility-assisted forwarding in partitioned networks. In *Proc. 2006 IEEE Communications Society on Sensor and Ad Hoc Communications and Networks (SECON)*, pages 226–235, 2006.

[SQBB10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[SQC12] D. Saez-Trumper, D. Quercia, and J. Crowcroft. Ads and the city: considering geographic distance goes a long way. In *Proc. 6th ACM conference on Recommender systems*, RecSys '12, page 187–194. ACM, 2012.

[TMML09] J. Tang, M. Musolesi, C. Mascolo, and V. Latora. Temporal distance metrics for social network analysis. In *Proc. 2nd ACM Workshop on Online Social Networks*, pages 31–36. ACM, 2009.

[TMML10] J. Tang, M. Musolesi, C. Mascolo, and V. Latora. Characterising temporal distance and reachability in mobile and online social networks. *SIGCOMM Comput. Commun. Rev.*, 40(1):118–124, 2010.

[TSM+10] J. Tang, S. Scellato, M. Musolesi, C. Mascolo, and V. Latora. Small-world behavior in time-varying graphs. *Physical Review E*, 81(5):055101, 2010.

[VdBCB+10] W. Van den Broeck, C. Cattuto, A. Barrat, M. Szomszor, G. Correndo, and H. Alani. The live social semantics application: a platform for integrating face-to-face presence with on-line social networking. In *Proc. 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 226–231, 2010.

[WGHB09] P. Wang, M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding the spreading patterns of mobile phone viruses. *Science*, 324(5930):1071–1076, 2009.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[WWA12a] M. J. Williams, R. M. Whitaker, and S. M. Allen. Decentralised detection of periodic encounter communities in opportunistic networks. *Ad Hoc Networks*, 10(8):1544–1556, 2012.

[WWA12b] M. J. Williams, R. M. Whitaker, and S. M. Allen. Measuring individual regularity in human visiting patterns. In *Proc. 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (SOCIALCOM-PASSAT)*, 2012.

[YG10] E. Yoneki and D. Greenfield. Inferring significance of meeting groups in human contact networks. In *Proc. 2010 European Conference on Complex Systems (ECCS 2010)*, 2010.

[YKUM05] G. Yavas, D. Katsaros, O. Ulusoy, and Y. Manolopoulos. A data mining approach for location prediction in mobile environments. *Data Knowl. Eng.*, 54(2):121–146, 2005.

[YYL10] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *Proc. 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, page 458–461. ACM, 2010.

[ZBY+12] X. Zhu, Y. Bai, W. Yang, Y. Peng, and C. Bi. SAME: a students' daily activity mobility model for campus delay-tolerant networks. In *Proc. 18th Asia-Pacific Conference on Communications (APCC)*, pages 528–533, 2012.

[ZCZ+10] V. W. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang. Collaborative filtering meets mobile recommendation: a user-centered approach. In *Proc. 24th AAAI Conference on Artificial Intelligence*, March 2010.

[ZZXM09] Y. Zheng, L. Zhang, X. Xie, and W-.Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. In *Proc. 18th international conference on World wide web*, WWW '09, page 791–800. ACM, 2009.