UNIVERSITY OF CALIFORNIA
Santa Barbara

Scalable approaches to communication and inference: Minimalistic strategies for measurement and coordination

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

by

Dineshkumar Karuppanna Gounder Ramasamy

Committee in Charge:

Professor Upamanyu Madhow, Chair

Professor João P. Hespanha

Professor Kenneth Rose

Professor Subhash Suri

December 2014

The Dissertation of
Dineshkumar Karuppanna Gounder Ramasamy is approved:

Professor João P. Hespanha

Professor Kenneth Rose

Professor Subhash Suri

Professor Upamanyu Madhow, Committee Chairperson

November 2014

Scalable approaches to communication and inference: Minimalistic strategies for measurement and coordination

by

Dineshkumar Karuppanna Gounder Ramasamy

To Amma, Appa and Nisha.

Acknowledgements

I consider myself lucky to have had a wonderful advisor and mentor in Professor Madhow. His vision and thoughtful comments have shaped much of this thesis and I am very grateful for that. His feedback on presentations and writing has changed the way I organize my thoughts when presenting to an audience. I thank Professors João Hespanha, Kenneth Rose and Subhash Suri for serving on my committee. My colleagues Andrew, Aseem, Eric, Francois, Hong, Sandeep, Sriram, Sumit, Babak, Maryam, Faruk, Hossein and Zhinous have been a fun group to work with and I sincerely appreciate their help over the years. I specifically thank Sriram for the numerous occasions he went out of his way to help me. I thank Professors Andrew Thangaraj and Anurag Mittal at IIT Madras for giving me opportunities to sample research as an undergraduate student.

Curriculum Vitæ
Dineshkumar Karuppanna Gounder Ramasamy

Education

November 2014	Doctor of Philosophy, Electrical and Computer Engineering University of California, Santa Barbara
March 2011	Master of Science, Electrical and Computer Engineering University of California, Santa Barbara
July 2009	Bachelor of Technology, Electrical Engineering Indian Institute of Technology Madras

Publications

Journals

- D. Ramasamy and U. Madhow. "Scalable and Efficient Geographic Routing in Mobile Ad Hoc Wireless Networks," *IEEE Transactions on Information Theory (submitted)*
- D. Ramasamy, S. Venkateswaran and U. Madhow. "Compressive Parameter Estimation in AWGN," *IEEE Transactions on Signal Processing*, April 2014

Conferences

- D. Ramasamy, S. Venkateswaran, and U. Madhow. "Inferring user interests from tweet times," *ACM Conference on Online Social Networks* (COSN), October 2013, Boston, MA
- D. Ramasamy, R. Ganti, and U. Madhow. "On the capacity of picocellular networks," *IEEE International Symposium on Information Theory*, July 2013, Istanbul, Turkey.
- D. Ramasamy, S. Venkateswaran, and U. Madhow. "Compressive estimation in AWGN: General observations and a case study," *Asilomar Conference on Signals, Systems and Computers*, November 2012, Asilomar, CA
- D. Ramasamy, S. Venkateswaran, and U. Madhow. "Compressive tracking with 1000-element arrays: A framework for multi-gbps mm wave cellular downlinks," *Allerton Conference on Communication, Control and Computing* (invited paper), October 2012, Monticello, IL
- D. Ramasamy and U. Madhow, "Can geographic routing scale when nodes are mobile?," *IEEE International Symposium on Information Theory*, July 2012, Boston, MA

- D. Ramasamy, S. Venkateswaran, and U. Madhow, "Compressive adaptation of large steerable arrays," *Information Theory and Applications Workshop* (ITA) (invited paper), Feb 2012, San Diego, CA

Abstract

Scalable approaches to communication and inference: Minimalistic strategies for measurement and coordination

Dineshkumar Karuppanna Gounder Ramasamy

Recent advances in technology have enabled large-scale systems in a wide variety of settings. We consider three examples which illustrate that by carefully reducing the size of problem from the very outset, we can provide efficient solutions to the engineering difficulties of communication and inference at scale.

The first example we provide is the problem of estimating a small number of continuous valued parameters from a high dimensional signal. Random projections have been shown to be effective in capturing sparse signals with few measurements. We study the effect of random projections for parameter estimation by focusing on two lower bounds on estimation error variance, the Ziv-Zakai bound (ZZB) and the Cramér-Rao bound (CRB). They reveal that when we ensure that *certain* geometries (shapes) in the signal manifold (the set of all possible values that the signal can take) are preserved, these bounds are also preserved up to an SNR penalty equal to the dimensionality reduction factor. We show how the SNR penalty results when viewed in conjunction with the threshold behavior of the ZZB can be used to accurately predict the *minimum* number of random

projections needed to avoid large estimation errors. We take up the problem of frequency estimation to illustrate the above ideas and give: (i) number of random projections needed to preserve the identified geometries and (ii) an algorithm which attains the CRB when the number of random projections exceeds the ZZB based prediction.

The second problem we consider is that of geographic routing in mobile ad hoc networks. In order to forward a packet, a relay node needs position estimates of its neighbors and the destination. When the number of nodes in the network grows, the overhead required to inform far away nodes of changes in one's position overwhelms the network. We address this problem by directing updates to only a small *subset* of nodes in the network. Accordingly, we give a routing protocol that accommodates scenarios where the source and/or relay nodes do not have estimates of the destination's location. The protocol parameters are chosen to guarantee that the overhead fits within network resources and to ensure that the length of every routing trajectory is bounded within a constant factor of the shortest path. We verify these analytical design prescriptions using simulations.

The third problem is an exploration of the value of time for efficient inference on Twitter. Twitter as an online social medium is defined by its real-time nature. We therefore ask whether interests (e.g., sport fandom) can be identified from a user's *tweet times alone* by exploiting the known timing of "events" associated

with the interest (e.g., game times of the sport team). Our results indicate that tweet times can be used to make reliable inferences on the interests of a large number of users. With a view to automate event identification, we develop an inference framework with minimal measurement (time of tweets) and processing requirements for accurately determining when a topic trends on Twitter from an aggregate Twitter feed related to the topic. We then dissect the identified trending times into groups, each associated with a different *subtopic*, using only userIDs of tweets made during trending times based on the intuition that when two events/trending-times share a large fraction of users, they likely correspond to the same subtopic. Our results from Twitter data obtained over six months illustrate that significant insights into topic-specific Twitter activity can be obtained within our frugal measurement and processing framework. These results suggest that time can be a compact and effective cue to cull data from massive online streams like Twitter.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many current and emerging engineered systems dwarf their predecessors in scale. This calls for minimalistic design approaches that extract the essential features of the problem at hand. Adopting such designs, enables us to decrease the problem scale to manageable levels, thereby giving room to carefully designed algorithms that limit the sacrifices in optimality due to this reduction in scale. In this thesis, we consider three examples illustrating such approaches:

- *Physical layer:* Channel estimation for 60GHz systems, where we give an architecture that sidesteps the difficulties that arise as a result of hardware scaling by restricting the nature and number of measurements that we take,

- *Network layer:* Maintaining routing state information for geographic routing in large mobile ad hoc wireless networks, where the resulting overhead can overwhelm the network. We propose a solution which drastically reduces

the coordination overhead by carefully choosing a minimal set of users to inform and

- *Application layer:* Inference on social networks, for which the vast scale of data that needs digesting motivates us to investigate the purchase from a compact and ubiquitous source of information, the time when people *act* (specifically, when a person tweets on Twitter).

While the problems differ in their details, the solutions we give demonstrate that by reducing the scale of the problem intelligently at the outset, we can avoid the pitfalls of scale, while also providing performance guarantees.

We now provide a summary of the aforementioned problems and state our contributions.

1.1 Compressive parameter estimation in AWGN

This dissertation makes contributions towards the general problem of estimating continuous parameters from random projections perturbed by AWGN. We first present the motivating problem of channel estimation for large 60GHz arrays.

Outdoor 60GHz cellular systems

The ever-growing demand for cellular data has strained existing wireless networks to their limits. We see the evidence for this in the finely tiered data packages offered by cellular service providers. This crunch in wireless resources is particularly severe in urban environments, where the density of users makes current approaches to tackling this shortage, namely increasing the density of cellular basestations unsustainable.

The wireless spectrum around the cellular bands ($\approx$ 2GHz) has mostly been taken up by civilian and defense applications. An orders of magnitude larger *unlicensed* spectrum is however available in the 60 GHz band (57-64 GHz in North America). With the emergence of commercial 60GHz indoor technologies, a potential approach to meet the cellular "capacity crunch" in outdoor urban scenarios is to use the 60GHz spectrum. This can for example be realized by tapping into the dense wireline backbone networks that are already available in urban settings.

Coarse RF beamforming

The rule of thumb for the area of an antenna element is λ^2, where λ refers to the wavelength of the EM wave ($\lambda \propto 1/f_c$; f_c is the carrier frequency). Therefore, at

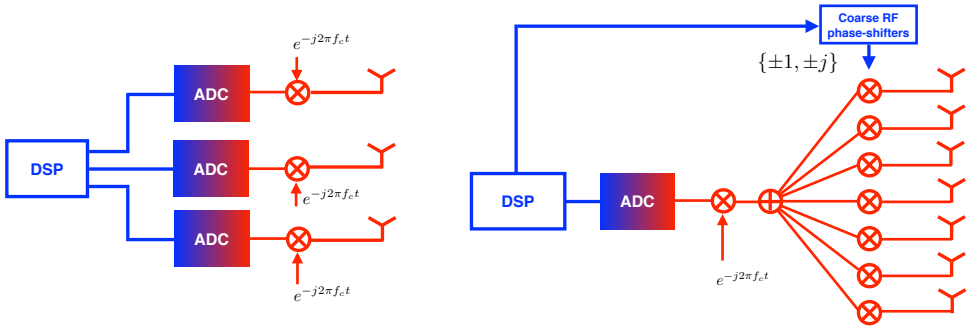

Figure 1.1: Left: Conventional beamforming with fine digital access to each antenna element; Right: Coarse RF beamforming with $\{\pm 1, \pm j\}$ phase shifters for each antenna element.

higher frequencies (smaller wavelengths), the amount of power that an individual antenna element can collect scales as λ^2 and takes an order of magnitude hit (by a factor of about a thousand when compared to the 2GHz band). An upside to this, however, is that we can now pack many antennas on a small area: for e.g., it is possible to fit a 32×32 (1024 element) array on the area equivalent to the palm of one's hand (a 6cm $\times$ 6cm patch). By employing such physically small, yet *electrically large* arrays we can compensate for the loss in power due to small antenna sizes at 60GHz. Such gains from multi-antenna systems are referred to as *beamforming gains*.

Current multi-antenna communication systems realize beamforming gains by dedicating a separate analog-to-digital converter (ADC) for each antenna and process the received signals digitally. By doing so, they ensure that the signals from the different antennas add constructively. We depict such an architecture

in Figure 1.1 (left). While this is feasible when antennas are few in number, this architecture, which dedicates a power-hungry ADC to each antenna element does not scale to the electrically large antenna arrays necessary for outdoor urban cellular systems at 60GHz.

We consider an approach which addresses this challenge by means of a hybrid minimalistic philosophy, wherein digitally controlled radio-frequency (RF) phase shifters are used to modulate the gain of each antenna element in a coarse manner. Following this, a *single* baseband chain (sum of the antenna responses) emerges from the array. This architecture is sketched in Figure 1.1 (right) and is referred to as RF beamforming.

Spatial channel estimation

It is desirable that we accommodate the following unique features of this estimation problem:

- Channel measurements are obtained by coarse phase control of each antenna. As a result, the kind of measurements that can be used for channel estimation is limited (if N is the number of antenna elements, the measurements are constrained to be projections onto vectors in $\{\pm 1, \pm j\}^N$).

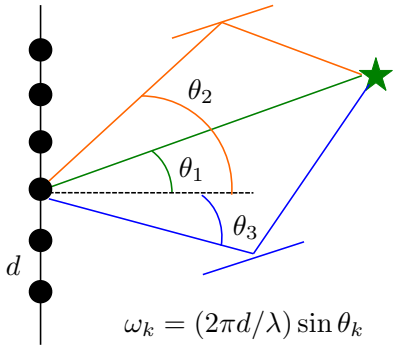

Figure 1.2: 60GHz spatial channel seen by a d-spaced uniform linear array is a mixture of sinusoids. The frequency of the sinusoid corresponding to each path is given by $\omega = 2\pi(d/\lambda)\sin\theta$, where θ is the angle of arrival of the path, a continuous quantity

- Even though the number of antennas is large, it is desirable that the number of measurements used for estimating the channel is minimal. This allows us to adapt to mobility in outdoor cellular systems.

- It has been empirically observed that 60GHz channels are sparse. i.e., they are composed of a few paths (say K). The channel response for each path can be concisely described by a few continuous parameters. For the example of the d-spaced linear array in Figure 1.2, they are the gains $\{g_l\}$ and the "spatial frequencies" $\{\omega_l\}$ of the K paths. Specifically, the channel response $x(n)$ seen by the n-th antenna element is given by

$$x(n) = \sum_{l=1}^{l=K} g_l e^{j\omega_l n}. \tag{1.1}$$

  The resultant channel is thus a mixture of sinusoids.

One approach to estimating the 60GHz channel could be to explicitly estimate this natural compact *parameterization* of the channel given by $\boldsymbol{\theta} = (g_1, \omega_1, \ldots, g_K, \omega_K)$, rather than the channel seen by each antenna element $\mathbf{x}(\boldsymbol{\theta}) = [x(1), \ \ldots, x(N)]^T$ individually (which entails estimating $N \gg 2K$ parameters).

Random projections are a natural fit for this estimation scenario because: (i) they are agnostic to the constraint that the measurement weights must come from a restricted set and (ii) they have been shown to be successful in exploiting sparsity and capturing information with a small number of measurements. Therefore, we investigate such *compressive* channel estimation strategies which take as input random projections of the channel response and estimate the parameters $\boldsymbol{\theta}$ directly.

We layout the notation we use to state our results: We want to estimate a K dimensional parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ from M random projections of an N dimensional signal $\mathbf{x}(\boldsymbol{\theta})$ in AWGN. We denote the compressive measurement matrix containing the random projection weights by $\boldsymbol{A}$.

Contributions

In this dissertation, we make the following contributions to general parameter estimation from compressive measurements (random projections) perturbed by

Additive White Gaussian Noise (AWGN), which we apply to our channel estimation setting:

- Bounds on the estimation error variance, such as the Ziv-Zakai bound (ZZB) and the Cramér Rao bound (CRB), relate the geometry of the signal manifold to the best achievable performance. We show that the ZZB is preserved up to an SNR penalty equal to the dimensionality reduction factor M/N when pairwise distances are approximately preserved. i.e.,

$$\|\boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}_1) - \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}_2)\| \approx C\,\|\mathbf{x}(\boldsymbol{\theta}_1) - \mathbf{x}(\boldsymbol{\theta}_2)\| \;\; \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$$

  We show that under weaker geometry preservation conditions (pairwise distances in the limit of $\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2$), the CRB is preserved up to the same SNR penalty of M/N.

- Non-linear estimation problems exhibit a threshold behavior with SNR and this threshold behavior is closely mirrored by the ZZB. We use the preceding SNR penalty results, which relate the number of compressive measurements with the effective SNR to give a *numerical* prediction of the minimum number of random projections needed to avoid gross estimation errors.

Towards the motivating problem of estimating sparse 60GHz channels from coarse phase shifts, this dissertation makes the following contributions ($\mathbf{x}(\boldsymbol{\theta})$ is

given by the mixture of sinusoids model in (1.1) and the entries of $\boldsymbol{A}$ are chosen uniformly and independently at random from $\{\pm 1, \pm j\}$):

- We characterize the minimum number of random projections needed to ensure that the geometry of the mixture of sinusoids problem (60GHz channel estimation for uniformly spaced linear arrays) is approximately preserved. While our characterization is complete when the number of sinusoids $K = 1$, we give partial results for $K \geq 2$.

- We provide a low-complexity algorithm for frequency estimation which attains the CRB when the number of random projections exceeds our ZZB-threshold based prediction.

While the motivation for our work on compressive frequency estimation has been spatial channel estimation for 60GHz networks, the results in this dissertation have direct implications for other related problems like OFDM channel estimation for wideband channels (those that fit the Tapped Delay Line Channel Model with a few taps).

1.2 Geographic Routing for Mobile Ad Hoc Networks

Conventional wireless networks such as cellular networks and WiFi networks are supported by a wired backhaul. The wireless portion of the network is used to provide last-mile connectivity. In disaster relief scenarios, such wired networks can be compromised. With the ubiquitous penetration of smartphones and tablets, a large and dense ad hoc network of wireless devices equipped with geographic positioning systems may yet be available.

Geographic routing is attractive for such ad hoc networks where the nodes know their own locations. In order to implement geographic routing, a node only requires the locations of its immediate neighbors and of the destination node in order to forward a message. When the nodes in the network can move, a node can maintain estimates of neighbors' locations easily by means of periodic local broadcasts. As the network scales, the overhead corresponding to global dissemination of information regarding the locations of moving destination nodes, however, becomes a bottleneck.

In this dissertation, we draw inspiration from prior work[7] where it was observed that it is possible to reduce the frequency of position updates to nodes which are farther away from the destination node without adversely affecting the

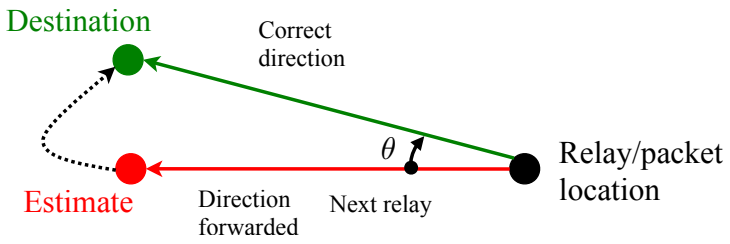

Figure 1.3: Rate of progress depends on the angle between the forwarded and correct directions (marked θ). When the distance between the destination and the packet location is large, the position estimate used can disagree with the true location by a proportionally larger amount while maintaining the quality of the routing decision made at the relay node (quantified by θ).

quality of routing decisions (this simple intuition is presented in Figure 1.3). We analyze the scalability of geographic routing protocols built on top of this basic intuition. For such protocols two competing objectives need to be addressed. They are the following:

- **Scalability:** We need to ensure that the volume of position updates does not exceed the amount that can be sustained using a fixed bandwidth (a portion of the available bandwidth can be preallocated for such overhead purposes). This necessitates that we reduce the volume of position updates made.

- **Reliability & Efficiency:** When we reduce the volume of updates, routing decisions made at each intermediate node are "approximate". We need guarantees that such approximation errors do not "add up" and that packets

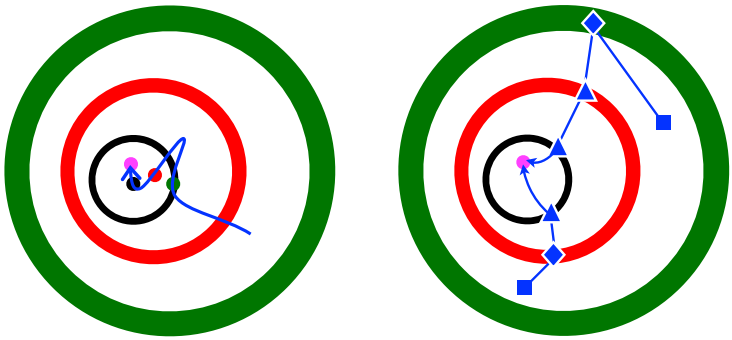

Figure 1.4: Left: Update rings corresponding to three consecutive ring indices. The position estimates (also centers of the update ring) are marked • on the destination's trajectory using corresponding colors. Right: Two packet trajectories (blue) starting from nodes marked ■ converging to the destination (magenta •).

are successfully delivered to their destinations. Further, it is necessary to quantify the loss in efficiency of the routing trajectories constructed using lax position-publish strategies (a measure of the efficiency of a route could be the length of the packet trajectory relative to length of the straight line from source to destination).

We give a scalable solution that provides the system designer with the choice of deciding where she wants to be in the above trade-off between overhead (bandwidth needed) and efficiency.

Our scalability and reliability computations assume that all nodes in the network execute Brownian motion with the same mean square velocity (or a random walk on a fine 2D lattice). While Brownian motion as a model for short-term mobility may seem unrealistic, we note that it is the movements over large time

scales which matter for both scalability and reliability. Using law of large numbers arguments, one can make the case that many local models for mobility "look like" Brownian motion when viewed over a sufficiently large time scale.

Contributions

The contributions made in this dissertation towards addressing the scalability and efficiency of geographic routing in mobile ad hoc networks are summarized as follows:

- A measure of the suboptimality of a routing decision is the angle between the correct direction of the destination and the direction along which a packet is forwarded (θ in Figure 1.3). One approach to reducing the position update overhead is to reduce the frequency of updates to nodes farther away, while also ensuring the angular "error" for *all* routing decisions is bounded. In order to do this, one needs to send position update to every node in the network. We show that the position-publish overhead needed to sustain such a scheme cannot be supported with fixed bandwidth.

- The position-publish protocol that we propose circumvents this scalability bottleneck by sending position updates to strictly a subset of the nodes in the network. Specifically, updates are made to annular regions centered

around the destination node (Figure 1.4 left). Counting outward from the destination node, the radii, thickness and lifetimes of the updates all grow exponentially with the ring index, but their exponents differ. We determine the constraints that these exponents need to satisfy for scalable position-publish.

- All nodes in the network do not possess position estimates corresponding to the destination node. Therefore, the routing protocol appends the "best" estimate of the destination node seen so far to the packet. Subsequent relay nodes use this estimate to make their forwarding decisions unless they have a better estimate of the destination node, in which case they overwrite the packets's estimate with theirs. The ring index corresponding to the update is used to decide whether the estimate is a better estimate or not. Representative packet trajectories corresponding to this routing strategy are sketched in Figure 1.4 (right). In order to ensure routing reliability, a packet must latch-on to the update corresponding to an update ring when it is relayed through it (we illustrate this for the trajectories drawn in Figure 1.4 right, where packets refine their estimates when they pass through the ring). Since, all nodes in the network are mobile, an update which we make to a ring tends to *diffuse* away with time. We determine the conditions that the

protocol parameters must satisfy in order to combat diffusion and ensure reliable routing.

- We identify constraints on the protocol parameters which limit the potential geometry of the update rings around each destination node to guarantee that the efficiency of routes constructed using the proposed routing protocol is bounded (from below) with high probability, even though a large fraction of the nodes in the network do not posses position estimates corresponding to the destination. We quantify routing efficiency by means of the reciprocal of the *route stretch*, which we define to be the ratio of the length of the packet trajectory to the shortest distance between the source and the destination and give guarantees on routing efficiency.

1.3 Inference from time on Twitter

Twitter as a medium for social interactions is defined by its real-time nature. Such timely interactions have been enabled by social media innovations like hashtags pioneered by Twitter. As a result, we expect the *timing* of tweets to be a useful signal for inference on Twitter. Existing inference methods that use tweet times also use the rich body of information in a tweet (like text, location, etc.,). For e.g., an early Twitter based news aggregator used text and author reputa-

tion scores for each tweet in addition to tweet times. Such systems, which pool information from multiple sources do not give us an understanding of the value inherent in time.

This dissertation explores two closely related problems with the view of clarifying the value of time for identifying user interests. The first is a study of whether the times when clearly demarcated events related to a topic transpire in the real world can be used to identify users who are interested in the topic. For the example of identifying fans of a sport team, these "event times" can be the times when this team plays its games. The second aims to discover and group trending/event times using aggregate feeds related to the topic (we use one such aggregate feed, obtained by asking Twitter to return the stream of public tweets containing a few keywords related to the topic). These groups of event times can then be used to identify users interested in the topic. We now summarize the contributions made in this dissertation for the two problems.

Contributions

This dissertation makes the following contributions towards understanding the value of time as a tool for inference on Twitter:

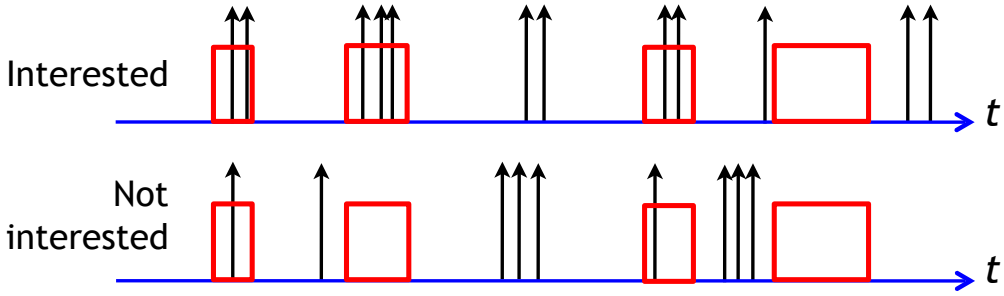

Figure 1.5: Tweet times of the user marked by arrows. Event times are marked in red. Top: Tweeting behavior of a person interested in the topic. Bottom: A person not interested

- **Interest from tweet times:** We expect users interested in a topic to tweet more often during associated event times than other times (sketched in Figure 1.5). We propose a simple Bayesian model for the tweet times of a user that incorporates this intuition and derive a statistic for interest in a topic. This statistic is merely a function of three *numbers*: (a) number of tweets during events, (b) total number of tweets and (c) fraction of the overall observation time that corresponds to events. Since people often interact with others who share similar interests, we extend our probabilistic framework to use the interest level estimates of other users with whom a person interacts (by referring to them in his/her tweets). Our results indicate that time can be used to make (computationally) cheap and yet reliable inferences over large user pools.

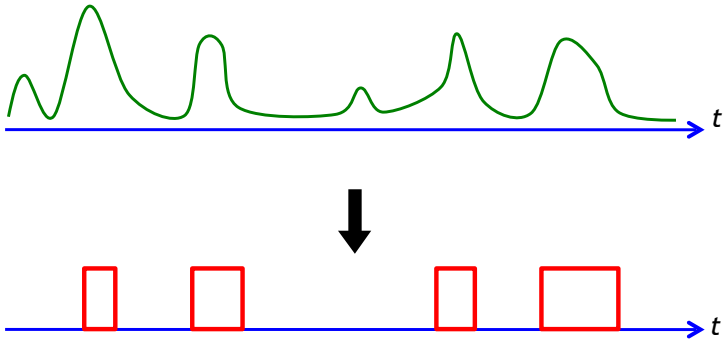

Figure 1.6: Top: The frequency of tweets on an aggregate feed corresponding a topic. Bottom: Event times/Trending times for the topic identified from this feed

- **Event detection from aggregate feeds:** Our notion of trending times/event times are the times when the level of activity on a topic-specific aggregate feed is abnormally high (Figure 1.6). Therefore, in order to identify events, we need a notion of what the baseline activity is. We model baseline activity using a time-varying Poisson process in order to capture the significant time-of-day effects on Twitter. A challenge in estimating baseline parameters is avoiding the upward bias due to event times. We give an algorithm for estimating the baseline which sidesteps this bias by borrowing ideas from outlier detection. We detect events by means of a constant false alarm rate (CFAR) test for increased tweet rate relative to the baseline model. This reduces to a simple comparison of the observed *number of tweets* over a particular time window with the number that one would expect on average due to baseline activity. Our results illustrate that it is possible to reliably

detect interesting activity from an aggregate feed using only the time of tweets.

- **Event clustering:** Once events are detected, we build a bipartite user-event graph, with an edge between a user and an event if the user tweets during the event. We employ spectral partitioning techniques [23, 30] on this graph to cluster the detected event times into sub-topics. We use wordclouds to visualize the results for feeds corresponding to 4 TV shows (using 6 months of observations). They show that in addition to identifying episode air times (as a distinct sub-topic), we discover other commercial and social tie-ins using such minimal means.

1.4 Outline

We review estimation error bounds and present our results on compressive parameter estimation in Chapter 2. We then focus on geometry preservation for compressive frequency estimation in Chapter 3. We present a scalable and efficient geographic routing protocol for MANETs in Chapter 4. In Chapter 5, we investigate inference from time on Twitter. We present our conclusions in Chapter 6.

Chapter 2

Compressive Parameter Estimation in AWGN

We consider the problem of estimating a K dimensional parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ from random projections of an N dimensional signal $\mathbf{x}(\boldsymbol{\theta})$ perturbed by Additive White Gaussian Noise (AWGN). In this chapter we arrive at a *precise* characterization of the number of random projections needed to avoid gross estimation errors.

When we make all N measurements, fundamental bounds on the estimation error variance, such as the Ziv-Zakai bound (ZZB) and the Cramér Rao bound (CRB), relate the geometry of the signal manifold to the best achievable performance. The ZZB depends on pairwise distances of the form $\|\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\| \ \forall \boldsymbol{\theta}, \boldsymbol{\theta}'$,

while the CRB depends on the norms of vectors in the "tangent planes": $\sum_k a_k \left(\partial \mathbf{x}(\boldsymbol{\theta})/\partial \theta_k\right)$, which are essentially pairwise distances $\|\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\|$ in the limit of $\boldsymbol{\theta}' \to \boldsymbol{\theta}$. We explain this in Section 2.3, when we review bounds on estimation error in AWGN.

As a result, when we make compressive measurements, the measurement matrix $\boldsymbol{A}$ enters the expressions for the ZZB and the CRB only via the norms of vectors of the form $\boldsymbol{A}\left(\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\right)$. This observation leads us to the main result of this chapter: We show that if the measurement matrix $\boldsymbol{A}$ satisfies the pairwise isometry property (PIP):

$$\|\boldsymbol{A}\left(\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\right)\| \approx C\,\|\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\|\,, \quad \forall \boldsymbol{\theta}, \boldsymbol{\theta}',$$

the ZZB with compressive measurements is approximately equal to the ZZB with all N measurements, except for an SNR penalty of M/N, where M stands for the number of compressive measurements made (when we scale $\boldsymbol{A}$ so that the noise variance is unaltered, each random projection captures only $(1/N)$-th of the signal energy and this accounts for the M/N SNR penalty). We prove an analogous result for the CRB when $\boldsymbol{A}$ guarantees tangent plane isometry

$$\left\|\boldsymbol{A}\sum_k a_k \left(\partial \mathbf{x}(\boldsymbol{\theta})/\partial \theta_k\right)\right\| \approx C \left\|\sum_k a_k \left(\partial \mathbf{x}(\boldsymbol{\theta})/\partial \theta_k\right)\right\|\,, \quad \forall \boldsymbol{\theta}, a_k,$$

which is a weaker requirement than pairwise isometry. We state these results as Theorems 2.1 & 2.3 in Section 2.4.

When the preceding isometries hold, we can use their relationship to the ZZB/CRB to obtain a tight prediction on the number of measurements necessary for successful compressive estimation. It is known that nonlinear estimation problems exhibit a threshold behavior with the SNR which is closely mirrored by the threshold behavior of the ZZB. We employ this observation to predict the number of measurements required to avoid performance floors, since the effective SNR with compressive measurements increases linearly with the number of measurements. This prescription is given in Section 2.4.4.

Outline: We start by surveying related work in Section 2.1. In Section 2.2, we state the compressive parameter estimation problem in AWGN and the isometry properties needed for successful estimation. In Section 2.3 we review bounds on parameter estimation in AWGN. The relationship between these estimation error bounds (CRB/ZZB) and the isometry properties are brought out in Section 2.4.

2.1 Related work

The goal of standard compressed sensing [14, 18] is to recover signals which are sparse over a finite basis with significantly fewer measurements than the dimension of the observation space. Signal recovery requires that the measurement matrix must satisfy the Restricted Isometry Property (RIP): the distance between any

two sparse signals must be roughly invariant under the action of the matrix. If the RIP is satisfied, sparse signals can be recovered efficiently using techniques such as Orthogonal Matching Pursuit (OMP) and ℓ_1-norm minimization.

The problem of compressive estimation of continuous-valued parameters was perhaps first investigated in [6]. However, it does not relate the pairwise ϵ-isometry property to estimation-theoretic bounds as we do here. Reference [6] also shows that compressive measurements guarantee pairwise ϵ-isometry for a signal manifold with probability $1-\rho$, as long as the number of measurements M satisfies

$$M = O\left(\epsilon^{-2}\log(1/\rho)K\log\left(NVR\tau^{-1}\epsilon^{-1}\right)\right), \tag{2.1}$$

where N and K are the dimensions of the observation and parameter space respectively and V, R, τ are properties of the signal manifold ($1/\tau$ is the condition number which is a generalization of the radius of curvature, R is the geodesic covering regularity and V is the volume). However, to the best of our knowledge, it is difficult to specify how $\{\tau, V, R\}$ scale with the parameters N and K in general. In this dissertation, we provide a self-contained derivation of the number of measurements required to preserve these isometries when the signal manifold consists of a mixture of sinusoids in Section 3.3. Compressive parameter estimation has also been studied in [61]; however, since the noise model there is adversarial, the

results are pessimistic for many practical applications in which a Gaussian model for the noise is a good fit.

The M/N SNR penalty due to compressive measurements has also been noted in [4], but we go further and make the connection between isometries and estimation bounds. Isometries and SNR loss for signal *detection* were considered in [31].

While we focus on compressive estimation based on a finite-dimensional signal, there has been significant research on the processing of continuous time signals exhibiting some measure of sparsity, sometimes termed "finite rate of innovation" (FRI) signals [38]. Sampling strategies for parameter estimation for such signals are studied in [11], using the CRB as the performance metric. The benefits of such good sampling strategies coupled with compressive processing at the analog front end are investigated in [41].

2.2 Compressive measurements

We begin by presenting the model for compressive measurements and providing the intuition behind two isometry conditions that are necessary for successful parameter estimation.

Consider the problem of estimating $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^K$ from noisy measurements of a differentiable manifold $\mathbf{x}(\boldsymbol{\theta}) \in \mathbb{C}^N$. The conventional estimation problem involves measuring all N elements of $\mathbf{x}(\boldsymbol{\theta})$ individually. In vector notation, the measurements are given by:

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\theta}) + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2 \mathbb{I}_N\right). \tag{2.2}$$

In contrast, with compressive measurements, we only observe $M \ll N$ noisy projections of the manifold $\mathbf{x}(\boldsymbol{\theta})$. Therefore, we have

$$\mathbf{y} = \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}) + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2 \mathbb{I}_M\right), \tag{2.3}$$

where $\boldsymbol{A} \in \mathbb{C}^{M \times N}$, which specifies the projection weights, is called the compressive measurement matrix. The elements of $\boldsymbol{A}$ are chosen independently from zero-mean distributions of variance $1/N$ for which certain concentration results (we comment on this later) are available. Examples of such distributions include Uniform$\{\pm 1/\sqrt{N}\}$, Gaussian and Uniform$\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$. When the matrix $\boldsymbol{A}$ satisfies certain isometry conditions, we can successfully estimate $\boldsymbol{\theta}$ from $M \ll N$ measurements. We first explain why these conditions are helpful intuitively and then define them formally.

The Maximum Likelihood (ML) estimator [59] of $\boldsymbol{\theta}$ for the model in (2.3) is given by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}'} \ \|\mathbf{y} - \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}')\| \tag{2.4}$$

$$= \arg\min_{\boldsymbol{\theta}'} \ \|\boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}) - \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}') + \mathbf{z}\| \,. \tag{2.5}$$

If the number of measurements is too small and $\boldsymbol{A}$ has a large nullspace, it is possible that $\|\boldsymbol{A}\left(\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\right)\| \approx 0$ even when $\|\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\|$ is large. Thus, with small amounts of noise $\mathbf{z}$, the optimizing parameter $\widehat{\boldsymbol{\theta}}$ could be drastically different from the true parameter $\boldsymbol{\theta}$, resulting in large errors. This problem can be avoided if the matrix $\boldsymbol{A}$ preserves the geometry of the estimation problem by ensuring that the distance between $\mathbf{x}(\boldsymbol{\theta})$ and $\mathbf{x}(\boldsymbol{\theta}')$ remains approximately unaltered under its action. Specifically, if we have,

$$\|\boldsymbol{A}\left(\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\right)\| \propto \|\mathbf{x}(\boldsymbol{\theta}) - \mathbf{x}(\boldsymbol{\theta}')\| \,, \quad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta, \tag{2.6}$$

we see from (2.5) that the ML estimate at high SNR from M compressive measurements roughly coincides with the estimate we would have obtained with (2.2), where we have access to *all* N measurements of $\mathbf{x}(\boldsymbol{\theta})$. The *pairwise ϵ-isometry property* captures this idea of distance preservation precisely.

Pairwise ϵ-isometry property: The matrix $\boldsymbol{A}$ satisfies the pairwise ϵ-isometry property ($\epsilon < 1$) for the signal model $\mathbf{x}(\boldsymbol{\theta})$ if

$$\sqrt{\frac{M}{N}}(1-\epsilon) \leq \frac{\|\boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}_1) - \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}_2)\|}{\|\mathbf{x}(\boldsymbol{\theta}_1) - \mathbf{x}(\boldsymbol{\theta}_2)\|} \leq \sqrt{\frac{M}{N}}(1+\epsilon), \quad \forall\ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta. \tag{2.7}$$

We now motivate the isometry constants $\sqrt{M/N}(1-\epsilon)$ and $\sqrt{M/N}(1+\epsilon)$. Let $\mathbf{w}_i^H$ denote the i-th row of $\boldsymbol{A}$. Consider a single random projection of a signal $\mathbf{v}$ onto the weights $\mathbf{w}_i$ that have been chosen independently from zero-mean distributions of variance $1/N$. The average energy in the projection is $1/N$ of the energy in the signal $\mathbf{v}$: $\mathbb{E}\left|\mathbf{w}_i^H\mathbf{v}\right|^2 = (1/N)\|\mathbf{v}\|^2$. Thus, M compressive measurements capture M/N of the signal energy on average: $\mathbb{E}\|\boldsymbol{A}\mathbf{v}\|^2 = (M/N)\|\mathbf{v}\|^2$. Thus, for compressive measurements, it is natural to define the pairwise isometry property with the constants $\sqrt{M/N}(1-\epsilon)$ and $\sqrt{M/N}(1+\epsilon)$.

When the elements of $\boldsymbol{A}$ are drawn from appropriate distributions, for any particular realization of the measurement matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\mathbf{v}\|^2$ concentrates around its expected value $(M/N)\|\mathbf{v}\|^2$ with high probability. Specifically, for any $\mathbf{v} \in \mathbb{C}^N$:

$$\Pr\left[\left|\frac{N}{M}\|\boldsymbol{A}\mathbf{v}\|^2 - \|\mathbf{v}\|^2\right| > \delta\right] < C\exp\left(-M\ c(\delta)\right), \tag{2.8}$$

with constants C and $c(\delta)$ that depend only on the distribution from which the elements of $\boldsymbol{A}$ are picked from. For example, when the elements of $\boldsymbol{A}$ are picked i.i.d from Uniform$\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$ or Uniform$\{\pm 1/\sqrt{N}\}$ or $\mathcal{N}(0, 1/N)$, we can

show that $C = 4$ and

$$c(\delta) = \delta^2/4 - \delta^3/6. \tag{2.9}$$

These concentration results are typically used to prove the pairwise isometry property (2.7). Refer [60] for a class of distributions (this includes all sub-gaussian distributions) for which such results are available.

We note that a particular instance of a randomly generated measurement matrix need not satisfy the pairwise isometry property for the signal manifold $\mathbf{x}(\boldsymbol{\theta})$. However, when the number of measurements M is sufficiently large, [6] shows that the pairwise ϵ-isometry property can be satisfied with arbitrarily high probability (the proof involves the use of concentration results (2.8) on carefully chosen samples on the manifold).

A weaker notion of distance preservation is the *tangent plane isometry property* that is particularly useful when we wish to refine an estimate $\widehat{\boldsymbol{\theta}}$ that is "close" to the true parameter value. In this case, since we are interested only in the ML cost surface around the true parameter $\boldsymbol{\theta}$, it suffices to preserve the geometry of the estimation problem in the vicinity of $\boldsymbol{\theta}$ by ensuring that the distances between $\mathbf{x}(\boldsymbol{\theta}')$ and $\mathbf{x}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}' \to \boldsymbol{\theta}$ are preserved under the action of $\boldsymbol{A}$. This is captured by the tangent plane isometry property defined as follows.

Tangent plane ϵ-isometry property: The matrix $\boldsymbol{A}$ satisfies the tangent plane ϵ-isometry property ($\epsilon < 1$) for the signal model $\mathbf{x}(\boldsymbol{\theta})$ if

$$\sqrt{\frac{M}{N}}(1-\epsilon) \leq \frac{\|\boldsymbol{A}\sum a_m \left(\partial \mathbf{x}(\boldsymbol{\theta})/\partial\theta_m\right)\|}{\|\sum a_m \left(\partial \mathbf{x}(\boldsymbol{\theta})/\partial\theta_m\right)\|} \leq \sqrt{\frac{M}{N}}(1+\epsilon)$$
$$\forall\ [a_1, a_2, \ldots, a_K]^T \in \mathbb{R}^K \backslash \{\mathbf{0}\},\ \forall \boldsymbol{\theta} \in \Theta \qquad (2.10)$$

By letting $\boldsymbol{\theta}_2 \to \boldsymbol{\theta}_1$ in the definition of the pairwise ϵ-isometry property, we see that a matrix $\boldsymbol{A}$ which satisfies the pairwise isometry property for the signal model $\mathbf{x}(\boldsymbol{\theta})$ also satisfies the tangent plane isometry, thereby confirming that tangent plane isometry is a weaker notion of distance preservation.

2.3 Parameter estimation in AWGN

We now review classical bounds on parameter estimation in AWGN that we relate to the isometry properties in the next section.

Consider the problem of estimating a parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^K$ from noisy observations of the differentiable manifold $\mathbf{s}(\boldsymbol{\theta}) \in \mathbb{C}^M$. The observations are given by:

$$\mathbf{y} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{z},\ \mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbb{I}_M). \qquad (2.11)$$

For this measurement model,

$$p(\mathbf{y}|\boldsymbol{\theta}) = (\pi\sigma^2)^{-M} \exp\left(\|\mathbf{y} - \mathbf{s}(\boldsymbol{\theta})\|^2/\sigma^2\right). \qquad (2.12)$$

For the observations $\mathbf{y}$, let $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ be an estimate of $\boldsymbol{\theta}$. Given a weight vector $\mathbf{a} \in \mathbb{R}^K$, classical bounds establish lower limits on the error in estimating $\mathbf{a}^T\boldsymbol{\theta}$, given by $\mathbb{E}\left(\mathbf{a}^T\widehat{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{a}^T\boldsymbol{\theta}\right)^2$, for a class of estimators $\widehat{\boldsymbol{\theta}}(\mathbf{y})$. What we have left unspecified is the set of quantities we take the expectation over, and depending on this, the bounds fall into one of two categories:

Deterministic, but unknown, parameters: One class of bounds do not use the prior distribution of $\boldsymbol{\theta}$, so that the parameter to be estimated $\boldsymbol{\theta}$ is best thought of as a deterministic but unknown quantity. The most popular such bound is the Cramér Rao Bound (CRB). For the CRB, the expectation is taken over the conditional distribution $p(\mathbf{y}|\boldsymbol{\theta})$, so that the bound is on $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}\left(\mathbf{a}^T\widehat{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{a}^T\boldsymbol{\theta}\right)^2$. The CRB typically depends on the parameter $\boldsymbol{\theta}$ and the most common version, which is what we use here, applies to estimators $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ which are unbiased (unbiased estimators $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is one that satisfies $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}\left\{\widehat{\boldsymbol{\theta}}(\mathbf{y})\right\} = \boldsymbol{\theta},\ \forall\boldsymbol{\theta}$).

Bayesian bounds: When we know the prior distribution $p(\boldsymbol{\theta})$ from which $\boldsymbol{\theta}$ is chosen, we can incorporate this information into the bounds. Such bounds are called Bayesian bounds and, in these cases, the expectation is taken over the joint distribution $p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. They establish lower limits on the Mean-Squared-Error (MSE) in estimating $\mathbf{a}^T\boldsymbol{\theta}$, given by $\mathbb{E}_{\mathbf{y},\boldsymbol{\theta}}\left(\mathbf{a}^T\widehat{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{a}^T\boldsymbol{\theta}\right)^2$. Among the Bayesian bounds, we are primarily concerned with the Ziv-Zakai bound (ZZB)

(we also briefly describe a version of the CRB, called the Bayesian CRB). Neither of these bounds (ZZB/BCRB) require the estimator to be unbiased.

The Ziv-Zakai Bound is known to be an accurate predictor of best possible estimation performance over a wide range of SNRs. Roughly speaking, it takes into account two sources of error: coarse error, when the estimate is not close to the true value of the parameter (essentially, making an error in hypothesis testing after binning the parameter space); and fine-grained error (the mean squared error from the true value when the estimate is in the right bin). At high SNR, the probability of the estimate falling into the wrong bin becomes negligible, and the Cramér Rao bound (CRB), which characterizes only fine-grained error, provides an excellent prediction of performance, while being easier to compute than the ZZB. We now state these bounds.

2.3.1 Cramér Rao Bound[59, 58]

Let $\mathbf{a} \in \mathbb{R}^K$. The variance of any unbiased estimator of $\mathbf{a}^T\boldsymbol{\theta}$, given by $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}\left(\mathbf{a}^T\widehat{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{a}^T\boldsymbol{\theta}\right)^2$, is lower bounded by $\mathbf{a}^T F^{-1}(\boldsymbol{\theta})\mathbf{a}$, where $F(\boldsymbol{\theta})$ is the Fisher Information Matrix (FIM). The (m,n)th element of the FIM is given by:

$$F_{m,n}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}\left\{\frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_m}\frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_n}\right\}. \tag{2.13}$$

For parameter estimation in AWGN (2.12), this simplifies to [59]

$$F_{m,n}(\boldsymbol{\theta}) = \frac{2}{\sigma^2}\Re\left\{\left(\frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \theta_m}\right)^H \frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \theta_n}\right\}, \tag{2.14}$$

where $\Re\{b\}$ denotes the real part of the complex number b.

Remark: Let $\mathbf{a} \in \mathbb{R}^K$ be any vector. Consider the quadratic form of the FIM $\mathbf{a}^T F(\boldsymbol{\theta})\mathbf{a}$. It can be shown that:

$$\mathbf{a}^T F(\boldsymbol{\theta})\mathbf{a} = \frac{2}{\sigma^2}\left\|\sum_{l=1}^{l=K} a_l \frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \theta_l}\right\|^2.$$

Therefore, all quadratic forms of the FIM depend on the signal manifold $\mathbf{s}(\boldsymbol{\theta})$ only through derivatives of the form $\sum a_l \left(\partial \mathbf{s}(\boldsymbol{\theta})/\partial \theta_l\right)$. These may be thought of as pairwise differences $\mathbf{s}(\boldsymbol{\theta}) - \mathbf{s}(\boldsymbol{\theta}')$ in the limit of $\boldsymbol{\theta}' \to \boldsymbol{\theta}$. We make use of this observation in Section 2.4.

2.3.2 Bayesian Bounds on Mean Square Error

To describe the Bayesian bounds, it is convenient to define the MSE matrix, $R(\widehat{\boldsymbol{\theta}})$ of the estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$. The m,n-th element of the MSE matrix $R(\widehat{\boldsymbol{\theta}})$ is given by $R_{m,n}(\widehat{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{y},\boldsymbol{\theta}}\{(\hat{\theta}_m - \theta_m)(\hat{\theta}_n - \theta_n)\}$. For a vector $\mathbf{a} \in \mathbb{R}^K$, the ZZB and BCRB provide bounds on $\mathbb{E}_{\mathbf{y},\boldsymbol{\theta}}(\mathbf{a}^T\widehat{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{a}^T\boldsymbol{\theta})^2$ which is simply $\mathbf{a}^T R(\widehat{\boldsymbol{\theta}})\mathbf{a}$.

Bayesian Cramér Rao Bound[59, 58]

For any weight vector $\mathbf{a} \in \mathbb{R}^K$ and estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ (not necessarily unbiased), the Bayesian Cramér Rao Bound (BCRB) lower bounds the MSE $\mathbf{a}^T R(\widehat{\boldsymbol{\theta}})\mathbf{a}$ by $\mathbf{a}^T B^{-1}\mathbf{a}$, where B is the Bayesian Information Matrix (BIM). The (m,n)th element of B is given by:

$$B_{m,n} = \mathbb{E}_{\boldsymbol{\theta}}\left\{F_{m,n}(\boldsymbol{\theta})\right\} + \mathbb{E}_{\boldsymbol{\theta}}\left\{\frac{\partial \ln p(\boldsymbol{\theta})}{\partial \theta_m}\frac{\partial \ln p(\boldsymbol{\theta})}{\partial \theta_n}\right\}. \tag{2.15}$$

(Extended) Ziv-Zakai Bound[9]

Since the ZZB is not as widely used as the CRB, we provide a brief review in Appendix A.1. Here, we simply state the bound. The ZZB bounds the MSE $\mathbf{a}^T R(\widehat{\boldsymbol{\theta}})\mathbf{a}$ and, for the AWGN measurement model (2.11), it is given by:

$$\mathbf{a}^T R(\widehat{\boldsymbol{\theta}})\mathbf{a} \geq \frac{1}{2}\int_0^\infty \mathcal{V}\left\{\max_{\boldsymbol{\delta}:\mathbf{a}^T\boldsymbol{\delta}=h}\int_{\boldsymbol{\phi}\in\mathbb{R}^K}\left(p(\boldsymbol{\phi})+p(\boldsymbol{\phi}+\boldsymbol{\delta})\right)f(\boldsymbol{\phi},\boldsymbol{\phi}+\boldsymbol{\delta})\,d\boldsymbol{\phi}\right\}h\,dh, \quad \forall\widehat{\boldsymbol{\theta}}(\mathbf{y}) \tag{2.16}$$

where $\mathcal{V}\{\ \}$ is the valley filling operation, defined as $\mathcal{V}\{g(h)\} = \max_{r\geq 0} g(h+r)$, and $f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ is the probability of error for the optimal detection rule in the following *hypothesis testing problem*:

$$\begin{aligned} H_1 &: \mathbf{y} = \mathbf{s}(\boldsymbol{\theta}_1) + \mathbf{z},\ \Pr(H_1) = \frac{p(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_1)+p(\boldsymbol{\theta}_2)} \\ H_2 &: \mathbf{y} = \mathbf{s}(\boldsymbol{\theta}_2) + \mathbf{z},\ \Pr(H_2) = \frac{p(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_1)+p(\boldsymbol{\theta}_2)}. \end{aligned} \tag{2.17}$$

Since $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbb{I}_M)$, this detection error probability is given by[59]:

$$f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{p(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_1) + p(\boldsymbol{\theta}_2)} Q\left(\frac{d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2} d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \ln \frac{p(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_2)} \right) + \frac{p(\boldsymbol{\theta}_2)}{p(\boldsymbol{\theta}_1) + p(\boldsymbol{\theta}_2)} Q\left(\frac{d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\sqrt{2}\sigma} - \frac{\sigma}{\sqrt{2} d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \ln \frac{p(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_2)} \right). \quad (2.18)$$

In the above expression, $Q(\ )$ stands for the CCDF of the standard normal distribution $\mathcal{N}(0, 1)$ and

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \|\mathbf{s}(\boldsymbol{\theta}_1) - \mathbf{s}(\boldsymbol{\theta}_2)\|. \quad (2.19)$$

Remark: While the expression for the ZZB is complicated, we only need two simple observations to prove the result we are interested in:

- With compressive measurements, the signal manifold $\mathbf{s}(\boldsymbol{\theta}) = \boldsymbol{A}\mathbf{x}(\theta)$ and the measurement matrix $\boldsymbol{A}$ enters the ZZB *only* through the pairwise SNRs $d^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/\sigma^2$.

- The minimum probability of detection error $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for the binary hypothesis testing problem (2.17) is a non-increasing function of the pairwise SNR $d^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/\sigma^2$.

We revisit these observations in Section 2.4.

2.3.3 Threshold behavior of ZZB

The ZZB typically exhibits a *threshold behavior* with SNR [58]. When the SNR is very low, the measurements carry little information about the parameters

we wish to estimate. Since the ZZB accounts for errors of "all magnitudes", it is usually large (depending primarily on the prior $p(\boldsymbol{\theta})$) and insensitive to small changes in SNR in this regime. However, at high SNRs, the variation of the ZZB with SNR is predictable. When the SNR and the ZZB are both expressed on a logarithmic scale, the ZZB falls off linearly with SNR, provided that the SNR is above a certain value, which is called the (asymptotic) *ZZB threshold* [9]. When the SNR exceeds the ZZB threshold, "large" estimation errors are unlikely, which is exactly when we would declare estimation of a continuous-valued parameter to be successful.

2.4 Relating the isometries to estimation bounds

We are now ready to relate the estimation error bounds for the compressive estimation problem to the corresponding bounds when we make all N measurements, provided that the compressive measurement matrix $\boldsymbol{A}$ satisfies appropriate isometry conditions.

Consider the general problem of estimating $\boldsymbol{\theta}$ from L measurements

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta}) + \mathbf{z}, \quad \boldsymbol{\theta} \in \Theta \tag{2.20}$$

where $\boldsymbol{\Phi}$ is any $L \times N$ complex-valued matrix and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbb{I}_L)$. The compressive estimation problem is subsumed in this model (obtained by setting $\boldsymbol{\Phi} = \boldsymbol{A}$,

whose elements are chosen i.i.d. from a zero-mean distribution of variance $1/N$ for which concentration results of the form (2.8) are available), as is the conventional problem of estimating $\boldsymbol{\theta}$ from all N measurements (obtained by setting $\boldsymbol{\Phi} = \mathbb{I}_N$, the $N \times N$ identity matrix). Note that, in both these cases, the per-measurement SNR $(1/L)\sum)_{k=1}^{k=L}\mathbb{E}|y_k|^2/\sigma^2$ is the same, since the rows of $\boldsymbol{A}$ have unit norm in expectation.

We prove two theorems that connect the fundamental estimation-theoretic bounds to the isometries defined in the previous section. First, we make a connection between the ZZB and the pairwise isometry property. As we observed in the remark under the statement of the ZZB, for the manifold $\mathbf{s}(\boldsymbol{\theta}) = \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta})$, the ZZB depends on the matrix $\boldsymbol{\Phi}$ only through the set of pairwise SNRs $\|\boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta}_1) - \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta}_2)\|^2/\sigma^2 \; \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. When the compressive measurement matrix $\boldsymbol{A}$ satisfies the pairwise isometry property (2.7), the pairwise SNRs with $\boldsymbol{\Phi} = \boldsymbol{A}$ are approximately M/N times the corresponding values with $\boldsymbol{\Phi} = \mathbb{I}_N$. Thus, the ZZB with compressive measurements is approximately the same as the ZZB with all N measurements, but at an SNR penalty of M/N. Theorem 2.3 proves this intuition rigorously.

Likewise, we can connect the CRB to the tangent-plane isometry property. We can show that the CRB depends on the measurement matrix only through norms of the vectors $\boldsymbol{\Phi}\sum_m a_m(\partial\mathbf{x}(\boldsymbol{\theta})/\partial\theta_m)$. Thus, if $\boldsymbol{A}$ satisfies the tangent-plane isometry

(2.10), the CRB with M compressive measurements is approximately equal to the CRB with all N measurements, but at an SNR that is lower by M/N. We prove this in Theorem 2.1.

While the connections established here between estimation-theoretic bounds and the corresponding isometries apply generally to compressive estimation in AWGN, showing that these isometries indeed hold requires a problem-specific analysis, as we illustate for sinusoidal mixtures in later sections. As with standard compressed sensing, the goal of such analyses is to characterize the number of measurements required for such isometries to hold with high probability for random measurement matrices.

2.4.1 Cramér Rao Bound

Let $F(\mathbf{\Phi}, \boldsymbol{\theta})$ denote the Fisher Information Matrix for the measurement model (2.20). For this measurement model the expression for FIM is given by (2.14) with $\mathbf{s}(\boldsymbol{\theta}) = \mathbf{\Phi}\mathbf{x}(\boldsymbol{\theta})$:

$$F_{m,n}(\mathbf{\Phi}, \boldsymbol{\theta}) = \frac{2}{\sigma^2} \Re \left\{ \left(\mathbf{\Phi} \frac{\partial \mathbf{x}(\boldsymbol{\theta})}{\partial \theta_m} \right)^H \mathbf{\Phi} \frac{\partial \mathbf{x}(\boldsymbol{\theta})}{\partial \theta_n} \right\}. \tag{2.21}$$

Theorem 2.1. *Let $\boldsymbol{A}$ be an $M \times N$ measurement matrix which satisfies the tangent plane ϵ-isometry property (2.10) for the signal manifold $\mathbf{x}(\boldsymbol{\theta})$. Then, the Fisher Information Matrix $F(\boldsymbol{A}, \boldsymbol{\theta})$, with compressive measurements (2.3) is related to*

the FIM with all N measurements as follows:

$$\begin{aligned} F(\boldsymbol{A},\boldsymbol{\theta}) &\preceq F\left(\sqrt{\tfrac{M}{N}}(1+\epsilon)\mathbb{I}_N,\boldsymbol{\theta}\right) \\ F(\boldsymbol{A},\boldsymbol{\theta}) &\succeq F\left(\sqrt{\tfrac{M}{N}}(1-\epsilon)\mathbb{I}_N,\boldsymbol{\theta}\right) \end{aligned} \quad \forall \boldsymbol{\theta}\in\Theta. \tag{2.22}$$

Proof. Consider the quadratic form $\mathbf{a}^T F(\boldsymbol{\Phi},\boldsymbol{\theta})\mathbf{a}$ for any $\mathbf{a}=[a_1\ \cdots\ a_K]^T\in\mathbb{R}^K$. We see that

$$\mathbf{a}^T F(\boldsymbol{\Phi},\boldsymbol{\theta})\mathbf{a} = \frac{2}{\sigma^2}\left\|\boldsymbol{\Phi}\sum_m a_m\frac{\partial \mathbf{x}(\boldsymbol{\theta})}{\partial\theta_m}\right\|^2. \tag{2.23}$$

Since the compressive measurement matrix $\boldsymbol{A}$ satisfies the tangent plane ϵ-isometry property (2.10) for the signal model $\mathbf{x}(\boldsymbol{\theta})$, we have that for all $\boldsymbol{\theta}\in\Theta$ and $\mathbf{a}\in\mathbb{R}^K$,

$$\left\|\boldsymbol{A}\sum_m a_m\frac{\partial \mathbf{x}(\boldsymbol{\theta})}{\partial\theta_m}\right\|^2 \le \frac{M}{N}(1+\epsilon)^2\left\|\sum_m a_m\frac{\partial \mathbf{x}(\boldsymbol{\theta})}{\partial\theta_m}\right\|^2. \tag{2.24}$$

Multiplying both sides by $2/\sigma^2$, we see that the LHS is $\mathbf{a}^T F(\boldsymbol{A},\boldsymbol{\theta})\mathbf{a}$, while the RHS corresponds to $\mathbf{a}^T F\left(\sqrt{M/N}(1+\epsilon)\mathbb{I}_N,\boldsymbol{\theta}\right)\mathbf{a}$. Therefore, we have that $\forall\boldsymbol{\theta}\in\Theta$,

$$\mathbf{a}^T F(\boldsymbol{A},\boldsymbol{\theta})\mathbf{a} \le \mathbf{a}^T F\left(\sqrt{M/N}(1+\epsilon)\mathbb{I}_N,\boldsymbol{\theta}\right)\mathbf{a},\ \forall\mathbf{a}\in\mathbb{R}^K. \tag{2.25}$$

This establishes the required upper bound on $F(\boldsymbol{A},\boldsymbol{\theta})$. The proof for the lower bound is analogous. □

2.4.2 Bayesian Cramér Rao Bound

Let $B(\boldsymbol{\Phi})$ denote the Bayesian Information Matrix for the measurement model (2.20). Let $p(\boldsymbol{\theta})$ be the prior on $\boldsymbol{\theta}$. For this measurement model the expression

for BIM is given by (2.15) with $\mathbf{s}(\boldsymbol{\theta}) = \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta})$:

$$B_{m,n}(\boldsymbol{\Phi}) = \mathbb{E}_{\boldsymbol{\theta}}\left\{F_{m,n}(\boldsymbol{\Phi},\boldsymbol{\theta})\right\} + \mathbb{E}_{\boldsymbol{\theta}}\left\{\frac{\partial \ln p(\boldsymbol{\theta})}{\partial \theta_m}\frac{\partial \ln p(\boldsymbol{\theta})}{\partial \theta_n}\right\}. \tag{2.26}$$

Corollary 2.2 (of Theorem 2.1)**.** *Let $\boldsymbol{A}$ be an $M \times N$ measurement matrix which satisfies the tangent plane ϵ-isometry property (2.10) for the signal manifold $\mathbf{x}(\boldsymbol{\theta})$. Then, the Bayesian Information Matrix $B(\boldsymbol{A})$ with compressive measurements (2.3) is related to the BIM with all N measurements as follows:*

$$B\left(\sqrt{\frac{M}{N}}(1-\epsilon)\mathbb{I}_N\right) \preceq B(\boldsymbol{A}) \preceq B\left(\sqrt{\frac{M}{N}}(1+\epsilon)\mathbb{I}_N\right) \tag{2.27}$$

Proof. Let $\mathbf{a} \in \mathbb{R}^K$. We see that $\mathbf{a}^T B(\boldsymbol{A})\mathbf{a}$ depends on the measurement matrix $\boldsymbol{A}$ only through quadratic forms of the FIM i.e., $\mathbf{a}^T F(\boldsymbol{A},\boldsymbol{\theta})\mathbf{a}$. When the tangent plane isometry condition (2.10) is satisfied, we have from Theorem 2.1 that $\mathbf{a}^T F(\boldsymbol{A},\boldsymbol{\theta})\mathbf{a}$ is bounded by $\mathbf{a}^T F(\sqrt{M/N}(1\pm\epsilon)\mathbb{I}_N,\boldsymbol{\theta})\mathbf{a}$ for all $\mathbf{a},\boldsymbol{\theta}$. It immediately follows that the quadratic forms of $B(\boldsymbol{A})$ are bounded by the corresponding quadratic forms of $B(\sqrt{M/N}(1\pm\epsilon)\mathbb{I}_N)$. □

2.4.3 Ziv-Zakai Bound

Let $Z(\boldsymbol{\Phi},\mathbf{a})$ denote the ZZB corresponding to the Mean-Squared-Error in estimating $\mathbf{a}^T\boldsymbol{\theta}$ for the measurement model (2.20). The expression for $Z(\boldsymbol{\Phi},\mathbf{a})$ is given by the right hand side of (2.16), with $d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \|\boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta}_1) - \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta}_2)\|$ (obtained by setting $\mathbf{s}(\boldsymbol{\theta}) = \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta})$).

Note that in (2.16), $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is the probability of detection error for the hypothesis testing problem (2.17) with $\mathbf{s}(\boldsymbol{\theta}) = \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta})$. We capture the dependence of this probability on the matrix $\boldsymbol{\Phi}$ by defining $g(\boldsymbol{\Phi}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ when $\mathbf{s}(\boldsymbol{\theta}) = \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta})$.

Theorem 2.3. *Let $\boldsymbol{A}$ be an $M \times N$ measurement matrix which satisfies the pairwise ϵ-isometry property (2.7) for the signal manifold $\mathbf{x}(\boldsymbol{\theta})$. Then, the ZZB $Z(\boldsymbol{A}, \mathbf{a})$, with the compressive measurements in (2.3), is related to the ZZB with all N measurements as*

$$Z\left(\sqrt{\frac{M}{N}}(1+\epsilon)\mathbb{I}_N, \mathbf{a}\right) \leq Z\left(\boldsymbol{A}, \mathbf{a}\right) \leq Z\left(\sqrt{\frac{M}{N}}(1-\epsilon)\mathbb{I}_N, \mathbf{a}\right). \tag{2.28}$$

Proof. As we observed in the remark at the end of the definition of the ZZB, $g(\boldsymbol{\Phi}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is a non-increasing function of the pairwise SNR $\|\boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta}_1) - \boldsymbol{\Phi}\mathbf{x}(\boldsymbol{\theta}_2)\|^2/\sigma^2$. When $\boldsymbol{A}$ satisfies the pairwise ϵ-isometry property (2.7), we can bound all the pairwise SNRs as follows:

$$\|\boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}_1) - \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}_2)\|^2/\sigma^2 \leq \frac{M}{N}(1+\epsilon)^2\|\mathbf{x}(\boldsymbol{\theta}_1) - \mathbf{x}(\boldsymbol{\theta}_2)\|^2/\sigma^2 \ \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta. \tag{2.29}$$

Combining these facts, we get $g(\boldsymbol{A}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \geq g(\sqrt{M/N}(1+\epsilon)\mathbb{I}_N, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, which is the probability of detection error with all N measurements, but at an SNR penalty of $(M/N)(1+\epsilon)^2$. Substituting these pointwise bounds in the expression for $Z(\boldsymbol{A}, \mathbf{a})$, we have that $Z\left(\boldsymbol{A}, \mathbf{a}\right) \geq Z\left(\sqrt{\frac{M}{N}}(1+\epsilon)\mathbb{I}_N, \mathbf{a}\right)$. The other inequality can be proved similarly. □

2.4.4 Number of measurements needed

These theorems show that, when the compressive measurement matrix $\boldsymbol{A}$ satisfies the pairwise isometry property, the CRB and the ZZB are well approximated by $\mathbf{a}^T F^{-1}(\sqrt{M/N}\mathbb{I}_N, \boldsymbol{\theta})\mathbf{a}$ and $Z(\sqrt{M/N}\mathbb{I}_N, \mathbf{a})$ respectively (for any $\mathbf{a}$). Thus, the estimation performance with the measurement matrix $\boldsymbol{\Phi} = \boldsymbol{A}$ is roughly the same as that with $\boldsymbol{\Phi} = \sqrt{M/N}\mathbb{I}_N$ (all N measurements, but with the signal component scaled by $\sqrt{M/N}$). Note that observations with $\boldsymbol{\Phi} = \sqrt{M/N}\mathbb{I}_N$ and per-sample noise variance σ^2 are *equivalent* to observations $\boldsymbol{\Phi} = \mathbb{I}_N$ (conventional measurements) but with an increased per-sample noise variance $\sigma^2(N/M)$ (easily seen by multiplying the observations with $\boldsymbol{\Phi} = \sqrt{M/N}\mathbb{I}_N$ by $\sqrt{N/M}$). Putting these observations together, we get a simple procedure for estimating the number of measurements M required for successful compressive estimation:

(1) For the case when we make all N measurements, $\mathbf{y} = \mathbf{x}(\boldsymbol{\theta}) + \mathbf{z}$ with $\mathbf{z} \sim \mathcal{CN}(0, \sigma^2\mathbb{I}_N)$, compute the ZZB as a function of σ^2. Find the ZZB threshold as described in Section 2.3 (the value of σ^2 *below* which $\log$ ZZB falls off linearly with $\log \sigma^2$). Denote this threshold by σ_{t}^2.

(2) Making M compressive measurements $\mathbf{y} = \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}) + \mathbf{z}$ with $\mathbf{z} \sim \mathcal{CN}(0, \sigma_0^2\mathbb{I}_M)$ is roughly equivalent to making the observations

$$\tilde{\mathbf{y}} = \mathbf{x}(\theta) + \tilde{\mathbf{z}}, \quad \tilde{\mathbf{z}} \sim \mathcal{CN}(0, \sigma_0^2(N/M)\mathbb{I}_M)$$

when $\boldsymbol{A}$ satisfies the pairwise isometry property. Thus, the number of measurements needed for successful compressive estimation is given by:

$$\sigma_0^2 \frac{N}{M} < \sigma_{\mathrm{t}}^2 \quad \text{or} \quad M > N \left(\frac{\sigma_0^2}{\sigma_{\mathrm{t}}^2} \right) \tag{2.30}$$

We reiterate that the above SNR criterion is not the only condition for successful compressive estimation: the number of measurements M must be large enough for the matrix $\boldsymbol{A}$ to satisfy the pairwise isometry property, so that we can invoke the SNR penalty arguments.

2.4.5 Remarks on model generality

While we describe our results in the context of the measurement model (2.3), they extend easily to variants commonly encountered in the compressed sensing literature, two of which we now discuss.

• For applications such as Direction of Arrival (DoA) estimation using large arrays [46], compressive measurements are acquired sequentially in time and every measurement is corrupted by independent *measurement* noise. Thus, the measurements satisfy

$$y_l = \mathbf{w}_l^T \left(\mathbf{x}(\boldsymbol{\theta}) + \tilde{\mathbf{z}}_l \right) = \mathbf{w}_l^T \mathbf{x}(\boldsymbol{\theta}) + z_l, \tag{2.31}$$

where $\tilde{\mathbf{z}}_l \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2 \mathbb{I}_N\right)$ and $z_l = \mathbf{w}_l^T \tilde{\mathbf{z}}_l \sim \mathcal{CN}(0, \sigma^2 \|\mathbf{w}_l\|^2)$. The key point here is that $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_M$ are i.i.d. and as a result $z_1, \ldots, z_M$ are independent. Letting

$\boldsymbol{A}$ denote the matrix with rows $\mathbf{w}_l^T$, $\mathbf{y} = [y_1 \ \cdots \ y_M]^T$ and $\mathbf{z} = [z_1 \ \cdots \ z_M]^T$, we have:

$$\mathbf{y} = \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}) + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{K}_1), \tag{2.32}$$

where $\mathbf{K}_1$ is a diagonal matrix whose diagonal entries are $\|\mathbf{w}_l\|^2, \ l = 1, \ldots, M$.

• For other applications, when we have access to a *single* noisy version of $\mathbf{x}(\boldsymbol{\theta})$ and compressive measurements are merely used as a *dimensionality reduction* tool, we have

$$\mathbf{y} = \boldsymbol{A}\left(\mathbf{x}(\boldsymbol{\theta}) + \tilde{\mathbf{z}}\right), \quad \tilde{\mathbf{z}} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbb{I}_N). \tag{2.33}$$

The same equation holds for the case when there are errors in modeling the manifold $\mathbf{x}(\boldsymbol{\theta})$ (given by $\tilde{\mathbf{z}}$) and we make M sequential *noiseless* projections. Letting $\mathbf{z} = \boldsymbol{A}\tilde{\mathbf{z}}$ we have

$$\mathbf{y} = \boldsymbol{A}\mathbf{x}(\boldsymbol{\theta}) + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{K}_2), \tag{2.34}$$

where $\mathbf{K}_2 = \boldsymbol{A}\boldsymbol{A}^H$.

Neither $\mathbf{K}_1$ and $\mathbf{K}_2$ are the identity matrix, hence these measurement models do not fit directly into the framework in (2.3). However, we can extend our results easily to these models by considering the whitened observations $\tilde{\mathbf{y}}_i = \mathbf{K}_i^{-1/2}\mathbf{y}, \ i = 1, 2$, and establishing bounds on the singular values of $\mathbf{K}_i$. When the elements of

$\boldsymbol{A}$ are chosen from a zero-mean distribution of variance $1/N$ (for which concentration results of the form (2.8) are available), the singular values of $\mathbf{K}_i$ concentrate around 1. As a result, an ϵ-isometry (tangent plane or pairwise) for $\boldsymbol{A}$ can be shown to translate to a mildly weaker $\epsilon_{\text{eff},i}$-isometry ($\epsilon_{\text{eff},i} \geq \epsilon$) for $\boldsymbol{A}_{\text{eff},i} = \mathbf{K}_i^{-1/2}\boldsymbol{A}$, the effective measurement matrix for the whitened measurements $\tilde{\mathbf{y}}_i$. All of our results now apply by simply replacing ϵ with $\epsilon_{\text{eff},i}$. This equivalence of the measurement model (2.33) and the general compressive model (2.3) has also been investigated in detail in [4]. The proof for the conditioning of both $\mathbf{K}_1$ and $\mathbf{K}_2$ involves using the concentration result (2.8) for $\sqrt{N/M}\boldsymbol{A}^H$ (see [60] for $\mathbf{K}_2$).

The concentration results for the singular values of $\mathbf{K}_2 = \boldsymbol{A}\boldsymbol{A}^H$ (which are the square of the singular values of $\boldsymbol{A}^H$) needs N to be somewhat larger than M. This is not an issue, since this is the regime of interest for compressive estimation. The diagonal matrix $\mathbf{K}_1$, on the other hand, is well-conditioned for much larger values of M (potentially larger than N).

Chapter 3

Compressive Frequency Estimation

In this chapter, we apply the results derived in Chapter 2 to the problem of frequency estimation from random projections of a mixture of sinusoids. In addition to channel estimation for 60GHz arrays which we briefly discussed in the introduction, this problem also appears in many other scenarios such as channel estimation for OFDM systems, range estimation in radar, etc., In Chapter 2, we showed that for parameter estimation from compressive measurements, we need to preserve the geometry of the estimation problem (provide pairwise and tangent-plane isometries). This chapter characterizes the number of random projections needed to give such geometry preservation results for frequency estimation. We

also verify the numerical prescription given in Chapter 2 for the number of random projections needed to avoid gross errors.

Outline: We begin in Section 3.1 by surveying prior work on compressive frequency estimation. We state the measurement model used in Section 3.2. Section 3.3 derives the number of compressive measurements needed to guarantee isometry conditions for the problem of frequency estimation from a mixture of K sinusoids of length N. We assume that we take M projections with the projection weights chosen uniformly and independently at random from the set $\{\pm 1, \pm j\}^N$. We show that:

(a) $M = O\left(K \log(NK\delta^{-1})\right)$ measurements suffice to provide tangent plane isometries, where δ depends on the frequency separation between the sinusoids in the mixture (δ vanishes when any two of the K frequencies approach one another).

(b) $M = O\left(K \log(NK\delta^{-1})\right)$ measurements suffice to provide pairwise isometries between two sets of frequencies $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_K)$ and $\boldsymbol{\omega}' = (\omega'_1, \omega'_2, \ldots, \omega'_K)$ that are "well-separated." Here δ depends only on the frequency separation between the sinusoids in the mixture of $2K$ sinusoids $(\boldsymbol{\omega}, \boldsymbol{\omega}')$, and vanishes when any two frequencies in $(\boldsymbol{\omega}, \boldsymbol{\omega}')$ approach one another. Therefore, with

$O\left(K \log(NK\delta^{-1})\right)$ compressive measurements, we can preserve the "well-separated" geometry of the frequency estimation problem.

The tangent plane isometry results (a) indicate that when the K frequencies in $\boldsymbol{\omega}$ themselves are "well-separated", compressive measurements preserve the "fine" geometry of the frequency estimation problem (and therefore the CRB). We strengthen these results in Section 3.4 for a single sinusoid ($K = 1$), exploiting the continuity of the sinusoidal manifold to show that $O(\log N)$ measurements suffice to guarantee pairwise isometry between sinusoids at *any* two frequencies ω, ω' (by merging the "well-separated" and "fine" regimes). In Section 3.6, we consider the problem of estimating the frequency of a sinusoid (channel composed of a single path: $K = 1$) and show that the criterion for prediction of the number of measurements, based on the threshold behavior of the ZZB given in Chapter 2, is tight, by evaluating the performance of an algorithm which closely approximates the MAP estimator. The algorithm works in two stages: first, from a discrete set of frequencies, we pick the one that fits the observations best and, then, we perform local refinements using Newton's method.

3.1 Related work

Reference [5] used the Johnson-Lindenstrauss (JL) lemma to provide a simple proof that $O(K \log N)$ random projections suffice to establish RIP for recovering K-sparse vectors in $\mathbb{R}^N$. We briefly summarize the key ideas, since we use an analogous approach in establishing pairwise isometry for the mixture of sinusoids in Section 3.3. The JL lemma states that, to approximately preserve the pairwise distances between P points after random projections (with the weights chosen from appropriate distributions, such as i.i.d. Uniform$\{\pm 1\}$ [3]), we need $O(\log P)$ such projections. However, to provide an RIP for compressive measurement matrices, the distances between any two K-sparse vectors must be preserved. Since the number of such vectors is infinite, the JL lemma cannot be applied directly. However, the desired RIP result is established in [5] by discretizing the set of K-sparse vectors sufficiently finely, applying the JL lemma to the resulting discrete set of points, and then exploiting continuity to provide isometries for the remaining points.

Algorithms to estimate the frequencies in a mixture of sinusoids from compressive measurements are proposed and evaluated in [19, 21]. Both of these papers assume that the sinusoids have a minimum frequency separation and [21] further assumes that the frequencies come from an oversampled DFT grid. They

propose variants of standard compressed sensing algorithms, such as Orthogonal Matching Pursuit (OMP) and Iterative Hard Thresholding (IHT), which rely on the sinusoids' frequencies not being too close. As mentioned earlier, restricting the frequency estimation to a discrete grid in this fashion results in performance floors, as studied in great detail in [15]. However, as we show in this dissertation, it is possible to avoid such performance floors, and to attain the CRB, by local refinements based on Newton-like algorithms after grid-based coarse estimation. A one-shot quadratic refinement is also proposed in [19] to improve estimates of off-grid frequencies.

We characterize the structure of compressive estimation here in terms of that of the original problem. However, in many cases, an estimation-theoretic understanding of the original problem is incomplete: in particular, for the mixture of sinusoids model, a characterization of the difficulty of the problem in terms of the minimum separation of frequencies in $\boldsymbol{\omega}$ remains an ongoing effort [56, 12, 57], as discussed in more detail below.

The problem of estimating frequencies in a mixture of sinusoids from noise-free compressive measurements is studied in [57]. While the frequencies can come from the $[0, 2\pi)$ continuum, [57] requires that they are "well-separated" (four times the DFT spacing of $2\pi/N$). When this condition is met, it is shown that atomic-norm denoising (cast as a semi-definite program) correctly estimates the frequencies in

the mixture. The same $4 \times (2\pi/N)$ frequency separation is shown to be necessary for recovering frequencies over a continuum with noisy measurements of all N samples (not compressive) in [56, 12]. It is interesting to note that even when all N samples are observed, the same minimum frequency separation is necessary for stable recovery. This falls in line with the observations that we make on the equivalence (except for an SNR penalty) of the "difficulty" in estimation using compressive measurements and uncompressed measurements (all N samples) by relating corresponding estimation error bounds.

In the algorithm description and numerical illustrations in this dissertation, we restrict attention to a single sinusoid in order to illustrate the fundamental features of compressive estimation. Our algorithmic approach (discrete grid followed by Newton refinement) extends easily to estimate the frequencies of multiple sinusoids[45, 46].

3.2 System model

Consider a manifold of signals which are linear combinations of K complex sinusoids $\sum_{l=1}^{K} g_l \mathbf{x}(\omega_l)$, where $g_l \in \mathbb{C}$ are complex gains and

$$\mathbf{x}(\omega) = \left[h_1 e^{-j\omega(N-1)/2} \ \cdots \ h_N e^{j\omega(N-1)/2}\right]^T \tag{3.1}$$

is a windowed sinusoid, with window weights given by $\{h_n\}$. Without loss of generality, we assume that the window weights are normalized so that $\sum_n |h_n|^2 = 1$. To avoid trivialities, we assume that more than one of the h_n's are non-zero.

We assume that we make M compressive measurements of the form

$$\mathbf{y} = \boldsymbol{A} \sum_{l=1}^{l=K} g_l \mathbf{x}(\omega_l) + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2 \mathbb{I}_N\right). \tag{3.2}$$

where the elements of the $M \times N$ matrix $\boldsymbol{A}$ are drawn uniformly and independently at random from $\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$. We refer the reader to Section 2.4.5 of Chapter 2 and Section 3.5 of this chapter for comments on the generality of this model. We wish to estimate the gains $\mathbf{g} = [g_1 \ \cdots \ g_K]^T$ and the frequencies $\boldsymbol{\omega} = [\omega_1 \ \cdots \ \omega_K]^T$. Therefore, the parameter to be estimated is $\boldsymbol{\theta} = (\mathbf{g}, \boldsymbol{\omega})$. In the forthcoming section, we characterize the number of measurements needed to give the necessary tangent plane and pairwise isometries.

3.3 Isometry conditions for frequency estimation from compressive measurements

We show that, for a mixture of K sinusoids, the number of measurements required depends on the conditioning of appropriately defined matrices, which in turn depends on the separation between the frequencies in the mixture. We

consider the special case of a single sinusoid, for which we can prove stronger results in Section 3.4.

3.3.1 Tangent plane isometry for a mixture of K sinusoids

Our first goal is to quantify the number of measurements needed to preserve the CRB for a given frequency support $\boldsymbol{\omega}$ (i.e., for all $\boldsymbol{\theta}$ that share this frequency support). We show that this is equivalent to guaranteeing ϵ-isometry for a *set* of tangent planes as follows. For any specific value of the unknown parameters – gain magnitude $\{|g_l|\}$, phases $\{g_l/|g_l|\}$ and frequencies $\{\omega_l\}$ (we split the complex gain in this manner in order to restrict attention to real parameters) – Theorem 2.1 guarantees that the CRB can be preserved (up to the M/N SNR penalty) by ensuring ϵ-isometry for the plane tangent to the manifold at this set of parameters. Therefore, to preserve the CRB for the frequency support $\boldsymbol{\omega}$, we need to guarantee ϵ-isometry for tangent-planes for all values that the gain magnitudes $\{|g_l|\}$ and the phases $\{g_l/|g_l|\}$ can take. We can show that the union of all such tangent planes is a subset of the span of the matrix $\mathbf{T}(\boldsymbol{\omega})$ (in $\mathbb{C}^N$), defined as

$$\mathbf{T}(\boldsymbol{\omega}) = \left[\mathbf{x}(\omega_1) \ \cdots \ \mathbf{x}(\omega_K) \ \tau\frac{d\mathbf{x}(\omega_1)}{d\omega} \ \cdots \ \tau\frac{d\mathbf{x}(\omega_K)}{d\omega}\right] \tag{3.3}$$

where $\tau = 1/\|d\mathbf{x}(\omega)/d\omega\|$ (note that τ does not depend on ω). Therefore, if the compressive measurement matrix $\boldsymbol{A}$ satisfies

$$\sqrt{\frac{M}{N}}(1-\epsilon) \leq \frac{\|\boldsymbol{A}\mathbf{T}(\boldsymbol{\omega})\mathbf{q}\|}{\|\mathbf{T}(\boldsymbol{\omega})\mathbf{q}\|} \leq \sqrt{\frac{M}{N}}(1+\epsilon) \ \forall \mathbf{q} \in \mathbb{C}^{2K}, \tag{3.4}$$

we can preserve the CRB (up to the SNR penalty) for a given frequency support $\boldsymbol{\omega}$. Furthermore, if the above relationship holds, we say that $\boldsymbol{A}$ satisfies the tangent plane ϵ-isometry property at $\boldsymbol{\omega}$.

Our first result is to show that the smallest singular value of the matrix $\mathbf{T}(\boldsymbol{\omega})$, given by $\delta = \min_{\mathbf{q}\in\mathbb{C}^{2K}} \|\mathbf{T}(\boldsymbol{\omega})\mathbf{q}\|/\|\mathbf{q}\|$, compactly characterizes the number of measurements needed to preserve tangent plane ϵ-isometry.

Theorem 3.1. *Let $\boldsymbol{A}$ be an $M \times N$ measurement matrix whose entries are drawn i.i.d. from* Uniform $\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$. *Let $\mathbf{T}(\boldsymbol{\omega})$ denote the tangent plane matrix (3.3) of sinusoids (3.1) with frequencies $\boldsymbol{\omega} = (\omega_1 \dots \omega_K) \in \mathbb{R}^K$. Let $\Lambda_T(\delta) = \{\boldsymbol{\omega} : \text{smallest singular value of } \mathbf{T}(\boldsymbol{\omega}) \geq \delta\}$. Then, for any $\epsilon > 0$, we have*

$$1 - \epsilon \leq \sqrt{\frac{N}{M}} \frac{\|\boldsymbol{A}\mathbf{T}(\boldsymbol{\omega})\mathbf{q}\|}{\|\mathbf{T}(\boldsymbol{\omega})\mathbf{q}\|} \leq 1 + \epsilon, \forall \boldsymbol{\omega} \in \Lambda_T(\delta), \mathbf{q} \in \mathbb{C}^{2K} \tag{3.5}$$

with high probability when $M = O(\epsilon^{-2} K \log(NK\ \epsilon^{-1}\delta^{-1}))$.

Remarks:

• The theorem states that the minimum number of measurements scales as the inverse of the smallest singular value δ. The singular values of $\mathbf{T}(\boldsymbol{\omega})$ are the square roots of the eigenvalues of $\mathbf{T}^H(\boldsymbol{\omega})\mathbf{T}(\boldsymbol{\omega})$, whose entries can be shown to depend only on the set of frequency *differences* $\omega_i - \omega_j, 1 \leq i, j \leq K$. Therefore, δ depends only on the set of frequency differences.

• The smallest singular value δ tends to zero when any two of the K frequencies (say ω_i and ω_j) get close, since the columns $\mathbf{x}(\omega_i)$ and $\mathbf{x}(\omega_j)$ (and hence the columns $d\mathbf{x}(\omega_i)/d\omega$ and $d\mathbf{x}(\omega_j)/d\omega$) approach each other, and the matrix $\mathbf{T}(\boldsymbol{\omega})$ becomes poorly conditioned. It is a natural question, therefore, to ask whether it is possible to provide a lower bound on δ, and hence an upper bound on the number of measurements required to give tangent plane isometries, by ensuring that the spacing between the constituent frequencies is large enough (larger than say $\Delta\omega$). We leave this as a topic for further investigation, since that characterization of the smallest singular value δ in terms of the minimum frequency separation $\Delta\omega$ is a feature of the original system with a full set of measurements rather than a problem inherent to compressive estimation. It is interesting to note that prior work on non-compressive frequency estimation [56, 12, 13], while not directly working with the parameter δ, also requires a minimum frequency separation for successful estimation (e.g., a separation of around four times the DFT spacing of $2\pi/N$) using N measurements.

• When the frequency support $\boldsymbol{\omega}$ is "*roughly*" known ahead of time (say $\boldsymbol{\omega} \approx \boldsymbol{\omega}_0$), such as in tracking scenarios encountered in radar (where frequencies correspond to directions of arrival), $\boldsymbol{A}$ need *only* preserve the norms of vectors in the span of $\mathbf{T}(\boldsymbol{\omega})$ for $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ (not all $\boldsymbol{\omega} \in \Lambda_T(\delta)$). Typically, the number of sinusoids K in the mixture is small. So, one can do better than the M/N SNR penalty that would be incurred if a compressive measurement matrix is used: In such a scenario, it may even be possible to preserve the CRB with no SNR degradation whatsoever. The equivalent problem of direction-of-arrival estimation is studied in [63]. The precise conditions on $\boldsymbol{A}$ so that the CRB is preserved with no SNR penalty are stated in [63]. This, however, requires knowing the very frequencies that we wish to estimate. Of course, this is not applicable to the one-shot estimation problem considered here, where we wish to preserve the CRB (up to the SNR penalty M/N) with a few measurements M, irrespective of what the particular realization of $\boldsymbol{\omega}$ is.

3.3.2 Pairwise isometry for a mixture of K sinusoids

Consider now the problem of quantifying the number of measurements needed to guarantee pairwise ϵ-isometry for a mixture of K sinusoids. We denote the matrix containing the sinusoids $[\mathbf{x}(\omega_1)\ \mathbf{x}(\omega_2)\ \ldots \mathbf{x}(\omega_K)]$ by $\mathbf{X}(\boldsymbol{\omega})$. From the defini-

tion of pairwise isometry in Section 2.2, compressive measurements must preserve the ML cost structure, thereby implying that

$$\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g} - \boldsymbol{A}\mathbf{X}(\boldsymbol{\omega}')\mathbf{g}'\| \approx \sqrt{M/N}\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g} - \mathbf{X}(\boldsymbol{\omega}')\mathbf{g}'\|,$$

for pairs of $(\mathbf{g}, \boldsymbol{\omega})$ and $(\mathbf{g}', \boldsymbol{\omega}')$ of interest. We are typically interested in all values of the gains $\mathbf{g}, \mathbf{g}'$ but may restrict the set of frequencies $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ to each come from a set Θ (for example, the set of K frequencies that are separated pairwise by at least $\Delta\omega$).

To simplify the problem, we only consider $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ that are "well-separated" (we comment on why this helps later). For example, we may restrict $\boldsymbol{\omega}'$ to $\Theta'(\boldsymbol{\omega}) = \Theta \backslash B(\boldsymbol{\omega}, \mu)$, where $B(\boldsymbol{\omega}, \mu)$ is a small ball of frequencies around $\boldsymbol{\omega}$. (A possible definition for the ball $B(\boldsymbol{\omega}, \mu)$ can be $B(\boldsymbol{\omega}, \mu) = \{\boldsymbol{\omega}' : \min_{1 \leq i,j \leq K} |\omega_i' - \omega_j| \leq \mu\}$). Suppose that we make enough measurements to guarantee pairwise ϵ-isometry for all $\boldsymbol{\omega} \in \Theta$ and $\boldsymbol{\omega}' \in \Theta'(\boldsymbol{\omega})$, no matter what value $\boldsymbol{\omega}$ takes. This implies that for any set of frequencies $\boldsymbol{\omega} \in \Theta$, we have preserved the cost-structure of the estimation problem at hypothesis frequencies $\boldsymbol{\omega}'$ that are "far-away" ($\boldsymbol{\omega}'$ outside $B(\boldsymbol{\omega}, \mu)$). Roughly, a good estimation algorithm should not incur frequency errors larger than μ at high SNRs.

We introduce some notation for the following discussion. Let $\tilde{\boldsymbol{\omega}} = [\boldsymbol{\omega}\ \boldsymbol{\omega}'], \tilde{\mathbf{g}} = [\mathbf{g}\ -\mathbf{g}']$ denote vectors of length $2K$ concatenating the gains and frequencies.

Also let $\mathbf{X}(\tilde{\boldsymbol{\omega}}) = [\mathbf{X}(\boldsymbol{\omega})\ \mathbf{X}(\boldsymbol{\omega}')]$ denote the $N \times 2K$ matrix containing all the sinusoids. Note that $\tilde{\mathbf{g}}$ can take any value in $\mathbb{C}^{2K}$ but $\tilde{\boldsymbol{\omega}}$ has a special structure: its first K entries $\boldsymbol{\omega}$ must belong to Θ and its last K entries come from a set $\Theta'(\boldsymbol{\omega})$ that depend on the first K values. As shorthand, we say that $\tilde{\boldsymbol{\omega}} \in \tilde{\Theta} = \{[\boldsymbol{\omega}\ \boldsymbol{\omega}'] : \boldsymbol{\omega} \in \Theta, \boldsymbol{\omega}' \in \Theta'(\boldsymbol{\omega})\}$. With this notation, the above pairwise isometry condition for a mixture of K sinusoids, which we desire can be written as

$$\sqrt{\frac{M}{N}}(1-\epsilon) \leq \frac{\|\boldsymbol{A}\mathbf{X}(\tilde{\boldsymbol{\omega}})\tilde{\mathbf{g}}\|}{\|\mathbf{X}(\tilde{\boldsymbol{\omega}})\tilde{\mathbf{g}}\|} \leq \sqrt{\frac{M}{N}}(1+\epsilon)\ \forall \tilde{\mathbf{g}} \in \mathbb{C}^{2K}, \tag{3.6}$$

for a particular $\tilde{\boldsymbol{\omega}} \in \tilde{\Theta}$. If the matrix $\boldsymbol{A}$ satisfies this relationship, we say that $\boldsymbol{A}$ guarantees ϵ-isometry (*just isometry, not pairwise*) for the frequency support $\tilde{\boldsymbol{\omega}}$ ($2K$ sinusoids).

Our goal is to quantify the number of measurements necessary for (3.6) to hold for all $\tilde{\boldsymbol{\omega}} \in \tilde{\Theta}$. While solving this problem in its entirety is difficult, we can break it down into two subproblems, the first of which we tackle. We explain the solution to this subproblem and then comment on the other. In analogy with our previous discussion of tangent plane isometry, let $\Lambda_p(\delta)$ denote the set of all frequencies $\tilde{\boldsymbol{\omega}}$ (chosen from anywhere in $\mathbb{R}^{2K}$, not just $\tilde{\Theta}$) such that the smallest singular value of $\mathbf{X}(\tilde{\boldsymbol{\omega}})$ is at least as large as δ. Suppose that we want $\boldsymbol{A}$ to guarantee ϵ-isometry for all $\tilde{\boldsymbol{\omega}} \in \Lambda_p(\delta)$ (as in (3.6) except that the set from which

$\tilde{\boldsymbol{\omega}}$ is chosen has changed). We show that $M = O\left(\epsilon^{-2}(2K)\log\left(N(2K)\epsilon^{-1}\delta^{-1}\right)\right)$ measurements suffice to provide such a guarantee with high probability.

Theorem 3.2. *Suppose that* $\boldsymbol{A}$ *is an* $M \times N$ *measurement matrix whose entries are drawn i.i.d. from* Uniform $\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$. *Let* $\mathbf{X}(\boldsymbol{\omega}) = [\mathbf{x}(\omega_1) \ \ldots \ \mathbf{x}(\omega_K)]$ *denote an* $N \times K$ *matrix of sinusoids (3.1) with* $\boldsymbol{\omega} = (\omega_1 \ldots \omega_K) \in \mathbb{R}^K$. *Let* $\Lambda_p(\delta) = \{\boldsymbol{\omega} : \textit{smallest singular value of } \mathbf{X}(\boldsymbol{\omega}) \textit{ is greater than or equal to } \delta\}$. *For any* $\epsilon > 0$ *and* $\delta > 0$, *we have*

$$1 - \epsilon \leq \sqrt{\frac{N}{M}} \frac{\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|} \leq 1 + \epsilon, \forall \boldsymbol{\omega} \in \Lambda_p(\delta), \mathbf{g} \in \mathbb{C}^K, \tag{3.7}$$

with high probability when $M = O\left(\epsilon^{-2}K\log\left(NK\epsilon^{-1}\delta^{-1}\right)\right)$.

Remarks:

• Returning to the problem posed in (3.6), suppose that the smallest singular value of $\mathbf{X}(\tilde{\boldsymbol{\omega}})$, further minimized over all values of $\tilde{\boldsymbol{\omega}} \in \tilde{\Theta}$ is $\sigma_{min} > 0$. Then, $\tilde{\Theta}$ is contained in $\Lambda_p(\sigma_{min})$ and using Theorem 3.2, $M = O\left(\epsilon^{-2}(2K)\log\left(N(2K)\epsilon^{-1}\sigma_{min}^{-1}\right)\right)$ measurements suffice to guarantee the required ϵ-isometry.

• While the singular values of $\mathbf{X}(\tilde{\boldsymbol{\omega}})$ depend only on frequency differences, we leave the question of quantifying σ_{min} (e.g., in terms of the minimum pairwise separation $\Delta\omega$ of frequencies for $\boldsymbol{\omega} \in \Theta$) and μ (the radius of the ball around each $\boldsymbol{\omega} \in \Theta$) as an open problem. The problem of lower bounding the singular values of the Fourier matrix ($\mathbf{X}(\boldsymbol{\omega})$, when choosing $\{h_n\}$ in (3.1) as the all-ones sequence)

as a function of minimum frequency separation has been investigated in [22] (using Gershgorin-type bounds). Similar ideas may be useful in our present context as well, but again, these are fundamental and difficult questions regarding the original frequency estimation problem (with a full set of measurements) that are beyond our scope here. We are, however, able to provide an explicit characterization for the special case of a single sinusoid in Appendix B.3.

• The previous remark also explains why we choose to restrict $\boldsymbol{\omega}'$ to $\Theta'(\boldsymbol{\omega}) = \Theta \backslash B(\boldsymbol{\omega}, \mu)$. The singular value of $\mathbf{X}(\tilde{\boldsymbol{\omega}})$ when $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Theta$ can be made arbitrarily small by allowing $\boldsymbol{\omega}' \to \boldsymbol{\omega}$. Thus, in this case, we cannot directly use Theorem 3.2 to quantify the number of measurements required. However, this does not necessarily mean that an isometry cannot be provided for closely spaced sinusoids. Indeed, we show in Appendix B.3 that, for $K = 1$, it is possible to provide an isometry no matter how close ω and ω' get.

Proof of Theorems 3.1 and 3.2: We give a proof of Theorem 3.2 along the lines of the proof in [6], where the authors extend the JL lemma (which gives the number of compressive measurements needed to preserve the geometry of a discrete point cloud) to a manifold by sampling the manifold and exploiting its continuity. Details of the proof can be found in Appendix B.1. A similar proof can be given for Theorem 3.1, which we briefly sketch in Appendix B.2.

3.4 Pairwise isometry for frequency estimation of a single sinusoid

In the preceding discussions, we quantify the number of measurements needed to give pairwise isometries for a mixture of K sinusoids in two distinct regimes: when the frequencies $(\boldsymbol{\omega}, \boldsymbol{\omega}')$ are "far apart" and in the limit of $\boldsymbol{\omega}' \to \boldsymbol{\omega}$ (tangent plane isometries). We now consider a single sinusoid ($K = 1$) and provide pairwise isometries for *all* frequency pairs. In order to do this, we consider two regimes of frequency pairs (ω_1, ω_2): closely spaced and well-separated. For the set of well-separated frequencies, say $\{(\omega_1, \omega_2) : |\omega_1 - \omega_2| > \psi\}$, we obtain a bound on the smallest singular value of $\mathbf{X}(\omega) = [\mathbf{x}(\omega_1)\ \mathbf{x}(\omega_2)]$ and use it in Theorem 3.2 to immediately infer the number of measurements needed to guarantee pairwise ϵ-isometry for sinusoids from this set. The challenge then is in providing a similar result for sinusoids whose frequencies are separated by less than ψ. We solve this problem in two stages: first, we use Theorem 3.1 to infer the number of measurements needed to guarantee tangent plane ϵ-isometries for all frequencies (loosely, pairwise isometries for $\omega_1 \to \omega_2$). We then use the continuity of the sinusoidal manifold to extend these tangent plane ϵ-isometries to a pairwise 2ϵ-isometry for closely-spaced frequencies $\{(\omega_1, \omega_2) : |\omega_1 - \omega_2| < \psi\}$.

Theorem 3.3. *Suppose that* $\boldsymbol{A}$ *is an* $M \times N$ *measurement matrix whose entries are drawn i.i.d. from* Uniform $\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$. *Let* $\mathbf{x}(\omega)$ *denote a sinusoid (3.1) of frequency* ω *with weights* $\{h_n\}$ *such that* $\sum |h_n|^2 = 1$. *Let* $H(\omega) = \sum_{n=1}^{n=N} |h_n|^2 e^{j\omega(n-(N+1)/2)}$ *be such that (i) the maxima of* $|H(\omega)|^2$ *that occur at frequencies other than* $\omega = 0$ *(side-lobes) are smaller than some constant* $D < 1$ *(independent of* N*) and (ii)* $|H(\omega)|^2$ *is non-increasing in* $(0, \pi/(2N))$. *Then, for any* $\epsilon > 0$,

$$1 - \epsilon \leq \sqrt{\frac{N}{M}} \frac{\|g_1 \boldsymbol{A}\mathbf{x}(\omega_1) - g_2 \boldsymbol{A}\mathbf{x}(\omega_2)\|}{\|g_1 \mathbf{x}(\omega_1) - g_2 \mathbf{x}(\omega_2)\|} \leq 1 + \epsilon, \ \forall g_1, g_2, \omega_1, \omega_2 \tag{3.8}$$

with high probability when $M = O(\epsilon^{-2} \log(N\epsilon^{-1}(1 - \tau\chi)^{-1}\zeta^{-1}\alpha^{-1}))$ *where* $\tau = 1/\|d\mathbf{x}(\omega)/d\omega\|$, $\chi = |dH(0)/d\omega|$, $\alpha = 1/(N\tau)$ *and* $\zeta = -\frac{N^{-2}}{2}\frac{d^2|H(0)|^2}{d\omega^2}$ *are parameters of the windowing sequence* $\{|h_n|^2\}$.

We give the proof of Theorem 3.3 in Appendix B.3. The condition that (i) $|H(\omega)|^2$ is monotonic in $(0, \pi/(2N))$ and (ii) all side-lobes peaks of $|H(\omega)|^2$ are smaller than an absolute constant $D < 1$ (the main-lobe peak $|H(0)|^2 = 1$ since $\sum |h_n|^2 = 1$) are mild. These conditions are satisfied by windowing sequences $\{|h_n|^2\}$ commonly used for spectral estimation, such as the all-ones, Hamming, Hanning, Triangular and Blackman sequences.

3.5 Remark on generality

We state and prove Theorems 3.1, 3.2 and 3.3 for compressive measurements with projection weights (elements of $\boldsymbol{A}$) taken from Uniform$\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$. In addition to the concentration results on $\|\boldsymbol{A}\mathbf{v}\|^2$ of the form (2.8) which we need to preserve the geometry of a discrete point cloud, we use the fact that the Frobenius norm $\|\boldsymbol{A}\|_F = \sqrt{M}$ w.p. 1 for this choice of distribution. When the elements of $\boldsymbol{A}$ are drawn from other distributions such as the gaussian distribution for which these concentration results on $\|\boldsymbol{A}\mathbf{v}\|^2$ are also available [60], $\|\boldsymbol{A}\|_F^2$, which is the sum of the square of all elements of $\boldsymbol{A}$, can be shown to fall within $M(1 \pm \delta)$ w.h.p. Therefore, the conclusions of Theorems 3.1, 3.2 and 3.3 also apply when the elements of $\boldsymbol{A}$ are drawn from these distributions.

3.6 Number of random projections needed for frequency estimation

In this section, we illustrate how to apply the results established in Chapter 2 to design compressive estimation strategies for frequency and phase estimation of a single sinusoid. We describe an algorithm which attains the CRB for compressive frequency estimation when given "enough" compressive measurements, and show

how to determine how many measurements are enough, based on the threshold behavior of the ZZB.

The measurements are given by

$$\mathbf{y} = e^{j\phi}\mathbf{\Phi}\mathbf{x}(\omega) + \mathbf{z} \tag{3.9}$$

where $\mathbf{x}(\omega) = \left[e^{-j\omega(N-1)/2}\ e^{-j\omega(N-3)/2}\ \cdots\ e^{j\omega(N-1)/2}\right]^T$ is an N-dimensional sinusoid with frequency ω, ϕ is its phase, $\mathbf{\Phi}$ is an $L \times N$ complex valued measurement matrix and $\mathbf{z} \sim \mathcal{CN}(0, \sigma^2\mathbb{I}_L)$. The parameters to be estimated ϕ and ω are both distributed uniformly over $[0, 2\pi]$. For this choice of $\{h_n\}$ (all-ones), when we apply Theorem 3.3 (after normalizing so that $\sum|h_n|^2 = 1$), we see that $M = O\left(\epsilon^{-2}\log\left(N\epsilon^{-1}\right)\right)$ compressive measurements suffice to satisfy the pairwise ϵ-isometry property w.h.p ($(1-\tau\chi)^{-1}\zeta^{-1}\alpha^{-1}$ in Theorem 3.3 is bounded for large N).

In Chapter 2, we denoted the parameter to be estimated by $\boldsymbol{\theta} = [\omega\ \phi]^T$ and the signal manifold $\mathbf{x}(\boldsymbol{\theta}) = e^{j\phi}\left[e^{-j\omega(N-1)/2}\ \cdots\ e^{j\omega(N-1)/2}\right]^T$. We now separate the contributions from the phase and frequency and use $\mathbf{x}(\omega)$ to denote a sinusoid with frequency ω and zero phase ($\phi = 0$).

When we make all N measurements (setting $\mathbf{\Phi} = \mathbb{I}_N$ in (3.9)), the CRB is well known [48]. The FIM in estimating $\boldsymbol{\theta} = [\omega \ \phi]$ is

$$F(\mathbb{I}_N, \boldsymbol{\theta}) = \frac{2}{\sigma^2} \begin{bmatrix} N(N^2-1)/12 & 0 \\ 0 & N \end{bmatrix}, \quad \forall \boldsymbol{\theta}. \tag{3.10}$$

In particular, the CRB on the variance of the frequency estimate (computed as $\mathbf{a}^T F^{-1}(\mathbb{I}_N, \boldsymbol{\theta})\mathbf{a}$ with $\mathbf{a} = [1 \ 0]$) is $\text{CRB}(\mathbb{I}_N, \boldsymbol{\theta}) = 6\sigma^2/(N(N^2-1))$. Note that the CRB is independent of $\boldsymbol{\theta}$.

We must be careful in computing the ZZB because the noiseless signal is a periodic function (with period 2π) of both the phase and the frequency. Thus, the errors in estimating these parameters must be appropriately defined (i.e., the difference between 0 and $2\pi - \epsilon$ is ϵ for small ϵ). The ZZB on the "periodic-MSE" of the frequency estimate is given by (using (27) in [8])

$$\begin{aligned} Z(\mathbb{I}_N, \mathbf{a}) &= \int_0^\pi \max_{\phi' \in [0, 2\pi]} Q\left(\frac{\|\mathbf{x}(0) - e^{j\phi'}\mathbf{x}(h)\|}{\sqrt{2}\sigma} \right) h \ dh \\ &= \int_0^\pi Q\left(\sqrt{\frac{N}{\sigma^2}\left(1 - \left|\frac{\sin(Nh/2)}{N\sin(h/2)}\right|\right)} \right) h \ dh. \end{aligned} \tag{3.11}$$

Suppose now that we make M compressive measurements (setting $\mathbf{\Phi} = \boldsymbol{A}$), choosing M large enough so that the measurement matrix $\boldsymbol{A}$ satisfies the pairwise ϵ-isometry property for the $\left\{e^{j\phi}\mathbf{x}(\omega)\right\}$ signal model. Then, from Section 2.4, we know that the Fisher information with compressive measurements $F(\boldsymbol{A}, \boldsymbol{\theta})$ is well-approximated by $F\left(\sqrt{M/N}\ \mathbb{I}_N, \boldsymbol{\theta}\right)$, the Fisher information with all N mea-

surements at an M/N SNR penalty. Given that we know $F(\mathbb{I}_N, \boldsymbol{\theta})$, computing $F(\sqrt{M/N}\,\mathbb{I}_N, \boldsymbol{\theta})$ is easy: we simply replace σ^2 in (3.10) by $\sigma^2(N/M)$.

When $\boldsymbol{A}$ satisfies the pairwise isometry property, we can show that the ZZB with periodic-MSE also satisfies Theorem 2.3. Therefore, the above arguments regarding the increase in the noise level by a factor of N/M hold true for the ZZB with periodic distortion too. Thus, we get the CRB and the ZZB with compressive measurements to be

$$\mathrm{CRB}(\boldsymbol{A}, \boldsymbol{\theta}) \approx \mathrm{CRB}(\sqrt{M/N}\,\mathbb{I}_N, \boldsymbol{\theta}) = 6\sigma^2/(M(N^2-1))\ \forall \boldsymbol{\theta},$$

$$Z(\boldsymbol{A}, \mathbf{a}) \approx Z(\sqrt{M/N}\,\mathbb{I}_N, \mathbf{a}) = \int_0^\pi Q\left(\sqrt{\frac{M}{\sigma^2}\left(1 - \left|\frac{\sin(Nh/2)}{N\sin(h/2)}\right|\right)}\right) h\ dh.$$

We now illustrate how to predict the number of measurements needed for successful compressive estimation based on the threshold behavior of the ZZB. Consider frequency estimation of a $N = 256$ sinusoid from all N measurements ($\boldsymbol{\Phi} = \mathbb{I}_N$) at a noise level σ^2. In Fig. 3.1, we plot the CRB and the ZZB for this estimation problem as a function of the per-measurement SNR $\stackrel{\triangle}{=} 1/\sigma^2$. For SNRs that are smaller than -30dB, we see from Fig. 3.1 that the ZZB is insensitive to changes in SNR, unlike the CRB which exhibits a linear falloff for all SNRs. However, when the SNR exceeds -10dB, the ZZB exhibits a linear falloff with SNR.

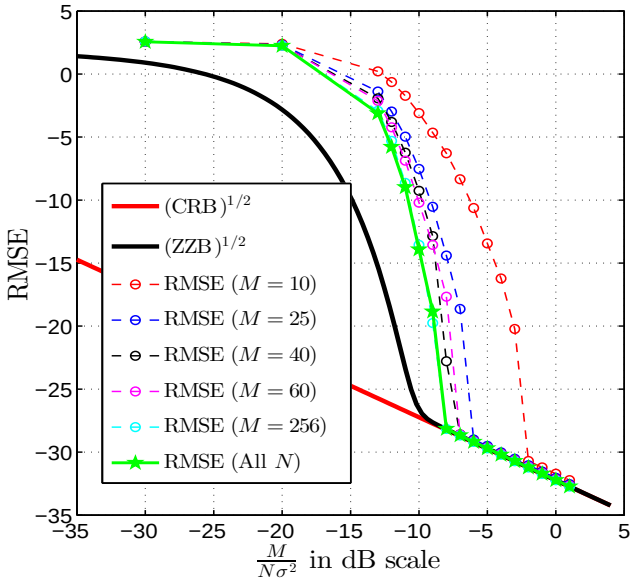

Figure 3.1: RMSE in dB scale for 5 compressive measurement matrices ($\mathbf{\Phi} = \boldsymbol{A}$) with $M = 10, 25, 40, 60, 256$ and the all N measurements case ($\mathbf{\Phi} = \mathbb{I}_N$) plotted against effective per sample SNR $M/(N\sigma^2)$. Overlaid are plots of $\sqrt{\mathrm{CRB}}$ and $\sqrt{\mathrm{ZZB}}$ for all N measurements ($\mathbf{\Phi} = \mathbb{I}_N$) corresponding to this effective SNR. The length of the sinusoid $\mathbf{x}(\omega)$ is $N = 256$.

If we now make M compressive measurements ($\mathbf{\Phi} = \boldsymbol{A}$), the effective SNR is given by $(1/\sigma^2)(M/N)$. We expect "good" estimation performance when this effective SNR exceeds the ZZB threshold, which translates to the following rule of thumb for the number of compressive measurements required:

$$M > N\sigma^2 \times \text{ZZB threshold SNR}, \tag{3.12}$$

Note that the ZZB threshold is computed for the *original* system with all N measurements ($\mathbf{\Phi} = \mathbb{I}_N$), independent of the compressive measurement matrix $\boldsymbol{A}$ and the noise level σ^2. For our specific example of a sinusoid of length $N = 256$, the preceding prescription translates to $M > N\sigma^2/10$, since the ZZB threshold is -10 dB.

We now describe an algorithm whose performance closely follows these predictions: the algorithm approaches the CRB (for a given effective SNR) when the effective SNR exceeds the *ZZB threshold*. This illustrates the efficiency of the algorithm, as well as the accuracy of our design guideline of "sufficient effective SNR."

3.6.1 Algorithm

Let $\mathbf{s}(\omega)$ denote the signal manifold $\mathbf{\Phi x}(\omega)$. Suppose that for the purposes of algorithm design, we ignore the fact that the unknown phase rotation $e^{j\phi}$ has unit

amplitude and estimate the complex gain g and the frequency ω according to the model

$$\mathbf{y} = g\mathbf{s}(\omega) + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbb{I}). \tag{3.13}$$

Let $\mathbf{y}_r(g, \omega)$ denote the residual measurements corresponding to a hypothesized (g, ω)-pair: $\mathbf{y} - g\mathbf{s}(\omega)$. The ML estimates of the gain and frequency $(\hat{g}, \hat{\omega})$ are obtained by optimizing the function

$$S(g, \omega) = -\left\|\mathbf{y}_r(g, \omega)\right\|^2/2 = \langle \mathbf{y}, g\mathbf{s}(\omega)\rangle - 0.5|g|^2\|\mathbf{s}(\omega)\|^2, \tag{3.14}$$

over $g \in \mathbb{C}, \omega \in [0, 2\pi]$ and $\langle \mathbf{x}, \mathbf{y}\rangle = \Re\{\mathbf{x}^H\mathbf{y}\}$. Performing a direct optimization over g and ω is difficult. Therefore, we resort to a two stage procedure, consisting of a detection phase and a refinement phase, which we describe now.

(i) *Detection phase*: First, we notice that for any ω, the optimizing g is given by $\left(\mathbf{s}(\omega)\right)^H \mathbf{y}/\|\mathbf{s}(\omega)\|^2$. Substituting this in the cost function $S(g, \omega)$, we see that the ML estimate of the frequency $\hat{\omega}$ should optimize $G(\omega) = \max_{g\in\mathbb{C}} S(g, \omega) = 0.5|\mathbf{y}^H\mathbf{s}(\omega)|^2/\|\mathbf{s}(\omega)\|^2$. We obtain a coarse frequency estimate by discretizing the frequencies uniformly into a set $F = \{0, 2\pi/(4N), \ldots, 2\pi(4N-1)/(4N)\}$ of size $4N$ and then choosing $q^\star \in F$ that maximizes $G(q), q \in F$. Since the frequency estimation error is substantial (on the order of $1/N$), we call this the detection phase. The gain estimate is given by $\hat{g} = \left(\mathbf{s}(q^\star)\right)^H \mathbf{y}/\|\mathbf{s}(q^\star)\|^2$

(ii) *Refinement phase*: In the second stage, we iteratively refine the gain and frequency estimates. Suppose that after the nth round of optimization, the gain and frequency estimates are given by $\hat{g}_n$ and $\hat{\omega}_n$ respectively (starting off with the estimates from the detection phase). In the $n+1$th round, we refine the frequency estimate by fixing the gain to $\hat{g}_n$ and locally optimizing $S(\hat{g}_n, \omega)$ around $\hat{\omega}_n$ using Newton's method:

$$\hat{\omega}_{n+1} = \hat{\omega}_n - \frac{\partial S(\hat{g}_n, \hat{\omega}_n)/\partial\omega}{\partial^2 S(\hat{g}_n, \hat{\omega}_n)/\partial\omega^2}, \qquad \text{where}$$

$$\begin{aligned}
\frac{\partial S(g,\omega)}{\partial\omega} &= \langle \mathbf{y}_r(g,\omega), g\left(d\mathbf{s}(\omega)/d\omega\right)\rangle\,, \\
\frac{\partial^2 S(g,\omega)}{\partial\omega^2} &= \left\langle \mathbf{y}_r(g,\omega), g\left(d^2\mathbf{s}(\omega)/d\omega^2\right)\right\rangle - |g|^2\|(d\mathbf{s}(\omega)/d\omega)\|^2\,.
\end{aligned}$$

Next, fixing the frequency estimate to $\hat{\omega}_{n+1}$, we get the updated gain after the $n+1$th round to be $\hat{g}_{n+1} = \left(\mathbf{s}(\hat{\omega}_{n+1})^H\mathbf{y}\right)/\|\mathbf{s}(\hat{\omega}_{n+1})\|^2$. Our numerical results are based on applying three such rounds of iterative optimization.

Results: We simulate the performance of the algorithm with $M = 10, 25, 40, 60$ and 256 compressive measurements across *effective per measurement SNRs* $M/(N\sigma^2)$ ranging from -30dB to 1dB using 5×10^4 trials (for each M, we use the same measurement matrix $\mathbf{\Phi} = \boldsymbol{A}$ for all SNR values). The elements of $\boldsymbol{A}$ are picked i.i.d from Uniform$\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$. We plot the Root-Mean-Squared-Error (RMSE) of the frequency estimate versus the effective SNR $M/(N\sigma^2)$ along with

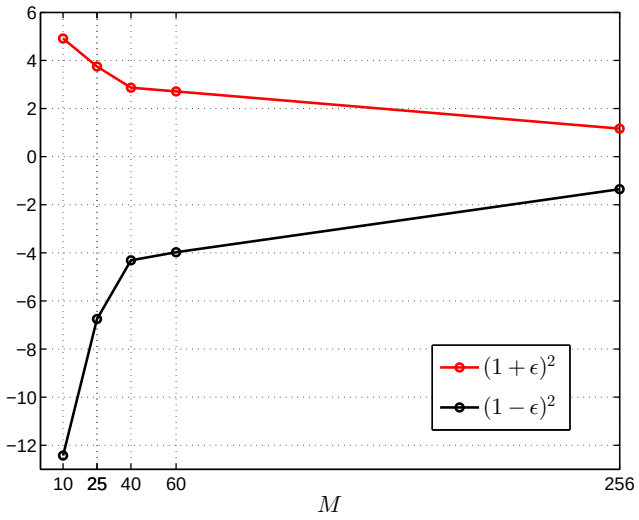

Figure 3.2: *Bounds* on pairwise SNR variation due to pairwise isometry constant ϵ (2.7) for the compressive measurement matrices used in Fig. 3.1. Isometry constant ϵ corresponds to the manifold $\{ge^{j\phi}\mathbf{x}(\omega)\}$ where $g \in \mathbb{R}^+$ and $\phi, \omega \in [0, 2\pi]$.

the CRB and ZZB in Fig. 3.1. We define the effective SNR beyond which the RMSE of the estimate exhibits a linear falloff with SNR in the log-log plot (similar to the ZZB at high SNRs) as the *RMSE threshold*. From our earlier discussions on the number of measurements needed for successful compressive estimation, we expect the RMSE threshold to exceed the ZZB threshold. From Fig. 3.1, we see that the RMSE thresholds for $M = 10, 25, 40, 60$ and 256 measurements are $-2, -6, -7, -7$ and -8dB respectively. All the RMSE thresholds are larger than the ZZB threshold of -10dB as expected. We also evaluate the algorithm for the all N measurements case ($\mathbf{\Phi} = \mathbb{I}_N$) and find that the RMSE threshold in this case is -8dB.

Differences in the isometry constant ϵ explain why the RMSE thresholds are different for different measurement matrices $\boldsymbol{A}$. With increasing number of measurements M, the isometry constant decreases. This trend is shown in Fig. 3.2 where we plot the bounds on the deviation of the pairwise SNRs from M/N, corresponding to $(1 \pm \epsilon)^2$, for the measurement matrices $\boldsymbol{A}$ used in our simulations. (*Note*: These isometry constants correspond to the manifold $\{ge^{j\phi}\mathbf{x}(\omega) : g \in \mathbb{R}^+, \phi, \omega \in [0, 2\pi)\}$ because the algorithm does not use the fact that $g = 1$). When we take few compressive measurements, pairwise SNRs can deteriorate significantly (ϵ is large) and, as a result, the RMSE threshold increases. The bounds on pairwise SNR variation (in Fig. 3.2) when we make 40 and 60 measurements do not differ by much. This illustrates the diminishing improvements in isometry per measurement beyond a point. For the all-N measurements case ($\boldsymbol{\Phi} = \mathbb{I}_N$), the isometry constant $\epsilon = 0$ by definition and therefore the RMSE threshold is close to the ZZB threshold.

When we set $M = N = 256$, the degradation in pairwise SNRs is smaller than 2dB. However, for this extreme case, the RMSE threshold is merely 1dB smaller than that for $M = 40$. This indicates that, for our example of frequency estimation for sinusoid of length $N = 256$, the isometry constant is small enough when we make 40 or more compressive measurements.

To summarize, when the number of measurements M is large enough for the isometry constant ϵ to be small, the number of measurements M necessary obeys the rule of thumb in (3.12), based on ZZB threshold computations for the original system. For our example $N = 256$ sinusoid, this translates to the rule of thumb $M \geq \max\{40, 25.6\sigma^2\}$.

Chapter 4

Scalable and Efficient Geographic Routing in Mobile Ad Hoc Wireless Networks

We investigate the problem of geographic routing in Mobile Ad Hoc Networks (MANETs). In order to implement geographic routing, a relay node needs estimates of the location of its neighbors and the destination node. When the nodes in the network can move, estimates of the locations of neighboring nodes can be maintained by means of periodic *local* broadcasts. Such local broadcasts can be easily accommodated with a constant bandwidth overhead. However, global dissemination of information regarding the locations of moving destination nodes becomes a bottleneck. As noted in prior work, it is possible to reduce the volume of position updates to distant nodes without excessively compromising route quality. While we build on top of this intuition, we show that sending updates to all nodes in the network does not scale even after accounting for this reduction. In

this chapter, we present a solution to this problem by carefully choosing a small subset of nodes in the network as the recipients of position updates made by a node. The main features of the solution presented are:

- Updates are sent to annular rings centered around potential destination nodes. The radii and thickness of these rings grow exponentially with the ring index. The frequency of updates on the other hand *decay* exponentially with ring index.

- Since all nodes in the network do not receive updates corresponding to the destination node, the "best" estimate seen by a packet so far is appended to it. Relay nodes that either do not possess updates corresponding to the destination or have updates of lesser quality use the packet's estimate. A relay node with a better estimate than that corresponding to the packet, overwrites the packet's estimate.

- It can happen that the source does not possess an estimate of the destination's position. In that case, the source node picks an arbitrary direction and forwards the packet along this direction. This direction is indicated in the packet and relays downstream forward along this direction until the packet reaches a relay with a position estimate corresponding to the destination (this relay overwrites the direction field with its estimate).

By choosing protocol parameters appropriately, we show that the proposed position-publish protocol is scalable while also ensuring that the accompanying routing protocol constructs efficient routes with high probability.

Outline: We start in Section 4.1 by reviewing prior art. In Section 4.2, we lay out the model and the accompanying scaling used in our computations. We focus on greedy geographic forwarding with location errors and give relevant definitions in Section 4.2.1. We derive necessary conditions on the communication radius to ensure that all greedy routing decisions "agree" with the straight line joining the relay at which they are made to the position estimate used to make the decision. We state this result as Theorem 4.1 in Section 4.2.1. We go over the cost of multicasting to a region (in terms of the number of transmissions needed), which is a building block for our scalability computations in Section 4.2.2. We show in Section 4.3 that the naive strategy of issuing updates (of necessary fidelity) to all nodes in the network does not scale. The proposed position-publish protocol, which overcomes this scalability bottleneck by issuing updates to a small subset of nodes is presented along with the accompanying routing protocol in Section 4.4. In Section 4.5, we provide necessary conditions on the protocol parameters for scalable position-publish and reliable & efficient routing. We report simulation results for one such choice in Section 4.6.

4.1 Related Work

Since we are concerned with large-scale networks, our communication model and notion of scalability are guided by the relevant asymptotic results of Gupta and Kumar [28] [29]. We postpone detailed discussion of these to Section 4.2.

The literature on MANET routing and on geographic routing (for stationary or mobile nodes) is vast, hence we restrict attention here to prior work that is most closely related to our approach (many of the references we cite provide good discussions on the state of the art). DREAM [7] considers geographic routing when the frequency of location updates is reduced as the distance from the updating node increases. While this intuition is the starting point for our scheme as well, we show that location updates made to all nodes as in DREAM are not scalable. A similar intuition is also behind the Hazy Sighted Link State (HSLS) algorithm in [52], in which link state updates are sent less frequently to distant nodes. HSLS is designed based on minimization of the sum of the overhead due to route suboptimality and location updates. However, the overhead computations in [52] show that HSLS is not scalable. An intuitive reason for this is that all nodes must have a roughly consistent view of the network for successful link state routing, whereas geographic routing only requires that an appropriate subset of nodes have location updates from a given destination node. GLS [35] is a spatially

hierarchical quorum based scheme for position lookups, but is not designed to work in networks with pervasive movement.

MLS[24] proposes a "lazy" hierarchical position lookup service in which updates are published to certain fixed geographical regions. It is similar in spirit to our scheme, in that it is able to guarantee a constant route stretch without requiring that all nodes in the network obtain location updates, but the updates in our scheme are published to regions which are different, in general, for different nodes. It is worth mentioning that MLS builds on an earlier scheme termed LLS[2], which structures location updates to areas centered around the destination node, as in our scheme. The key difference of [24, 2] from our work is they do not relate the routing overhead to network transport capacity, and do not provide means to vary the tradeoff between route stretch and overhead.

Prior work [28] derives the scaling of the communication radius needed for network connectivity. The critical radius needed for successful greedy geographic forwarding is derived in [62]. We derive necessary conditions on the communication radius to ensure that a routing decision made at a relay node using an arbitrary position estimate does not deviate "too much" from the straight line joining the relay to this estimate. Unlike [62], our results hold even when the position estimates using which forwarding decisions are made do not correspond to current locations of network nodes. The node distribution in [62] is assumed

to be given by a Poisson Point Process (PPP) of uniform density. While we make the assumption that the number of nodes is fixed and distributed uniformly and independently at random, we note that our scaling results also hold for the PPP model in [62].

For the routing scheme presented in this dissertation and the preceding references, mobility is a nuisance that increases routing overhead. However, when delay in message delivery is not an issue, Grossglauser and Tse have shown in [26] that mobility can actually help us get around the transport capacity limits derived by Gupta and Kumar [29]. In a similar spirit, mobility can be exploited to reduce the overhead of location updates, as argued in [27][20]. However, this is not the regime of interest to us, since we are interested in delivering packets to their destinations with minimal delay.

4.2 System Model

We consider a network of n nodes in the two-dimensional plane. The deployment region is a square of area n from which initial node positions are picked uniformly and independently at random. Therefore, the nominal node density is fixed at one node per unit area.

Connectivity: We assume that the communication radius for all n nodes is fixed at $r = r(n)$, and that it is chosen so that the network is connected. It is shown in [28] that connectivity requires that r must scale so that

$$\pi r^2 = (1 + \epsilon) \log n, \tag{4.1}$$

for any constant $\epsilon > 0$. While such a choice corresponds to a communication radius $r(n) = \Theta\left(\sqrt{\log n}\right)$, we note that we can scale down to a constant communication radius by scaling the deployment region as $n/\log n$ (along with suitably scaling other parameters) rather than as n.

Scalability: We use the protocol model for interference proposed in [29] for our scalability computations: Each transmission precludes the reception of any other transmission within a disc of radius $(1 + \Delta)\, r$, where r is the communication radius (Δ being an absolute constant). Thus, if the bandwidth available for communication is W, then the maximum number of simultaneous useful transmissions available per time slot, denoted by $T_A(n)$, scales as $\Theta\left(Wn/r^2(n)\right)$. Denoting by $T_U\left(n\right)$ the average of the total number of simultaneous transmissions needed per unit time to sustain a protocol (overhead) across all nodes, we employ the following definition for the scalability of a protocol.

Definition. *We refer to a protocol as* ***scalable*** *if* $T_U(n) = O\left(n/r^2(n)\right)$. *In this case, the overhead needed for the protocol can be accommodated with a suitably chosen constant bandwidth* W.

The preceding definition assumes that the load induced on the network as a result of position updates is uniform in space and time (this holds for our mobility model, which is described next).

Mobility Model: Every node in the network is mobile, executing 2D Brownian motion of mean square velocity $2\sigma^2$, with reflection at the boundaries of the deployment region (assumed to be square for convenience). We note that for the choice of square deployment region with initial node positions picked independently and identically at random from the uniform distribution, 2D Brownian motion with reflection at the boundaries results in *instantaneous* node positions (marginals in time) also given by the uniform distribution, with each node's position being independent of the other nodes in the network.

While we choose the Brownian motion model for its analytical tractability, we note that our scalability results hold more broadly: scalability depends on how distant nodes perceive the mobility of a destination node, and a large class of randomized models for *local* mobility look like Brownian motion when viewed from far away and at large time scales. For example, consider a version of the random

waypoint model[33] in which each node chooses a new speed $V_l \geq 0$ independently and identically from a distribution and direction Φ_l uniformly over $[0, 2\pi]$ for a duration D_l, where the times $D_l > 0$ are independent and identically distributed random variables. It can be shown that, over large time scales, this model can be viewed as Brownian motion with mean square velocity $(\mathbb{E}V_1^2\ \mathbb{E}D_1^2)/\mathbb{E}D_1$.

4.2.1 Greedy geographic forwarding with location errors

The routing protocol that we consider is the following: When a packet arrives at a node which is not the intended destination node, this node forwards the packet to the neighbor that is the closest to the current estimate of destination node's position (this position estimate may be available at the relay node or may have been appended to the packet by an earlier relay node). We refer to such a local routing strategy as **greedy geographic forwarding**. To facilitate this, we assume that every node has perfect knowledge of the location of its neighboring nodes.

We want to ensure that greedy geographic forwarding with imperfect location estimates is *reliable* and that successful routes are *efficient*. In the forthcoming discussions, we formally define these properties of greedy geographic forwarding protocols.

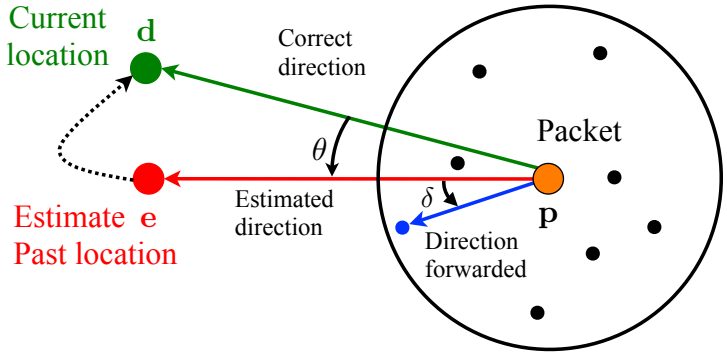

Figure 4.1: Rate of progress depends on the angle between the forwarded and correct directions. The circle around the packet is the neighborhood of the relay node (given by the communication radius).

Definition. *We refer to a routing protocol as* ***reliable*** *if it delivers packets to their destination with high probability (w.h.p.). We note that greedy geographic forwarding with* perfect *location information is reliable when* ϵ *in* (4.1) *exceeds* $\epsilon_0 \approx 1.6$ *[62].*

When information about the destination's location is imperfect, the natural approach is to route the packet along the best estimate of the direction of the destination, possibly updating this estimate after each hop, until we get close enough that the destination is within the communication radius. If the angle between the correct and the forwarded direction is θ, then the progress towards the destination per unit distance traveled is $\cos\theta$, so that we would like θ to be small.

We now observe that we can afford to be sloppier in our estimate of the destination's location when we are farther away. Let us denote the distance between points $\mathbf{a}$ and $\mathbf{b}$ by $\ell(\mathbf{a}, \mathbf{b})$.

Definition. *We define the* ***uncertainty*** U *of the position estimate* $\mathbf{e}$ *of the destination at* $\mathbf{d}$ *available to a packet at* $\mathbf{p}$ *as the ratio of distances* $U = \ell(\mathbf{d}, \mathbf{e})/\ell(\mathbf{d}, \mathbf{p})$.

When we fix the uncertainty of the available estimate to U, it is easy to show that the worst case (largest) value of θ is given by $\sin\theta = U$. Thus, if we wish to ensure that the angle between the estimated direction and the correct direction is less than θ, we can allow for the localization error to be larger when the packet is farther away (i.e., $\ell(\mathbf{d}, \mathbf{e})$ can be as large as $\ell(\mathbf{d}, \mathbf{p}) \times \sin\theta$).

Definition. *We define the ratio of the length of the source-destination packet trajectory to the source-destination distance to be the route* ***stretch****.*

Definition. *The reciprocal of route stretch is a measure of routing efficiency and we refer to a routing protocol with a bounded stretch as an* ***efficient*** *protocol.*

Bounded uncertainty leads to bounded stretch: Now, suppose that the uncertainty seen by a packet is always less than $U_{\max} < 1$, so that the worst case angle between the correct and estimated directions always satisfies $\theta \leq \arcsin(U_{\max})$. This implies that $\cos\theta \geq \sqrt{1 - U_{\max}^2}$ and the route stretch will be bounded by $1/\sqrt{1 - U_{\max}^2}$.

The preceding argument assumes that the forwarded direction *perfectly matches* the estimated direction. However, the neighbor of the relay node node (at $\mathbf{p}$) which is the closest to the estimate $\mathbf{e}$ is never (with probability one, for our model of network nodes as points) on the line joining $\mathbf{p}$ and $\mathbf{e}$. Therefore, we expect some disagreement between the actual direction along which a packet is forwarded and the *desired* direction corresponding to the estimate $\mathbf{e}$ used (we sketch this in Figure 4.1). The amount of this disagreement depends on the availability of neighbors around the relay node along the estimated direction and thereabouts. Providing route stretch guarantees while taking this variability into account requires that we choose a large enough communication radius (4.1) by an appropriate choice of ϵ. This ensures that w.h.p., the amount of disagreement between the forwarded direction and the desired direction (denoted by δ in Figure 4.1) is small for all greedy forwarding decisions. We summarize this in the following theorem. We note that this theorem holds for an *arbitrary* position estimate, which need not correspond to the current location of any of the n nodes in the network. The theorem assumes that at any time instant all n nodes are distributed uniformly and independently at random over a square of area n. This assumption holds true for the deployment region and mobility model considered.

Theorem 4.1. *For any $0 < \delta \leq \pi/3$, the following statement holds w.h.p. The maximum disagreement between the direction along which a packet is forwarded and the desired direction given by the straight line joining the relay with the estimate using which this greedy routing decision is being made is at most δ, when ϵ in the choice of communication radius $r = \sqrt{(1+\epsilon)/\pi \log n}$ is a large enough constant so that*

$$1 + \epsilon > \pi/(\delta - \sin\delta) \tag{4.2}$$

and the estimate is at least $2r$ away from the relay node

Corollary 4.2. *Suppose that all nodes within $2\left(1 - U_{max}\right)^{-1} r$ of one another know each other's locations perfectly. When the uncertainty seen by packets is bounded by $U_{max} < 1$, routing with* imperfect estimates *is* reliable *if ϵ satisfies* (4.2) *for some $0 < \delta < min\{\pi/3, \pi/2 - \arcsin U_{max}\}$.*

Remark: When uncertainty is smaller than $U_{\max}$ and $\ell(\mathbf{p}, \mathbf{d}) > 2\left(1 - U_{\max}\right)^{-1} r$, it can be shown that $\ell(\mathbf{p}, \mathbf{e}) > 2r$. Therefore, from Theorem 4.1 we have that, if ϵ satisfies (4.2), the stretch of the segment of the trajectory from the packet source up until a distance of $2\left(1 - U_{\max}\right)^{-1} r$ from the destination node is bounded by

$$1/\cos\left(\arcsin\left(U_{\max}\right) + \delta\right) \leq \left(\sqrt{1 - U_{\max}^2} - \sqrt{2}\delta\right)^{-1}.$$

Proofs of Theorem 4.1 and Corollary 4.2: We give the proofs of both the theorem and its corollary in Appendix C.1. The main ingredients of the proof of Theo-

rem 4.1 are the following: (i) We identify $\lceil 2\pi/\delta \rceil$ "anchor" regions in the neighborhood every node (within the communication radius $r(n)$). We show that when all anchor regions around a node are occupied, greedy forwarding decisions made by this node are such that the angle between the forwarded and estimated directions is smaller than δ (ii) We use the union bound and show that if $\pi r^2(n) = (1+\epsilon)\log n$ with ϵ satisfying (4.2) all anchor regions in the network are occupied w.h.p.

A natural approach to guarantee a worst case route stretch is to employ a position-publish protocol that maintains uncertainty below a level $U_{\max} < 1$ throughout the network. We show in the next section that such protocols would not scale. Before doing that, we round out this section by quantifying the cost of multicasting information to a specific region (which is a basic building block for the position-publish protocol discussed here).

4.2.2 Cost of Multicast

We note here for future use that the *optimal* number of transmissions needed to multicast a message to all nodes in a connected region A, $C(A) = \Theta(|A|/r^2)$.

To see this, we note that, in order for every node in A to have listened to the message at least once, the area $|A|$ has to be "tiled" by circles of area πr^2. Thus, $C(A) = \Omega(|A|/r^2)$.

To provide an upper bound, we need a constructive scheme that multicasts messages to all nodes in A. When $\pi r^2 = (1+\epsilon)\log n$, with ϵ a sufficiently large constant, we can use a result from [29] to tile the area *a priori* into $\Theta(|A|/r^2)$ tiles such that there exists at least one node per tile and every node in a tile can communicate with every other node in its tile and all the nodes in its neighboring tiles. Thus, the resultant network of tiles is connected. We designate one node per tile (say the one with the smallest node ID) to transmit and listen while others merely listen. The first instant when the designated node in each tile receives the multicast message, it airs the message to all nodes in its range in a collision free manner (using an appropriate MAC) by means of a wireless broadcast. Thus, the message is multicast to all nodes in A with $\Theta(|A|/r^2)$ transmissions, which proves that $C(A) = \Theta(|A|/r^2)$.

Note that, even though nodes are mobile, the tiling of the network can be done *a priori* as in [29], and a node leader elected based on the node of smallest ID occupying the tile (this only requires nodes to have information regarding their neighbors, which any position-publish protocol provides).

We have been able to use the results in [29] because for the initial node deployment and mobility model considered, the instantaneous distribution of nodes in the network is given by the uniform distribution, with each node's position being independent of the other nodes in the network

4.3 A Non-Scalability Result

While maintaining the uncertainty guarantees an upper bound on route stretch, we now show that maintaining a uniform uncertainty throughout the network, which requires updating all nodes in the network, does not scale.

In order to maintain uncertainty of at most $U_{\max}$, location updates from a particular node (say v) must reach all nodes that are a distance z away from it if it moves a distance roughly equal to $(U_{\max} z)/(1 + U_{\max})$. For our Brownian motion model, the mean time to move this distance is $(U_{\max}^2 z^2)/\left(2\sigma^2 (1 + U_{\max})^2\right)$, and the average frequency of updates to these nodes is the reciprocal of this time. The area of a small ring at distance z is $2\pi z \, \mathrm{d}z$ and, as shown in Section 4.2.2, the minimum number of transmissions needed to inform all nodes in this ring is $C\,(2\pi z \, \mathrm{d}z) = \Theta\,(z \, \mathrm{d}z/r^2)$.

Remembering that the diameter of the network is $\Theta\,(\sqrt{n})$, the average number of transmissions allocated to a node v per unit time t_U must satisfy:

$$\begin{aligned} t_U &\geq \frac{2\sigma^2 C\,(\pi k_1 r^2)}{k_2^2 r^2} + \int_{k_1 r}^{k_3\sqrt{n}} \frac{2\sigma^2 C\,(2\pi z \, \mathrm{d}z)}{k_4 z^2} \\ &= \sigma^2 \frac{k_5}{r^2} + \sigma^2 \frac{k_6}{r^2} \log\left(\frac{k_3\sqrt{n}}{k_1 r}\right) \end{aligned} \tag{4.3}$$

for some constants k_1, k_2, k_3, k_4, k_5 and k_6. The first term corresponds to broadcasts to a circle of radius bigger than r to ensure all nodes have accurate lists of neighbors, while the second term corresponds to the location updates to distant

nodes aimed at preserving uncertainty. The inequality in (4.3) is because we have ignored the rate needed to preserve updates in space (other network nodes are mobile and so updates made to a certain region in space will not be available in that region indefinitely). So $T_U(n) = n \times t_U = \Omega\left((\sigma^2 n \log n)/r^2\right)$.

Maintaining uniform uncertainty does not scale: The ratio of required overhead to sustainable capacity is therefore given by

$$T_U(n)/T_A(n) = \Omega\left(\left(\sigma^2 \log n\right)/W\right)$$

which blows up (albeit slowly) for large n. Thus, a strategy of maintaining an upper bound on uncertainty throughout the network (and thus bounding route stretch) does not scale.

Clearly, in order to provide guarantees on route stretch, the angle between the true and estimated directions towards the destination cannot be too large. But what we have just shown implies that we must appropriately choose a *subset* of nodes to update in order to reduce the routing overhead enough that the protocol can scale. This observation motivates the proposed protocol described in the next section.

4.4 Proposed Protocol

We now describe a *scalable* position-publish protocol, and an accompanying greedy geographic forwarding protocol which works with imperfect position estimates and is *reliable* and *efficient*. Before presenting the details of the position-publish protocol and the routing protocol, we provide an overview of the position-publish protocol and state the necessary conditions on the protocol parameters for scalable position-publish and for reliable and efficient routing.

4.4.1 Overview of the position-publish protocol

We give a summary of the position-publish protocol executed by a typical node that is a potential destination (we call this the *destination node* henceforth) while deferring the details to Section 4.4.3. The destination node directs its updates to geographic regions structured as annular rings around its current position, indexed as $i = 0, 1, ..., K$. The position-publish algorithm is executed in a parallel fashion for each ring index.

An update ring corresponding to index i has inner radius of the r_i and thickness $d_i \ll r_i$. Therefore, the geographical region to which an update is issued is specified by the center $\mathbf{c}$ of the ring and its ring index i. An update issued to the i-th ring is retained for a duration of T_i by the nodes that receive this update

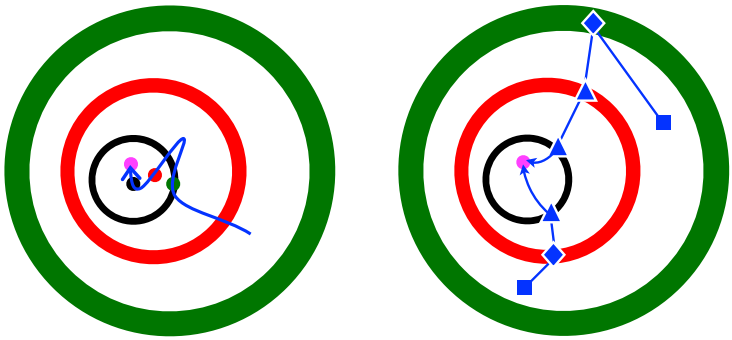

Figure 4.2: Left: Update rings corresponding to three consecutive ring indices $l-1$ (black), l (red) and $l+1$ (green). The position estimates (also centers of the update ring) are marked $\bullet$ on the destination's trajectory (blue) using corresponding colors. The current position of the destination is the magenta $\bullet$. Right: Two packet trajectories (blue) starting from nodes marked $\blacksquare$ in between the l-th and $(l+1)$-th update rings converging to the destination (magenta $\bullet$). The packets are launched in arbitrary directions and acquire their first estimate (bootstrap) inside the l-th and $(l+1)$-th update rings respectively (marked $\blacklozenge$). They progressively refine their estimates when they cut through lower indexed rings (marked $\blacktriangle$).

after which it is discarded. We refer to this time duration over which a particular update is retained as its *lifetime* and those updates whose lifetimes have come to pass as *expired* updates. The parameters r_i, d_i and T_i, which define the ring index i all scale exponentially with the ring index i.

The position-publish protocol proactively publishes position updates to those ring indices whose updates are at the cusp of expiry, thereby ensuring that an update ring (which has not yet expired) corresponding to each of the $K+1$ update indices encircles the destination node. We depict the typical configuration of update rings around the destination node in Figure 4.2 (left). Three update

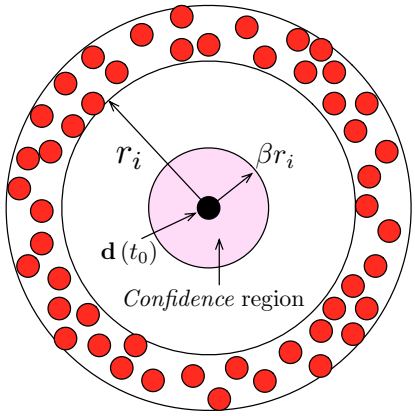

Figure 4.3: Update made to the i-th ring at time t_0 by the destination node (in black). The nodes shaded red receive this update.

rings corresponding to consecutive ring indices $\{l-1, l, l+1\}$ are highlighted in Figure 4.2. The position-publish algorithm runs in parallel for different ring indices (with different typical lifetimes $\{T_{l-1}, T_l, T_{l+1}\}$). As a result, the three rings are centered around different points on the destination's trajectory.

Denoting the position of the destination at time t by $\mathbf{d}(t)$, the position estimate $\mathbf{e}$ of an update points to the location of the destination node at the time of issue (i.e., an update made at time t_0 satisfies $\mathbf{e} = \mathbf{d}(t_0)$). Each update also comes with a guarantee on the quality of its position estimate $\mathbf{e}$, which can be tuned by a parameter β satisfying $0 < \beta < 1$: The destination node is understood to remain within a circle of radius βr_i around this position estimate $\mathbf{e}$ (i.e., $\ell(\mathbf{d}(t), \mathbf{e}) < \beta r_i$ until this update expires). We refer to this region as the *confidence region* of the update. Figure 4.3 illustrates a typical update ring and the *confidence region*

associated with it. Confidence region guarantees are essential for the reliability and efficiency of the accompanying routing protocol. Two representative packet trajectories, which make use of the information disseminated by the position-publish protocol are sketched in Figure 4.2 (right).

4.4.2 Protocol parameter choices

We summarize the regime of operation of the proposed protocol in Theorem 4.3. We provide a proof of Theorem 4.3 in Section 4.5 via Lemmas 4.4, 4.5 & 4.6.

The inner radius r_i, thickness d_i and timer duration T_i of update rings grow exponentially with the ring index i as

$$r_i = r_0\alpha^i, \quad d_i = d_0\alpha^{\mu i}, \quad T_i = T_0\alpha^{\gamma i}.$$

The zero-order ring defined by r_0, d_0 and T_0, the ring scaling exponents α, μ and γ and the confidence region parameter β are the tunable parameters of the proposed position-publish protocol. Let $U_{\max}$ denote the maximum uncertainty seen by a packet after it acquires an initial estimate of the destination's position (we refer to this process of acquiring an initial estimate, as *bootstrapping*, when we explain the routing protocol in Section 4.4.4). We show in Appendix C.3 that

$$U_{\max} = \alpha\beta/(1-\beta)\,. \tag{4.4}$$

We assume hereon that all nodes within $2\left(1-U_{\max}\right)^{-1} r$ of one another know each other's positions perfectly (via *local* broadcasts), where r is the communication radius, chosen to satisfy $\pi r^2 = (1+\epsilon)\log n$. Such local broadcasts are necessary for maintaining neighbor lists and can be accommodated within a constant bandwidth.

Theorem 4.3. *The proposed position-publish protocol is* scalable *and the associated routing protocol is* efficient *and* reliable *when* $r_0/\beta, d_0$ *and* $\sqrt{T_0}/\sigma$ *scale as* $\Theta(r)$, $\alpha > 1$, $0 < \beta < 1/(1+\alpha)$, $1/3 < \mu < 1$, $1+\mu < \gamma < min\{2, 4\mu\}$ *and* $1+\epsilon > \pi/(\delta - \sin\delta)$ *for some constant* $0 < \delta < min\left\{\pi/3, \pi/2 - \arcsin U_{max}\right\}$.

An example of parameter choices satisfying Theorem 4.3: (i) confidence region parameter $\beta = 0.25$, (ii) order-zero ring specified by $r_0 = r/\beta$, $d_0 = 2r$, $T_0 = (1/8)\left(\beta r_0/\sigma\right)^2$ and (iii) ring scaling parameters $\alpha = 2$, $\mu = 0.55$, $\gamma = 1.95$. We perform simulations for this choice of parameters and present results in Section 4.6.

4.4.3 Position-publish protocol

We now present the details of the position-publish protocol. There are two kinds of location updates: normal and abnormal updates.

Normal update: A normal update published at time t_0 to ring i (of radius r_i) specifies the center of the ring as the current location of the destination node. i.e.,

$\mathbf{c} = \mathbf{d}(t_0)$. The update points to the current location $\mathbf{e} = \mathbf{d}(t_0)$ and has a lifetime of T_i after which the nodes which receive the update discard it.

Abnormal update: An abnormal update is sent when the destination leaves the confidence region for a prior *normal* update before the timer for the latter update expires. For example, for the normal update at time t_0 described above, if the destination node crosses the boundary of the confidence region at time $t_1 < t_0 + T_i$ (i.e., $\ell(\mathbf{d}(t_1), \mathbf{e}) = \ell(\mathbf{d}(t_1), \mathbf{d}(t_0)) > \beta r_i$), then we send an abnormal update to the ring centered at the *prior* update. That is, we send an update specifying the current location $\mathbf{e} = \mathbf{d}(t_1)$ to a ring of index i centered at $\mathbf{c} = \mathbf{d}(t_0)$ with a timer $T_i - (t_1 - t_0)$ (spanning the remaining lifetime of the invalidated update).

When we choose the protocol parameters within the regime prescribed in Theorem 4.3, the probability of abnormal updates tends to zero as the ring index increases. However, we include abnormal updates to ensure that stretch guarantees are met.

Triggers for new updates: A normal update is performed whenever the timer for a prior normal update expires. This is depicted in Figure 4.4 (left). When a destination node moves out of the confidence region of a *normal* update whose timer has not expired, then *two* updates are performed: a normal update to a ring centered around the current location and an abnormal update centered around the old location (at which the invalidated, but as yet unexpired, normal

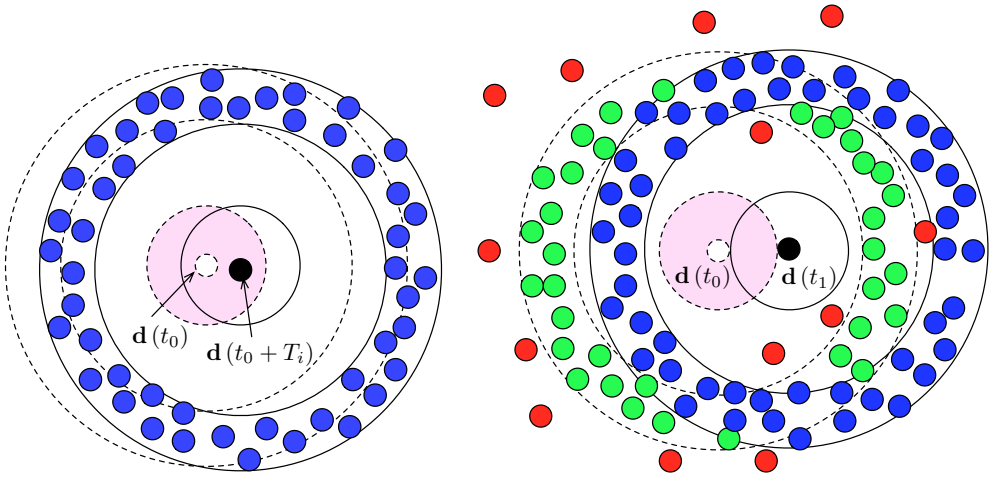

Figure 4.4: Left: Typical scenario of the destination node staying within the *confidence region* of the update made at t_0, when it expires at $t_0 + T_i$. A new *normal* update of lifetime T_i is made to the ring i and is received by the blue relay nodes. Right: An unlikely situation where at time $t_1 < t_0 + T_i$, the destination node leaves the *confidence region* of the update before it expires, thus requiring an *abnormal* update (received by the green relays) of lifetime $T_i - (t_1 - t_0)$ and a *normal* update of lifetime T_i (received by the blue relays). Relays marked red, *outside* the two update rings possess stale unexpired updates made at time t_0 and these updates can be applied to packets only if these relays re-enter the ring centered at $\mathbf{d}(t_0)$.

update was made). The abnormal update lasts for the remaining lifetime of the invalidated update. This is shown in Figure 4.4 (right). When a destination node moves out of the confidence region of an *abnormal* update whose timer has not expired, then *one* update is performed: an abnormal update centered around the old location which lasts for the remaining lifetime of the invalidated abnormal update. Abnormal updates prevent invalidated updates from influencing packet trajectories. The destination node maintains a list of updates published by it, so that it can publish new updates when these updates time out or when their guarantees are invalidated. Updates whose guarantees have been invalidated, are deleted from this list once the aforementioned compensatory action (of issuing new updates) is taken.

Spatial validity of updates: An update (whose timer has not yet expired) can only be used for geographical forwarding if the relay node is in the ring to which the update was made (specified by its ring center $\mathbf{c}$ and ring index i). Thus, once a node moves out of that ring, it can no longer use the information it received about the destination's location when in the ring. It will use this information if it moves into the update ring again. While this may seem overly restrictive, this constraint on the spatial validity of updates enables us to use abnormal updates to reinstate the confidence region guarantees needed for reliable & efficient routing.

Update propagation: In order to limit the traffic generated by a location update, the destination node sends the update packet in a specified direction until it hits the ring it is intended for, at which point it "expands" into a multicast message. Specifically, the destination launches the packet in an arbitrarily chosen direction **u**, which is indicated in the packet. Each intermediate node examines the packet to see if it is in the specified ring. If not, it simply forwards the packet in the direction **u**. Once the packet reaches a node in the update ring, that node repackages the update as a multicast packet for all nodes in the update ring. All nodes that receive this multicast message store the position update for the destination node (overwriting previous updates for the destination node with the most recent update). While large "holes" in the deployment region can disrupt update propagation and expansion, we note that this can be handled when we assume that the shape of the *deployment region* (expected to be static) is known to all potential destination nodes. The destination nodes can thus choose launch direction(s) **u** so as to avoid disruption of update propagation and expansion.

4.4.4 Routing protocol

We now consider the problem of routing a packet to a destination which proactively publishes its location as described in Section 4.4.3. The packet contains a

field indicating the destination identity, the "best" estimate of its location and the ring index & time of update corresponding to this estimate. Intermediate nodes use this field for geographic forwarding, and are allowed to overwrite it if they have a "better" active estimate of the destination's location. An estimate is active only when the relay node's current location satisfies the spatial validity constraints of the update. An estimate is considered "better" only if its ring index is smaller than that of the packet or if its ring index is the same as that the packet, but the update is more recent than that of the packet. The ring index is given more importance than the time of update because of the guarantees given by the destination node through its layered update scheme.

If the source node does not possess an active update, then it chooses a random direction to relay the packet along: this is indicated in the packet by means of a vector indicating this direction (in the position estimate field), time of update $-\infty$ and ring index ∞. Until the packet reaches a node with an active update all intermediate nodes relay the packet along this direction. When the packet hits a node with an active estimate, it is said to have *bootstrapped*. If the packet reaches the boundary of the network before bootstrap, it bounces off the boundary by reflection (by a boundary node changing the direction field).

The parameters α and β of the position-publish protocol limit the potential geometry of *normal* update rings around the destination node and are chosen to

ensure that before the packet reaches the estimate it possesses (say corresponding to the ring index l+1), it meets a smaller indexed update ring (the ring index l) and acquires the corresponding estimate (we detail this in Appendix C.3). Therefore, the amount by which the packet's estimate $\mathbf{e}$ can disagree with the true location of the destination $\mathbf{d}$, which can be no larger than the radius of the confidence region corresponding to the present estimate βr_{l+1}, progressively decreases after bootstrap and the packet eventually reaches the estimate corresponding to the ring indexed 0, whose confidence region guarantee ensures that the destination is no further than βr_0 away from this location. This motivates the choice of the radius of the 0-order ring, $r_0 = r/\beta$ in Section 4.5, thereby ensuring successful packet delivery (we assume that nodes within $2\left(1 - U_{\max}\right)^{-1} r > 2r > \beta r_0 = r$ know each other's locations perfectly). Two such converging packet trajectories are sketched in Figure 4.2 (right).

4.5 Scalability, Reliability and Efficiency

We now derive the design guidelines for protocol parameter choices summarized in Theorem 4.3 in Section 4.4.2. Recall that the inner radii of the update rings scale up exponentially with ring index: $r_i = r_0\alpha^i$, where $\alpha > 1$. So does the ring thickness, but at a slower rate: $d_i = d_0\alpha^{\mu i}$, $0 < \mu < 1$. The timer durations also

scale up exponentially: $T_i = T_0\alpha^{\gamma i}$. The behavior of the protocol depends on the parameters: r_0, d_0, T_0, α, β, γ and μ. In this section, we present three lemmas constraining the parameters for each of our design objectives: Lemma 4.4 for scalability of the position-publish scheme, Lemma 4.5 for routing reliability, and Lemma 4.6 for routing efficiency. Theorem 4.3 simply represents the intersection of the conditions for these three lemmas.

We start off by choosing the radius of the confidence region of the innermost update region, βr_0, equal to the communication radius r (i.e., $r_0 = \Theta(\sqrt{\log n})$). This ensures that before the packet is forwarded to the node closest to the estimate corresponding to this zero-order ring, which can disagree with the true location of the destination by at most $\beta r_0 = r$, the packet is within a distance of $2r$ of the destination, thereby acquiring the true location (we assume nodes within $2\left(1 - U_{\max}\right)^{-1} r > 2r$ of one another know each other's positions perfectly by means of local broadcasts). For this choice of r_0, the number of rings K scales as $O\left(\log n\right)$. To see this, we note that we need the radius of the outermost update ring $r_K = r_0\alpha^K$ to roughly equal the network diameter $\sqrt{2n}$ (the deployment region is a square of area n), and this yields $K = \Theta\left(\log\left(n/r_0^2\right)\right)$.

4.5.1 Position-publish Scalability

Computing the average cost of updates to a particular ring index i is the key step to computing the routing overhead. For proving scalability, it is the behavior for large i that is the most relevant. For 2D Brownian motion, it can be shown that the probability of exiting a circle of radius βr_i within the timer duration T_i tends to zero, as long as T_i grows slower than r_i^2. For $T_i = T_0\alpha^{\gamma i}$ and $r_i = r_0\alpha^i$, this is satisfied for large i as long as $\gamma < 2$, which we henceforth take as a constraint. This implies that the rate of abnormal updates tends to zero (with ring index i), and that the update rate F_i for ring i is approximately $1/T_i$ for large i. In fact, $F_i \leq \Phi/T_i$ for all i, where Φ depends only on $\beta r_0/\left(\sigma\sqrt{T_0}\right)$. We choose T_0 so that $r/\left(\sigma\sqrt{T_0}\right) = \Theta(1)$ and as a result $F_i = \Theta(1/T_i)$ for all i.

As described in Section 4.4.3, a position update to ring i goes on a straight line until it hits the ring, and then is multicast in the ring. The area of the i-th ring ($i \geq 1$) is $A_i = \pi\left((r_i + d_i)^2 - r_i^2\right) = \pi d_i^2 + 2\pi r_i d_i = \Theta(r_i d_i)$ (since the radius r_i scales faster than the thickness d_i). From Section 4.2.2, we know that the number of transmissions to multicast in this area is $C(A_i) = \Theta(|A_i|/r^2)$. Proceeding along the straight line takes $\Theta(r_i/r)$ transmissions, which can be ignored in comparison to the preceding. Thus, the number of transmissions for an update to the i-th ring is $\nu_i = \Theta(r_i d_i/r^2)$. The average rate of transmissions corresponding to updates for

a typical destination node, which we term the *average overhead rate,* is therefore given by

$$t_U = \sum_{i=1}^{K} F_i \nu_i = \Theta\left(\sum_{i=1}^{K} \frac{r_i d_i}{r^2 T_i}\right)$$

Plugging in the scaling for r_i, d_i and T_i, we obtain that the average overhead rate is given by

$$t_U = \Theta\left(\frac{r_0 d_0}{r^2 T_0} \sum_{i=1}^{K} \alpha^{(1+\mu-\gamma)i}\right)$$

As n gets large, so does the number of rings K, so that the preceding summation converges when $\alpha^{1+\mu-\gamma} < 1$. Since we need $\alpha > 1$ in order to exponentially expand the rings, we must have $1 + \mu - \gamma < 0$ as a necessary condition. As n grows, we have already noted that r_0 scales as $\Theta(r/\beta)$ and T_0 as $\Theta\left(r^2/\sigma^2\right)$. For this scaling, we show in Section 4.5.2 that we need $d_0 = \Theta(r)$ for reliable routing. We now have that

$$T_U = n \times t_U = \Theta\left(\sigma^2/r^2\right)$$

which matches the throughput available per node T_A for a fixed bandwidth.

Lemma 4.4. *The proposed position-publish protocol is* scalable *when* $1 + \mu < \gamma < 2$, $\beta r_0 = \Theta(r)$, $d_0 = \Theta(r)$ *and* $T_0 = \Theta\left(r^2/\sigma^2\right)$.

4.5.2 Routing Reliability

We have analyzed the update protocol to determine conditions for scalability. We now analyze the routing protocol to determine conditions that ensure reliable routing. After an update is made to nodes in a ring, some of these nodes may leave the ring. When a packet being routed to the destination hits the ring, therefore, the relay nodes it sees may be ones which moved in after the currently active update was made. According to the proposed routing protocol, when the packet meets such nodes which have estimates of the destination's location worse than its own (including not having any estimate of the destination's location), it simply continues in the direction it is going. Thus, in order for a packet to take advantage of an active update for ring i once it hits it, it suffices that at least one of the nodes it meets as it is cutting through the ring has an active update corresponding to ring i. If this does not happen, we say that the packet has "missed" the i-th ring. The lifetime T_i of normal updates must be short enough that the probability of a miss tends to zero, which imposes additional conditions on the protocol parameters, as we show here.

The worst case scenario for missing a ring is the following scenario: (i) The packet is relayed radially across it, since it meets fewer relay nodes along the ring, and hence a smaller probability of meeting a node with an active update (ii) The

time since update issue is T_i (just prior to update expiry). This is because the density of relay nodes with updates inside the update ring (the region where the update is spatially valid) decreases as time from update issue increases and the packet is least likely to meet a relay node with an active update just prior to update expiry. We consider this worst case scenario and use the following asymptotics (for outer rings; large i): $r_i/\sqrt{T_i} \to \infty$, $d_i/r_i \to 0$ and $d_i \big/ \left(\sigma\sqrt{T_i}\right) \to 0$ to give an upper bound on the miss probability $P_{\text{miss}}(i)$:

$$\log P_{\text{miss}}(i) \lessapprox - d_i^2 \Big/ \left(\sigma r \sqrt{2\pi T_i}\right) . \tag{4.5}$$

We provide the details of this derivation in Appendix C.2. Since $\sigma\sqrt{T_0} = \Theta(\beta r_0) = \Theta(r)$, if we ensure that d_0 grows at least as fast as r, we can bound the probability of missing the innermost ring ($i = 0$) for all network sizes. Thus, we set $d_0 = \Theta(r)$.

We prove in Appendix C.3 that when no misses occur, the confidence region guarantees of active updates ensure that the uncertainty seen by a packet after bootstrap is bounded by $U_{\max} = \alpha\beta/(1-\beta)$. When we make the following assumptions: (i) β is chosen so that $U_{\max} < 1$ (ii) all nodes broadcast their current positions to other nodes within $2\left(1 - U_{\max}\right)^{-1} r$ (iii) ϵ in $\pi r^2 = (1+\epsilon)\log n$ is a large enough constant given by (4.2) for some $\delta < \min\left\{\pi/3, \arcsin U_{\max}\right\}$, we can invoke the corollary of Theorem 4.1 to guarantee routing reliability. Therefore, the overall probability of routing failure $P_{\text{miss}}^{\text{nett}}$ can be upper bounded by the prob-

ability of missing at least one ring $P_{\text{miss}}^{\text{one}}$ (so that the uncertainty can no longer be bounded by $U_{\max}$). Using the union bound, we have

$$P_{\text{miss}}^{\text{one}} \leq \sum_{i=0}^{i=K} P_{\text{miss}}(i) \lessapprox \sum_{i=0}^{i=K} \exp\left(-d_i^2 \Big/ \left(\sigma r \sqrt{2\pi T_i}\right)\right). \tag{4.6}$$

Since $d_0^2/\left(r\sigma\sqrt{T_0}\right) = \Theta(1)$, we note that $P_{\text{miss}}^{\text{one}}$ is bounded for all network sizes (and thus, so is $P_{\text{miss}}^{\text{nett}}$) if $\gamma < 4\mu$. By choosing $d_0^2/\left(r\sigma\sqrt{T_0}\right)$ to be large enough, we can drive the bound on $P_{\text{miss}}^{\text{one}}$ given by (4.6) below any desired level.

Lemma 4.5. *The proposed routing protocol is* reliable *when* $\gamma < 4\mu$, $0 < \beta < 1/(1+\alpha)$, $d_0 = \Theta(r)$, $T_0 = \Theta\left(r^2/\sigma^2\right)$ *and* ϵ *in* (4.1) *is large enough so that* (4.2) *holds for some* $\delta < min\{\pi/3, \pi/2 - \arcsin U_{max}\}$.

4.5.3 Routing Efficiency

In order to bound the route stretch, we must account for the fact that, since location updates are sent to only a subset of nodes, the source node need not have an active update for the destination. In this case, the packet travels an additional distance in an arbitrarily chosen direction until it hits a node with an active update and we refer to this process of acquiring an initial estimate as the bootstrapping process. Our bound on route stretch must account for this additional distance. Once the packet does encounter a node with an active update, we use uncertainty, which we show is bounded by $U_{\max} = \alpha\beta/(1-\beta)$ provided the packet does not

miss smaller indexed rings thereafter, to bound the route stretch. The resulting bound on the route stretch is given by (we assume that ϵ in the choice of r is large enough so that the direction along which a packet is forwarded closely matches that corresponding to the position update used; i.e., ϵ chosen to satisfy (4.2) for small δ):

$$\sqrt{1+\left(\sqrt{\frac{\alpha^2\left(1+\beta\right)^2}{\left(1-\beta\right)^2}-1}+\frac{\alpha\left(1+\beta\right)}{\sqrt{\left(1-\beta\right)^2-\alpha^2\beta^2}}\right)^2} \tag{4.7}$$

Details of the derivation can be found in Appendix C.3. We note that this bound is finite if $U_{\max}=\alpha\beta/(1-\beta)<1$.

Lemma 4.6. *Packet trajectories are within a constant* stretch *factor of the source-destination distance (*efficient*) when routing is* reliable *(no additional constraints).*

4.6 Simulation results

We perform simulations of the position-publish and routing protocol for a particular destination node, for the following scenario: Number of nodes $n=1.8\times10^6$, node density 1, mobility model being 2D Brownian motion with parameter $\sigma^2=1$ and the deployment area is a square of side $\sqrt{n}$. We choose the communication radius r to be $\sqrt{(1+\epsilon)\log n/\pi}$ and report results for both $\epsilon=0$ and $\epsilon=2$. The parameters of the position-publish protocol are: (i) confidence region parameter

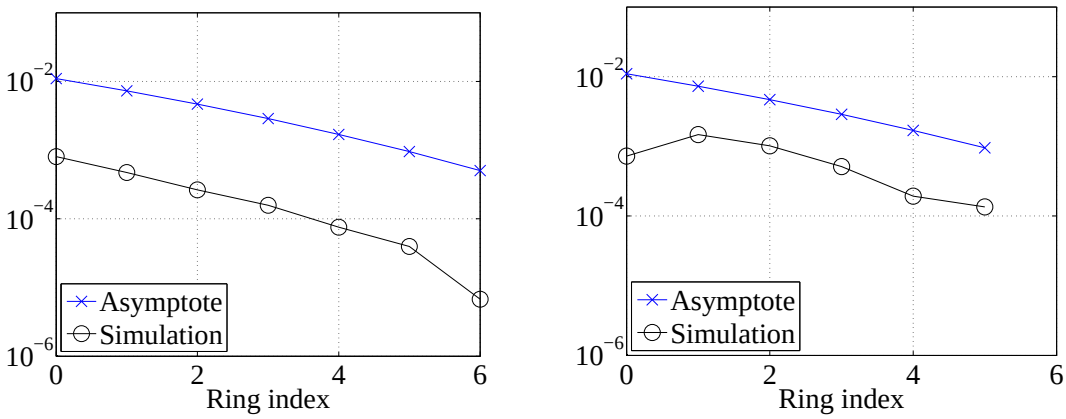

Figure 4.5: Probability of missing a ring for r corresponding to $\epsilon = 0$ (left) and $\epsilon = 2$ (right) for radial traversal just prior to update expiry (worst case)

$\beta = 0.25$, (ii) order-zero ring specified by $r_0 = r/\beta$, $d_0 = 2r$, $T_0 = (1/8)\,(\beta r_0/\sigma)^2$ and (iii) ring scaling parameters $\alpha = 2$, $\mu = 0.55$, $\gamma = 1.95$.

To get a concrete sense of what these numbers mean, we choose the units of distance so that the communication radius is $r = 100$m (for $\epsilon = 0$). The deployment area is 63 km by 63 km, with a node density of 458 nodes per square km. Now we choose units of time so that RMS motion over one second $\sigma\sqrt{2}$, to be 10m (consistent with vehicular speeds), we see that the lifetime of updates to the update rings of radii $3.2, 6.4, 12.8$ & 25.6 km ($i = 3$ to $i = 6$) and thickness $0.63, 0.92, 1.35$ & 1.97 km are $0.40, 1.55, 5.98$ & 23.10 hours respectively.

Probability of a packet missing updated nodes in a ring: We first compare the asymptotic estimate of the worst case miss probability in (4.5) with simulations of the worst case (radial trajectory & just prior to update expiry) in Figure 4.5.

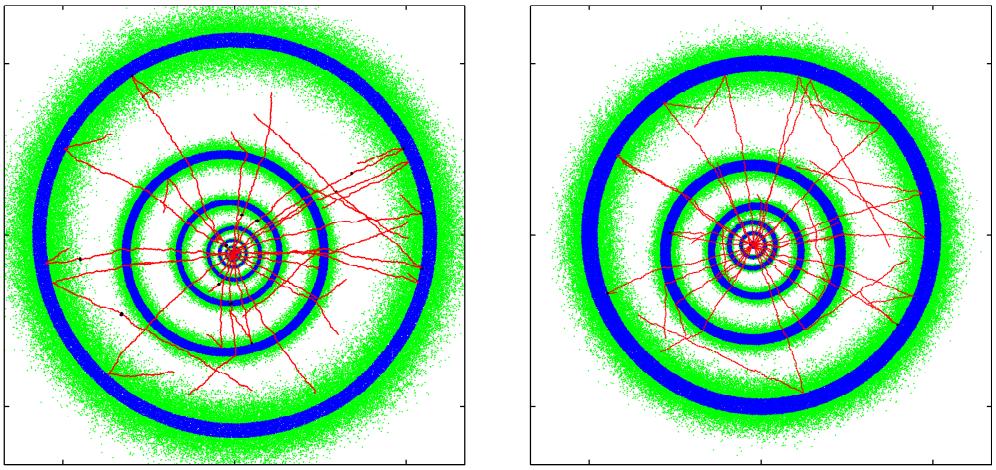

Figure 4.6: Overlaid routing trajectories (converging to a destination node) at a snapshot of the network for the proposed routing protocol for two different communication radii (smaller on the left). Blue dots indicate nodes with active position updates for the destination node; green dots indicate nodes with active position updates that however cannot be used because these nodes do not satisfy the spatial validity constraint of their position updates; red dots indicate relay nodes along packet trajectories using greedy geographic forwarding; black dots indicate greedy face traversal around voids [34].

Note that the match is better for $\epsilon = 2$, because the average number of relay nodes seen by a packet decreases with an increase in ϵ and approaches the lower bound of d_i/r used in the derivation of the upper bound on $P_{\text{miss}}(i)$ ((C.3) in Appendix C.2).

Trajectories: We plot a sampling of the simulated trajectories in Figure 4.6 (Left: $\epsilon = 0$, Right: $\epsilon = 2$) (we ignore edge effects by focusing only on trajectories that start inside the outermost update ring while noting that network boundaries can be handled either by using a specialized update ring, or via packet "reflec-

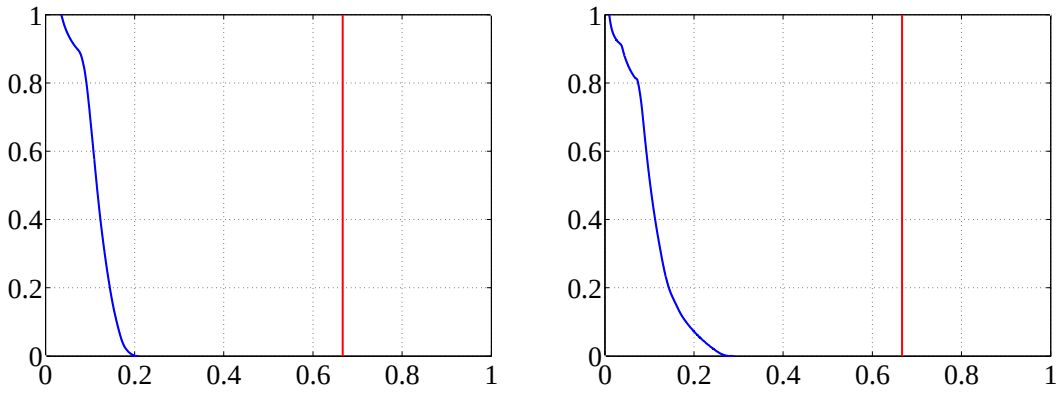

Figure 4.7: CCDF of uncertainty seen by packets as they cut through the network (after bootstrap) for $\epsilon = 0$ (left) and $\epsilon = 2$ (right). The red vertical line demarcates the upper bound on uncertainty (4.4)

tion" at the boundaries). We note that, for larger communication radius ($\epsilon = 2$), the trajectories are straighter, as there are many nodes available in each direction around a relay node. For $\epsilon = 0$, which is below the threshold [62] for asymptotic success of greedy geographic forwarding, trajectories hit voids frequently. However, using the standard technique of greedy left hand traversal of voids [34] (these segments of the trajectory are marked in black), route failure rates are reduced to a small level.

Uncertainty: The uncertainty of the position estimate is designed to be less than $\alpha\beta/(1-\beta) = 2/3$ once the packet bootstraps. From our simulations, we find that the uncertainty seen by a packet after bootstrap remains smaller than this value. The CCDFs of the location uncertainty after bootstrap are presented in Figure 4.7.

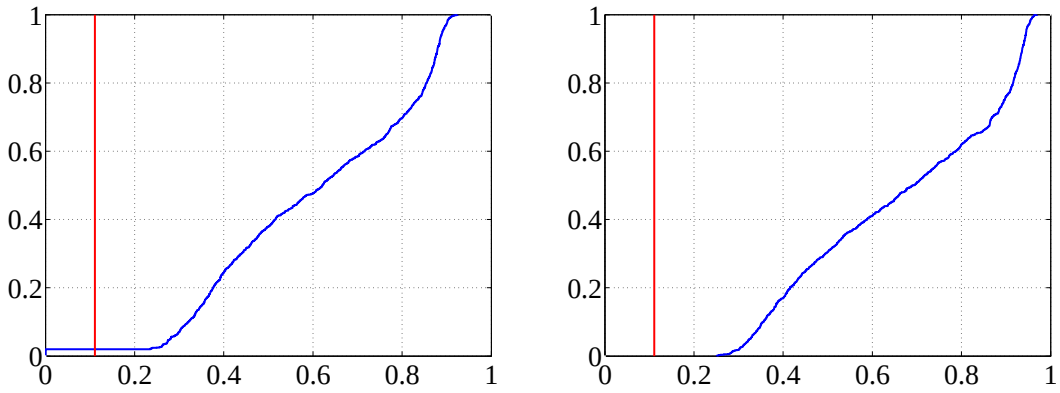

Figure 4.8: CDF of reciprocal stretch, i.e., the ratio of source-destination distance to the length of the packet trajectory, for communication radius r corresponding to $\epsilon = 0$ (left) and $\epsilon = 2$ (right). Red vertical line corresponds to the worst case stretch (4.7)

Route Stretch: Figure 4.8 plots the CDF of the *reciprocal* of the route stretch attained, with reciprocal stretch equal to zero indicating a routing failure (edge effects are ignored by focusing on trajectories which start inside the outermost update ring). From (4.7), the route stretch is bounded by 9 and the corresponding reciprocal stretch is marked via the red vertical line. Note that all successful routes satisfy this guarantee, and that all routes are successful for $\epsilon = 2$. For $\epsilon = 0$ (greedy geographic forwarding not guaranteed to work), a small fraction of route failures do occur.

Chapter 5

Inference from time on Twitter

In this chapter, we first explore whether the *timing* of a user's tweets tells us something about her/his interests, by comparing it against the known timing of external events associated with a particular interest. We take the example of baseball fandom to illustrate that the time of tweets can be a useful tool for interest inference on Twitter.

We then tackle the problem of understanding the temporal evolution of a topic feed on Twitter with a view to leverage this knowledge for identifying user interests. We use minimal measurements derived from the tweets in the feed: their time and authorship (the *when* and the *who* of each tweet). The main features of this problem are the following:

(a) Twitter feeds exhibit significant time-of-day effects. i.e., even when no interesting activity takes place, the variations in tweet intensity as a function of time are pronounced

(b) The duration of interesting activity (and their tweet intensity) observed is diverse, with some events lasting around 20 hours, while others are brief (10 minutes)

(c) Within the broad area, we expect people who share specific interests to tweet at roughly the same times, when conversations related to their shared interest take place on Twitter.

The framework proposed in this chapter accommodates the above features and has two major components: (i) Identify and isolate the times when interesting activity takes place in the feed using tweet times. We call such times *event times*. We address the facets of the problem (a) and (b) when we detect events (ii) Dissect, using tweet times and authorship, this set of events into subsets, each of which likely correspond to different subtopics. We exploit attribute (c) in order to discover such relationships.

Outline: We start by reviewing prior work on mining Twitter feeds in Section 5.1. We state the model we use for tweet times of a user in Section 5.2.1. We model a user interested in the topic as one who is likely to tweet at a higher rate during "event times" associated with the topic than at other times. This leads to a statistical measure for a user's interest which is the Bayesian posterior probability, based on measured tweet times, of the user's tweet rate during event times be-

ing higher than at other times. Under our model, this probability only depends on the *number* of tweets (rather than on their exact timing) during events and during other times. This makes the statistic, which we derive in Section 5.2.2, attractive for inference in large-scale systems, both in terms of measurement and computation. We present our model for interest-overlap among "neighboring" users and derive a statistic for pooling interest-level estimates of neighbors under this model in Section 5.2.3. In Section 5.2.4, we present results for inferring baseball fandom. We define event times as those times which see abnormally high activity. In order to identify these event times, we need to define what normal is: i.e., model baseline activity. By using a time-varying Poisson process for modeling baseline activity, we are able to capture the significant time-of-day effects in aggregate Twitter feeds. This model is presented in Section 5.3.1. We review constant false alarm rate (CFAR) detection under the Poisson model in Section 5.3.2 and present our event detection algorithm in Sections 5.3.3 and 5.3.4. We follow this up with the baseline estimation algorithm in Section 5.3.5. We round out Section 5.3 by presenting event detection results corresponding to 4 TV show feeds in Section 5.3.6. In Section 5.4, we identify relationships between events using tweet authorship information (user ID) and dissect the event times identified in Section 5.3 into subtopics. For this purpose, we build a user-event bipartite graph, where we place an edge between a user and an event if the user tweets

during the event. We embed the random-walk commute times between events on this bipartite graph using spectral methods. We summarize these ideas in Sections 5.4.1 and 5.4.2. We explain our event clustering algorithm which makes use of these embeddings in Section 5.4.3 and present implementation considerations in Section 5.4.5. We conclude Section 5.4 by presenting wordclouds and timelines of significant subtopics (for the 4 TV show feeds) in Section 5.4.6. We choose to illustrate the power of our inference framework via TV shows (we present results for four shows) because clearly demarcated periodic events such as episode air times provide unambiguous validation of our approach, while less predictable tie-ins with other cultural and commercial phenomena provide an opportunity to demonstrate that our approach can lead to automated discovery of interesting phenomena (all without text analysis).

5.1 Related Work

Prior work on mining Twitter feeds has mostly been fed by text analysis. TwitterStand [51] maintains a news stand by parsing through different tweet feeds. The timing of tweets has been used here to help in the clustering of tweets into different news groups. The authors in [50] build a system that can locate events such as an earthquake in space and time from tweets (using tweet location and

times). However, unlike the solution proposed herein, both[51, 50] rely mainly on text analysis, with tweet times being used only in the later stages. PET[36] tracks the evolution of events, and users' interest in them, as a function of time. Unlike our approach, PET uses text analysis, and does not use the specific tweet time or its relation to external events (PET analyzes tweets collected daily to infer the evolution of topics from day to day). A method of training a classifier to do sentiment analysis of individual tweets is proposed in [44]. Here smileys are used in a bootstrapping mechanism to build a corpus of words along with an associated sentiment (positive or negative) for each word. The preceding references do not explicitly aim to mine for the interests of *a user*, which is the focus of our work. A system that employs Wikipedia as an external corpus to do word associations is proposed in [40] for mining *broad* interests on a *per user* basis. In [16], the authors observe that in identifying political affiliation of a 1000 hand-labeled users, the structure of the re-tweet graph is more useful than the text in the tweets themselves. They arrive at this conclusion by implementing a text based classifier and comparing it with the results obtained merely by identifying the community structure in the re-tweet graph.

There is by now a significant body of work on Twitter event detection for summarizing topic-specific Twitter feeds. The method proposed in [54, 55] relies on abrupt local *changes* in tweet counts to detect activity in a topic-specific feed.

TwitInfo [39] identifies events using an algorithm inspired by TCP's congestion control algorithm, by keeping track of the mean and variance of tweet counts. The authors in [43] present a system to detect and summarize significant events in a single sport event. The event detection algorithm proposed maintains an estimate of the mean and variance of the *slopes* of tweet counts (differences in adjacent tweet counts). These estimates are used to threshold observed slopes in order to detect events. The preceding systems examine the content of tweets that fall within events, so as to provide text-based summaries of *individual* events corresponding to a single extended event such as a soccer game or Presidential debates. In contrast to these local change detection approaches, we use observations over a longer period to build a baseline statistical model, and then use this to identify extended events of interest as well as shorter burst of activity. In addition, we provide an approach for teasing out relationships *between* identified events without any content analysis.

Temporal dynamics have been used for clustering news topics in Twitter based news aggregator systems[51, 64]. Identifying the newsworthiness of a tweet is a fundamental feature of such systems. As a result they rely heavily on the text of tweets. In fact TwitterStand[51] uses the distribution of words in a tweet first to filter out "noisy tweets," using the time of tweets later as one of the cues when clustering newsworthy tweets into topics. EDCoW [64] uses the distribution of

words in each tweet to detect important words, and employs a short-term temporal analysis of these words (over a 24 hour period) to extract trending themes (a bag of words) from a Twitter feed obtained by crawling around 20,000 "active" users. The correlation between the time courses of usage for different words is used in [64] for clustering newsworthy words. While this is similar in spirit to our approach, we cluster events based on *users* with similar tweet times, and hence do not require any content analysis.

A complementary work of interest is the study of temporal patterns of user attention in online social media in [65], which shows that *short-term* temporal dynamics (over a few days) of a large number of topics (hashtags) on Twitter can be well-approximated by a small set of shapes. While our focus here is on detecting and dissecting activity over *large* time duration (order of a year) for a *single topic,* exploring whether the shapes of the events we detect are similar to those in [65] is certainly an interesting topic for future work.

5.2 Inferring interest from time

5.2.1 Tweet times model

We start by presenting a probabilistic model for tweet times of a user over an observation time window (this need not consist of contiguous intervals) . Our basic

premise is the following: a user who is interested in topic X (say the SFGiants baseball team) tweets more often at times when X is in the "public eye" (SFGiants play a baseball game) than at other times. Thus, we partition the observation window into two complementary sets:

1. **Event times** are times within the observation window when X is in the "public eye" (which, according to our hypothesis, stimulates users interested in X to engage in conversations on Twitter).

2. **Non-event times:** All other times over the observation window.

This partitioning, along with the behaviors we expect for users who are interested (or not) in the topic X, is shown in figure 5.1.

The tweet times of a user are modeled as a homogeneous Poisson process of rate λ_1 tweets per unit time during event times and an *independent* homogeneous Poisson process of rate λ_0 tweets per unit time during non-event times. As depicted in figure 5.1, we expect that $\lambda_1 > \lambda_0$ for users interested in topic X.

A homogeneous Poisson process is parameterized by a single parameter, its rate λ. Such a parsimonious model for the tweet times of a user has two advantages: robustness (heterogeneity among twitter users may make more detailed usage profiles, such as allowing for tweet rates dependent on the time of day, counterproductive) and simplicity (e.g., the decision statistics we obtain require

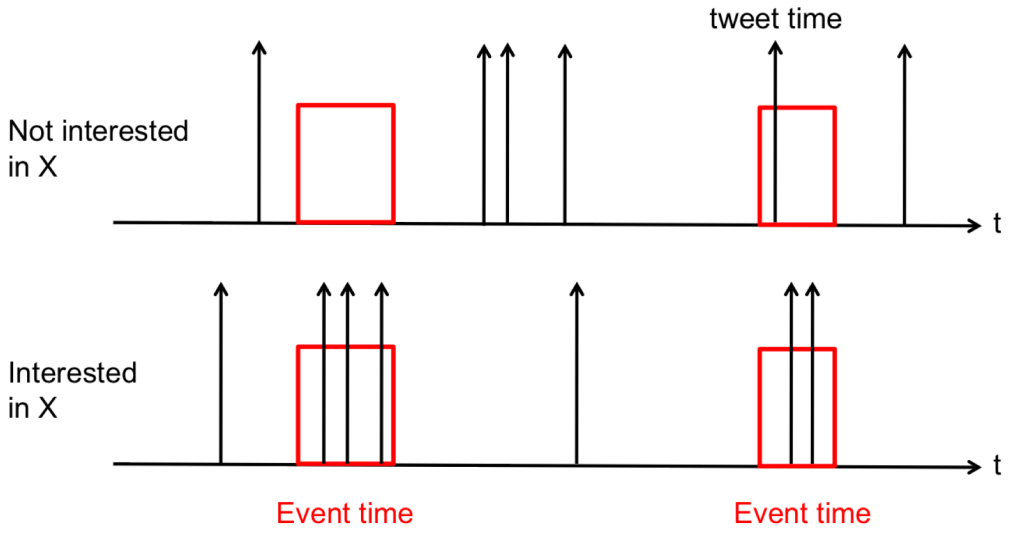

Figure 5.1: Tweet times of the user marked by arrows. Event times are marked in red. All other times are non-event times. Top: Tweeting behavior of a person not interested in X. Bottom: A person interested in X

aggregate tweet counts rather than individual tweet times). For a Poisson process of constant rate λ tweets/unit time, the number of tweets N made in a time interval of length T (need not be contiguous) is a Poisson random variable with mean $\lambda \times T$. i.e., the probability that the user puts out n tweets in T time units is given by

$$\Pr\left[N = n | \, \lambda\right] = \frac{e^{-\lambda T} \left(\lambda T\right)^n}{n!}, \;\; n = 0, 1, 2, \ldots, \infty.$$

Further, under the Poisson model, the number of tweets put out by the user in non-overlapping time intervals are independent random variables.

5.2.2 Inferring interest levels from tweet times

We propose a statistic that measures our confidence in the assertion that the user tweets more *frequently* during event times than other times. i.e., his/her tweet rate during event times is larger than the rate at other times. This statistic is our metric for the user's interest level in the topic X. We use knowledge of the event and non-event times to estimate the probability distributions of the corresponding tweet rates λ_1 and λ_0 from the tweet times of the user, and then compute the statistic from these posterior distributions.

Under our Poisson model, the posterior distribution of λ_1 given the tweet times depends only on the total number of tweets put out by the user during event times, which we denote by N_1. The tweet times themselves do not matter. Similarly, to make probabilistic inferences on λ_0, all we need is the total number of tweets during non-event times, denoted by N_0. In the language of estimation theory, N_1 and N_0 are *minimal sufficient statistics* for estimation of λ_1 and λ_0, respectively. Let the total time span of the event times and non-event times be T_1 and T_0 respectively.

Continuing with our minimalism in modeling, we assume a *non-informative* prior on the rates λ_1 and λ_0, assuming that the prior density $p(\lambda_1, \lambda_0) \propto 1/\sqrt{\lambda_1 \lambda_0}$ for all $\lambda_1 > 0, \lambda_0 > 0$ (the corresponding marginal priors are $p(\lambda_i) \propto 1/\sqrt{\lambda_i}$ for

$\lambda_i > 0$, $i = 0, 1$). Of course, this prior cannot exist over an infinite support, since densities must integrate to one, but this is a standard trick in Bayesian estimation when the ground truth on priors is difficult to determine. This joint prior is the Jeffreys non-informative prior on the rate parameters (λ_1, λ_0) [32]. In our case, accurately estimating priors for each topic X would require the ground truth on the interests of a large number of users, which goes counter to our objective of mining for these interests. Furthermore, we would need to constantly revise our ground truth data set for a heterogeneous population of Twitter users with dynamically evolving interests, which is clearly infeasible.

Since we assume that the two Poisson processes corresponding to the event times and non-event times are independent, the corresponding counts N_1 and N_0 are conditionally independent given λ_1, λ_0. Putting this together with our assumption of non-informative prior on λ_1, λ_0, we obtain, using Bayes' rule, that the posterior distributions of λ_1, λ_0 also factor and are given by $p(\lambda_i|N_i) \propto \Pr[N_i|\lambda_i]\, p(\lambda_i) \propto {\lambda_i}^{N_i - 0.5} e^{-T_i \lambda_i}$, $\lambda_i > 0$. Normalizing the posteriors so they integrate to one (which we can do even though we employed improper priors), we obtain

$$p(\lambda_i = x|N_i) = \begin{cases} \frac{T_i (T_i x)^{N_i - 0.5} e^{-T_i x}}{\gamma(N_i + 0.5)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{5.1}$$

where $\gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the gamma function. The statistic which we propose to quantify the user's interest level in the topic X is $Z = \Pr[\lambda_1 > \lambda_0 | N_1, N_0]$. We declare a user to be interested in X when Z exceeds a certain threshold. Thus, we conclude that the user is interested in X, when we are "confident enough" that his/her tweet rate during event times is larger than that during non-event times. Given the observations N_1, N_0, T_1, T_0, the statistic Z can be computed using the posteriors (5.1) as follows:

$$\begin{aligned} Z &= \Pr[\lambda_1 > \lambda_0 | N_1, N_0] \\ &= \iint_{x>y} p(\lambda_1 = x | N_1)\, p(\lambda_0 = y | N_0)\ dx\ dy. \end{aligned} \tag{5.2}$$

5.2.3 Exploiting user interactions

We use social networks to engage in conversation with others who share our interests. If a user has interacts with others who are interested in the topic X, we expect that the probability that he/she is also interested in X is higher than for a randomly picked user. We present a method that relies on this simple intuition to improve our estimates of the interest level of a "tagged" user using interest level estimates of other users mentioned in his/her tweets. During the observation time window, this tagged user may mention other users using their twitter handle (for example, the official SFGiants twitter handle `@SFGiants`, or another individual

`@johnadams2001`) in his/her tweets. We call such users the "neighbors" of the tagged user. Since we will be combining the Z statistics of multiple users, we need to pay attention to scaling. In particular, we expect that we would weight the tagged user's Z statistic higher than that of his/her neighbors. We now describe a framework for motivating such scaling.

Notation: Let the index 0 denote the tagged user and the indices $i = 1, \ldots, M$ denote the neighbors. From the number of tweets during event times $N_1(i)$ and non-event times $N_0(i)$ of the i-th user ($\mathbf{N}(i)$ denotes the pair $(N_1(i), N_0(i))$), we arrive the statistic (5.2) which we denote by Z_i. Let $\lambda_1(i), \lambda_0(i)$ denote the tweet rates of the i-th user in the event and non-event times and Y_i denote the event that the i-th user tweets more frequently during event times than other times. i.e., $Y_i = 1$ if $\lambda_1(i) > \lambda_0(i)$ and $Y_i = 0$ otherwise (note that $Z_i = \Pr[Y_i = 1|\mathbf{N}(i)]$). Let C_i represent the true interest of i in the topic X (C_i takes the value 1 if this user is interested in X and 0 otherwise).

A user who is not interested in X may still happen to tweet more often during event times. Likewise, a user interested in X may happen to tweet less frequently during event times than at other times. Therefore, we first relate Y_i to C_i in a probabilistic manner to derive a function of the Z_i statistic for each user i (i.e., the tagged user and his/her neighbors) such that, when combined across users to make an inference regarding the tagged user, no one user has too big an influence.

We then discuss a model for the dependence between the tagged users and his/her neighbors which motivates combining these individual statistics.

For the first step, let $p_t = \Pr[Y_k = 1|C_k = 1]$ denote the probability that a user interested in topic X is *timely* (i.e., tweets more frequently during event times than at other times), and let $p_f = \Pr[Y_k = 1|C_k = 0]$ denote the probability of *false alarm* (i.e., a user not interested in X happens to tweet more frequently during event times). We now compute the likelihood ratio of user k's interest in topic X based on its own measurements, defined as

$$\phi_k = \frac{\Pr[\mathbf{N}(k)|C_k = 1]}{\Pr[\mathbf{N}(k)|C_k = 0]},$$

in terms of the statistic $Z_k = P[Y_k = 1|\mathbf{N}(k)]$, which we already know how to compute from Section 5.2.2.

Under our uninformative prior, it is easy to show that $\Pr[Y_k = 1] = \Pr[\lambda_1 > \lambda_0] = \frac{1}{2}$, Conditioning on Y_k and using the conditional independence of $\mathbf{N}(k)$ and C_k given Y_k (the Markov structure in figure 5.2):

$$\begin{aligned}\Pr\left[\mathbf{N}(k)|C_k\right] &= \Pr\left[Y_k = 1|C_k\right]\Pr\left[\mathbf{N}(k)|Y_k = 1\right] + \Pr\left[Y_k = 0|C_k\right]\Pr\left[\mathbf{N}(k)|Y_k = 0\right] \\ &= 2\Pr[\mathbf{N}(k)]\Big(\Pr\left[Y_k = 1|C_k\right]Z_k + \Pr\left[Y_k = 0|C_k\right](1 - Z_k)\Big),\end{aligned}$$

where we have used $Z_k = \Pr\left[Y_k = 1|\mathbf{N}(k)\right]$. Using the above, we obtain that

$$\phi_k = \frac{p_t Z_k + (1 - p_t)(1 - Z_k)}{p_f Z_k + (1 - p_f)(1 - Z_k)} = \frac{1 + p_t\left(\frac{Z_k}{1-Z_k} - 1\right)}{1 + p_f\left(\frac{Z_k}{1-Z_k} - 1\right)}.$$

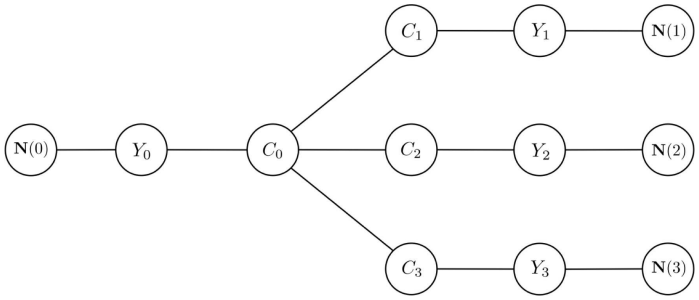

Figure 5.2: Markov structure of the user interests C_i, the tweet rate differentials $Y_i = \lambda_1(i) > \lambda_0(i)$ and the number of tweets $\mathbf{N}(i) = (N_1(i), N_0(i))$. The index 0 refers to the tagged user while $1, 2, 3$ denote the neighbors of this user

This effectively corresponds to soft thresholding the raw likelihood ratio $\Pr[\lambda_1 > \lambda_0]/\Pr[\lambda_1 \leq \lambda_0] = Z_k/(1-Z_k)$ between an upper limit of $1/p_f$ and a lower limit of $1-p_t$. Both ϕ_k and the raw likelihood ratio are monotone increasing in Z_k. Thus, for a single user (as considered in the Section 5.2.2), threshold rules based on any of these statistics are equivalent. However, when combining across multiple users, the soft thresholding in ϕ_k is important for robustness, since it ensures that no one user has too large an influence on the outcome.

Let us now consider the second step: relating the interests of the tagged user and his/her neighbors. We expect that it is more likely that the neighbors are interested in X when the tagged user is interested in X than when the tagged user is not: Denoting $\Pr[C_k = 1|C_0 = 1]$ by α and $\Pr[C_k = 1|C_0 = 0]$ by β, we expect that $\alpha \gg \beta$. It is actually the difference in α and β that affects how we combine these statistics, rather than their raw values. For example, even if α is small (e.g.,

0.1, so that there is only a 10% probability of the neighbor of a fan also being a fan), if $\beta = 10^{-4}$, then we still get very useful information from the neighbors' measurements.

We make a simplifying assumption on the structure of interactions among neighbors: The true interests of the neighbors, $\{C_i, i > 0\}$, are independent when conditioned on the interest status of the tagged user C_0 : $\Pr[C_1, \ldots, C_M|C_0] = \prod \Pr[C_i|C_0]$. This is illustrated via the Markov structure depicted in figure 5.2 (in the figure $M = 3$). This assumption is violated when a neighbor of the tagged user refers to another neighbor of the tagged user in his/her tweets (therefore introducing additional dependencies between the two neighbors). However, as we will see when we discuss the results, this simple structure by itself gives us considerable gains over just using the interest level estimates Z_0 of the tagged user alone.

Our statistic that incorporates information from the neighbors is the following log likelihood ratio:

$$S = \frac{\Pr\left[\mathbf{N}(0), \mathbf{N}(1), \ldots, \mathbf{N}(M)|C_0 = 1\right]}{\Pr\left[\mathbf{N}(0), \mathbf{N}(1), \ldots, \mathbf{N}(M)|C_0 = 0\right]}.$$

From the Markov structure in figure 5.2, we observe that the true interests of the neighbors C_i given that of the tagged user C_0 are independent. This observation

leads to the following simplification:

$$S = \log \frac{\Pr[\mathbf{N}(0)|C_0 = 1]}{\Pr[\mathbf{N}(0)|C_0 = 0]} + \sum_{k=1}^{M} \log \frac{\Pr[\mathbf{N}(k)|C_0 = 1]}{\Pr[\mathbf{N}(k)|C_0 = 0]}.$$

From Bayes' rule, for the neighbors,

$$\begin{aligned}\Pr\left[\mathbf{N}(k)|C_0\right] &= \Pr\left[\mathbf{N}(k), C_k = 1|C_0\right] + \Pr\left[\mathbf{N}(k), C_k = 0|C_0\right] \\ &= \Pr\left[\mathbf{N}(k)|C_k = 1\right]\Pr\left[C_k = 1|C_0\right] + \Pr\left[\mathbf{N}(k)|C_k = 0\right]\Pr\left[C_k = 0|C_0\right].\end{aligned}$$

Using the above, we obtain that:

$$\frac{\Pr[\mathbf{N}(k)|C_0 = 1]}{\Pr[\mathbf{N}(k)|C_0 = 0]} = \frac{\alpha\phi_k + (1-\alpha)}{\beta\phi_k + (1-\beta)}.$$

Therefore, the statistic S depends only on the likelihood ratios ϕ_k of the tagged user and his/her neighbors, as follows:

$$S = \log \phi_0 + \sum_{k=1}^{M} \log \frac{1 + \alpha(\phi_k - 1)}{1 + \beta(\phi_k - 1)}.$$

While we can tune the parameters α and β to get good performance with this statistic, in practice, we have found the following modified rule, using a single parameter to scale down the sum of the neighbors' log likelihood ratios, to work well:

$$\tilde{S} = \log \phi_0 + \kappa \textstyle\sum_{k=1}^{M} \log \phi_k. \tag{5.3}$$

In our numerical results, therefore, we report on the performance of this modified statistic, with $\kappa = 1/6$ (found to work well empirically).

5.2.4 Results

We test our statistical framework by trying to identify whether a user is a fan of the San Francisco Giants (SFGiants) baseball team from the user's tweet times (we also briefly report on analogous results for the NY Yankees). The times when SFGiants played Major League Baseball (MLB) games are used as a natural candidate for event times. We also include a 15 minute window on either side of each game in our definition of event times to account for the buzz before and after each game when fans are expected to tweet heavily.

Dataset description: The data set is a 10% random sampling of all *public* tweets over a month (May-June) in the summer of 2011. In this one month window, SFGiants played 29 games. Each tweet, apart from its brief text, is tagged with an user ID, the time when this tweet was made and the user IDs of twitter handles mentioned in the tweet (if any).

Ground truth: In order to characterize the effectiveness of the statistic that we propose, we need to know the fandom of users on whose tweet times we apply the statistic. For this purpose, we searched the text of all tweets (in our dataset) that were made in the first and last 10 minutes of all SFGiants games for keywords associated with this baseball team. The keywords that we used were: `sfgiants`, `#sfgiants`, `rowand`, `#rowand`, `lincecum` and `#lincecum`. We identified 640 users

in this manner. We assume that these users who used the keywords associated with the SFGiants baseball team are indeed their fans. We also picked a random set of 1000 users who appear in our dataset (they tweeted at least once in this one month window). None of these randomly picked users used the preceding keywords in their tweets and we assume that they are not fans of SFGiants.

For all of the above users (fans and non-fans) we keep a list of the times at which they put out tweets in this one month window. We use these times to evaluate the statistic (5.2) for these users. We also keep a list of user IDs for each of these users and this list gives our per user neighbor list. The entries in this list are the users who are mentioned in the tweets of the tagged user over the one month time window (his/her neighbors). In order to compute the statistic (5.3) which uses estimates of the interest levels of the neighbors, we also compile a list of the tweet times of the neighbors of every user.

Interests from user times: We evaluate the statistic Z in (5.2) from the tweet times of the 640 fans and 1000 non-fans. When computing Z, we account for an average of ten hours of sleep daily. We do this by scaling the total one month time window $T_1 + T_0$ by 14/24 and computing the total sleep compensated non-event times via $T_0' = (14/24) \times (T_1 + T_0) - T_1$. We assume that the user is awake during event times (thus leaving T_1 as it is). Let $\hat{\lambda}_i = N_i/T_i$ denote the empirical estimate of λ_i. We threshold the Z statistic at different values and plot

the number of correctly detected fans versus the false alarms (number of randomly picked users misclassified as fans) in figure 5.3 (blue curve, top). Contrast this with naive ratio of empirical tweet rate estimates $\hat{\lambda}_1/\hat{\lambda}_0 = (N_1T_0)/(N_0T_1)$ that is plotted in black. When we are interested in small false alarm rates, $\hat{\lambda}_1/\hat{\lambda}_0$ metric is not useful: for a false alarm rate of 10/1000 we detect a mere 51/640 fans when we use $\hat{\lambda}_1/\hat{\lambda}_0$, whereas, we are able to detect 137/640 fans using the statistic Z. However, when we are willing to tolerate more false alarms ($> 40/1000$), we see that the performance of Z is comparable to that of empirical tweet rate ratios $\hat{\lambda}_1/\hat{\lambda}_0$.

Incorporating neighbor tweet times: From the tweet times of the neighbors of the 640 fans and the 1000 randomly picked users, we compute their interest level statistic Z (again accounting for a per day average of ten hours of sleep). We then use the interest level estimates of the tagged user and his/her neighbors to compute the statistic $\tilde{S}$ in (5.3). To compute ϕ_k from the individual interest levels Z_k, we choose $p_t = 0.9$ and $p_f = 10^{-20}$. We threshold the statistic $\tilde{S}$ at different values and plot as before the number of correctly detected fans versus the false alarms in figure 5.3 (red curve, top). From the figure, we see that for any fixed false alarm rate, we are able to detect more fans via the consolidated statistic $\tilde{S}$ than the interest level Z_0 of the tagged user alone. For example, for a false alarm rate of 10 in a 1000, we are able to improve the detection accuracy for SFGiants

from 138/640 using Z_0 alone, to 233/640 using the consolidated statistic $\tilde{S}$ with $\kappa = 1/6$.

We run an identical analysis for 623 fans of the New York Yankees baseball team (identified in a manner similar to the SFGiants fans). These results are plotted in figure 5.3 (bottom). We see the same trend with the Yankees, with $\tilde{S}$ outperforming Z.

When interpreting the results summarized in figure 5.3, we must bear in mind the importance of operating at low false alarm rates. The proportion of "fans,", or users interested in any particular topic, is expected to be small. For example, suppose 10% of the overall user population are fans. Then, for a moderately large false alarm rates of 10%, the number of misclassified non-fans is 9% of the user pool. This can overwhelm the pool of correctly classified fans, which is at most 10% for our example. This is the well known *multiple comparisons* problem, for which the natural regime of interest is low false alarm rates. From Figure 5.3, we see that we are able to detect a significant fraction of fans for false alarm rates as small as 1%.

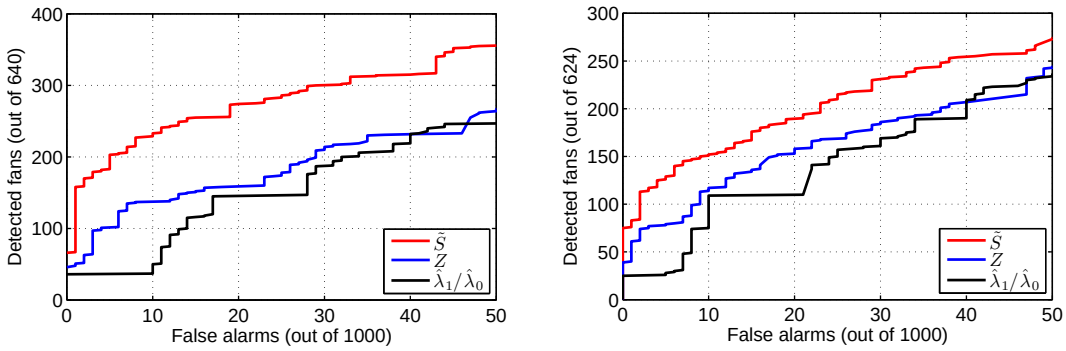

Figure 5.3: Number of correctly detected fans plotted versus the number of randomly picked users misclassified as fans for the statistics $Z, \tilde{S}$ and $\hat{\lambda}_1/\hat{\lambda}_0$. Left: SFGiants and Right: Yankees

5.3 Topic-specific trending times

In order to identify the times when a "unusually high" level of activity related to a topic takes place on Twitter, we must first build a model for baseline/default tweet rates. In this section, we describe how we build our baseline model, introduce the notion of *virtual time* tailored to our goal of detecting events corresponding to increased activity, and describe our multi-resolution approach to detecting these events.

5.3.1 Poisson model

Our baseline model for "background chatter" related to a keyword or set of keywords is a Poisson Point Process (PPP) whose rate is a function of the time-

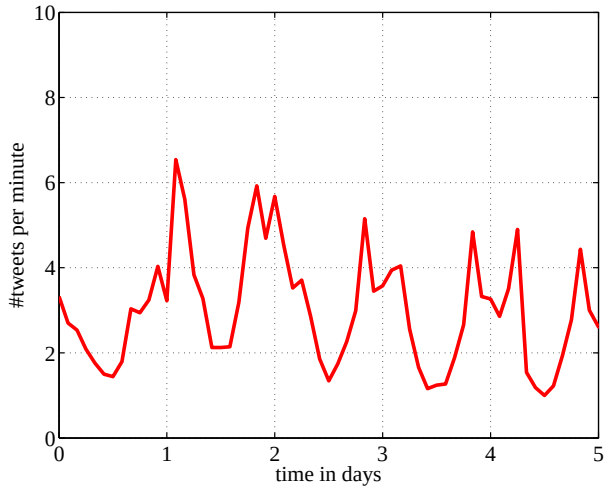

Figure 5.4: Rate of tweets that match the query `simpsons` or `bartsimpson` over a 5 day period measured using 2 hour windows

of-day. Figure 5.4 shows an example of the strongly periodic trend (with a period of 24 hours) in topic-specific tweet rates. The PPP model is consistent with the following intuitively plausible assumptions for baseline activity: (i) a user's past tweet times (matching the keyword query) are independent of future tweet times; (ii) a user's tweet times do not depend on the tweet times of any other user; (iii) the pool of users who can potentially use one of the keywords we track is large; (iv) the chance that a particular user posts a tweet with a keyword at any given time is negligibly small (since our notion of time is continuous).

Let t denote absolute time (for example, the UNIX epoch) and $[t]$ the time-of-day, both expressed in hours. Thus, $[t] \in [0, 24)$ equals t modulo 24. We model the number of search query matching tweets over a time interval corresponding

to background chatter by a Poisson Point Process of rate $B([t])$. If no activity were taking place over an interval $\mathcal{S}$, the distribution of the number of tweets corresponding to baseline activity is given by a Poisson random variable of mean

$$m = \Lambda(\mathcal{S}) = \int_{\mathcal{S}} B([t])\,\mathrm{d}t. \tag{5.4}$$

An interesting event is said to occur when the observed number of tweets in such an interval exceeds m by a "significant" amount. But how should we define "significant," and how should we choose the interval length? In order to see this, we first review some basic detection-theoretic tradeoffs for a single Poisson random variable.

5.3.2 CFAR Detection under a Poisson model

Let X denote a random variable modeled as Poisson with nominal mean m, corresponding to background activity over an interval as in (5.4. Suppose that our decision rule is that an interesting event happens when $X > a$, for some $a > m$. The *false alarm rate* associated with this decision rule (i.e., the probability that this realization actually does come from the null hypothesis) is therefore given by

$$P_F(a, m) = P[X \geq a] = \sum_{k=a}^{\infty} \frac{m^k}{k!} e^{-m} \tag{5.5}$$

A simple criterion for choosing the threshold a is to maintain the false alarm rate below some threshold p_f. In detection theory, such tests are referred to as

Constant False Alarm Rate (CFAR) tests [53]. The advantage of this approach is that it requires only a model for the nominal distribution, and can leave deviations from the nominal unmodeled. This is well-matched to our application, where we observe that tweet rates during event times are heterogeneous and difficult to model.

An explicit expression for the false alarm rate is given by

$$P_F(a, m) = \frac{\Gamma(\lfloor a \rfloor + 1, m)}{\Gamma(\lfloor a \rfloor + 1, 0)}, \tag{5.6}$$

where $\Gamma(\ ,\ )$ refers to the (upper) incomplete gamma function given by $\Gamma(n, x) = \int_x^\infty t^{n-1} e^{-t} \mathrm{d}t$. However, for the purpose of our design, it is useful to develop additional insight into how P_F behaves as a function of a and m.

For additive deviations from the mean of order $\sqrt{m}$, we can employ a Gaussian approximation to see that

$$P_F(m + \alpha\sqrt{m}, m) \approx Q(\alpha)$$

where $Q(x) = \int_x^\infty \frac{e^{-u^2/2}}{\sqrt{2\pi}} du$ denotes the complementary CDF of a standard Gaussian random variable. The Gaussian approximation becomes tight for large m, as long as α does not grow with m. However, in order to make the false alarm rate low, we must make α large, and once we make it large enough (e.g., comparable to $\sqrt{m}$), we fall out of the Gaussian regime. We therefore consider multiplicative

deviations from the mean, setting $a = \gamma m$, where $\gamma > 1$. In this regime, we can use a Chernoff bound to obtain that

$$P_F(\gamma m, m) \leq \exp\left(-A(\gamma)m\right) \ , \quad \gamma > 1 \tag{5.7}$$

where the exponent of decay $A(\gamma) = \gamma \log \gamma - \gamma + 1$ is a positive and increasing function of γ, for $\gamma > 1$, and is asymptotically exact as m gets large.

What (5.7) tells us is that we need the *product* $A(\gamma)m$ to be large if we want to drive the false alarm rate down. Thus, if we want to detect unusual activity over a smaller time window over which the nominal number of tweets m is small, then we can only hope to detect larger multiplicative deviations (i.e., larger γ and hence $A(\gamma)$) while maintaining a desired value for P_F.

Furthermore, since detection performance depends only on m and γ, it is useful to divide time into bins that correspond to fixed values of m. Since our background process is time-varying, as discussed next, we find it convenient to transform it into an equivalent time-invariant process by introducing the concept of *virtual time,* such that the nominal mean over a virtual time bin depends only on its length.

5.3.3 Virtual time

We scale the time axis (depending on the time-of-day $[t]$) so that the frequency of tweets corresponding to background chatter $B(t)$ is a constant λ tweets per v-hour (we add the prefix "v-" to units of time in this new notion of time to stress that they are not real units of time but *virtual* ones), where λ is the *average* hourly baseline rate:

$$\lambda = \frac{1}{24} \int_{y=0}^{y=24} B(y)\ \mathrm{d}y. \tag{5.8}$$

Let

$$f(t) = t - [t] + F([t]) \tag{5.9}$$

denote the mapping from true time to virtual time, where $F(t)$, $0 \leq t < 24$ is given by

$$F(t) = \frac{1}{\lambda} \int_{y=0}^{y=t} B(y)\ \mathrm{d}y, \tag{5.10}$$

We reserve the notation t for the native notion of time and denote the virtual time $f(t)$ by x. The background tweet rate in virtual time equals λ, since $\lambda \mathrm{d}x = B(t)\mathrm{d}t$. A virtual time interval of length Δ corresponds to a nominal mean tweet count of $m = \lambda\Delta$, and we can now ask whether the observed number of tweets significantly exceeds it using the CFAR test described in the previous section. Once we detect the virtual event times, we map it back to native time: the virtual

time $x = f(t)$ is a continuous and increasing function of t, and is invertible when $B(t) > 0$.

Let E denote the set of observed tweet times (in native time), and $H = \{f(t) : t \in E\}$ the corresponding virtual times. We define the tweet count function for the virtual time interval $\mathcal{A}$ as

$$y(\mathcal{A}, H) = \sum_{x \in H} I_{\mathcal{A}}(x),$$

where $I_{\mathcal{A}}(x)$ is the indicator function of the set $\mathcal{A}$. We compare the tweet count function to appropriate thresholds for event detection, as described next.

5.3.4 Event detection

In virtual time, the nominal mean number of tweets over an interval $[a, a+\Delta)$ is $m = \lambda\Delta$, and we define an interesting event to occur, or the topic to be "trending," if the tweet count over the interval exceeds γm, where $\gamma > 1$. As discussed in Section 5.3.2, the false alarm rate depends on γ and m, and we must choose their combination so as to attain a given false alarm level p_f. Define $\vartheta(m, p_f)$ as the smallest value of γ such that $P_F(\gamma m, m) \leq p_f$. Instead of using the asymptotic approximation (5.7), we compute $\vartheta(m, p_f)$ by numerically inverting (5.6). These values can be pre-computed for false alarm rates of interest for various values of m. Figure 5.5 plots $\vartheta(\lambda\Delta, p_f)$ versus the mean parameter $m = \lambda\Delta$ for different false

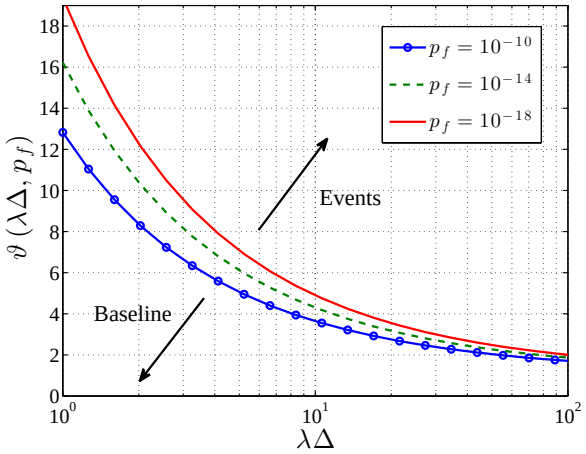

Figure 5.5: Smallest factor $\vartheta(\lambda\Delta, p_f)$ by which the frequency of tweets during trending times must increase for us to be able to detect it plotted against the average number of tweets expected $\lambda\Delta$.

alarm rates. For example, we can read off from the figure that for $\gamma = 2.5$ (which we take to be the smallest γ of interest in our experiments) and $p_f = 10^{-14}$, we need $m = \lambda\Delta \geq 40$, so that we can restrict attention to windows of width $\Delta = 40/\lambda$ v-hours.

Event detection algorithm

We can now design the size of the windows in virtual time that we should use to test for unusual activity. We need to test first for bursts of intense activity (large γ) over small time intervals (hence small $m = \lambda\Delta$). Once we detect and eliminate these, we should expand the window (larger $m = \lambda\Delta$) to detect less intense activity (smaller γ). The order in which we do this is important: an

intense but brief burst of activity which falls next to a period of baseline activity could be classified as a longer event if we do not first detect and eliminate the short burst. This leads to the following multi-resolution algorithm.

Algorithm: We take a multi-resolution approach in Algorithm 1, testing for activity in time windows of the form $[k\Delta, (k+1)\Delta)$ for increasing values of Δ. Since we were interested in activity of intensity greater than 2.5 times the baseline intensity (i.e., $\gamma \geq 2.5$), we use test windows of duration at most $40/\lambda$ v-hours. Specifically, we choose Δ of the form $\Delta_0 2^r,\ r = 0, 1, 2, 3$ with $\Delta_0 = 5/\lambda$ v-hours so that the largest test windows spanned $40/\lambda$ v-hours.

5.3.5 Learning baseline tweet profiles

We now describe how to estimate the baseline chatter frequency, given by $B(t),\ 0 \leq t < 24$ from the set of observed tweet times E (one may use a small subset of the observed tweet times for this training phase; we use the entire observation window of around 6 months). We assume that $B(t),\ t \in [0, 24)$ is a piecewise constant function of the form

$$B(t) = \sum_{l=0}^{l=24/\Delta-1} \phi(l) I_{[l\Delta,(l+1)\Delta)}(t), \tag{5.11}$$

for the time duration Δ hours (which we call coherence time). It is convenient to choose Δ so that it divides 24 hours (we use $\Delta = 40$ minutes).

Algorithm 1 Identify topic specific trending times

inputs Observed tweet times H (discrete set), observation window $\mathcal{W} \subseteq [x_{\text{start}}, x_{\text{end}})$, baseline rate λ, window sizes $\Delta_1 < \Delta_2 < \cdots < \Delta_K$ and false alarm rate p_f (we use 10^{-14}).

initialize $\mathcal{R} = \{\ \}$

for $l = 1$ to $l = K$ **do**

  for $r = \lfloor x_{\text{start}}\ / \Delta_l \rfloor$ to $r = \lfloor x_{\text{end}}\ / \Delta_l \rfloor$ **do**

    $\mathcal{S} = [r\Delta_l, (r+1)\Delta_l) \cap \mathcal{W}$ {the observation window $\mathcal{W}$ need not be contiguous}

    $\mathcal{S} = \mathcal{S} \setminus \mathcal{R}$ {discard portions declared active earlier}

    if $y(\mathcal{S}, H) \geq \lambda\,|\mathcal{S}| \times \vartheta\,(\lambda\,|\mathcal{S}|\,, p_f)$ **then**

      $\mathcal{R} = \mathcal{R} \cup \mathcal{S}$

    end if

  end for

end for

return $\mathcal{R}$

We start by pooling tweet counts corresponding to the same time-of-day $[k\Delta, (k+1)\Delta)$ into sets $\mathcal{N}(k)$ for $k = 0, 1, \ldots, (24/\Delta) - 1$. For each such set $\mathcal{N}(k)$, we want an estimate $\widehat{\phi}(k)$ of the baseline frequency $\phi(k)$. We explain the ideas behind our algorithm by focussing on one of the sets $\mathcal{N}(k)$, dropping the dependence on k in subsequent discussions. As per our piecewise constant baseline model, background chatter in the set $\mathcal{N}$ are realizations of Poisson $(\phi\Delta)$. If there was no topic specific activity at all, then the maximum likelihood (ML) estimate of ϕ is given by $(1/\Delta) \times \text{mean}\,(\mathcal{N})$. The sets $\mathcal{N}$ also contain tweet counts corresponding to topic specific trending times. The tweet counts during trending times can be orders of magnitude larger than those at times corresponding to baseline chatter. Therefore, one must account for such *outliers* while estimating ϕ.

The method we use to estimate the baseline profile ϕ resembles "inward" procedures for univariate outlier detection ([10] gives a brief survey of outlier detection methods). We assume that outliers only correspond to trending times (which by definition see heightened activity; i.e., $\gamma > 1$). Therefore, the small values in $\mathcal{N}$ are likely to correspond to the baseline. Assuming that the proportion of baseline entries is high, a robust estimate of ϕ is given by $\widehat{\phi} = (1/\Delta) \times \text{median}\,(\mathcal{N})$. We use this estimator for ϕ because (i) the median of a Poisson distribution is approximately equal to its mean $\phi\Delta$ (ii) the empirical median is more resilient to

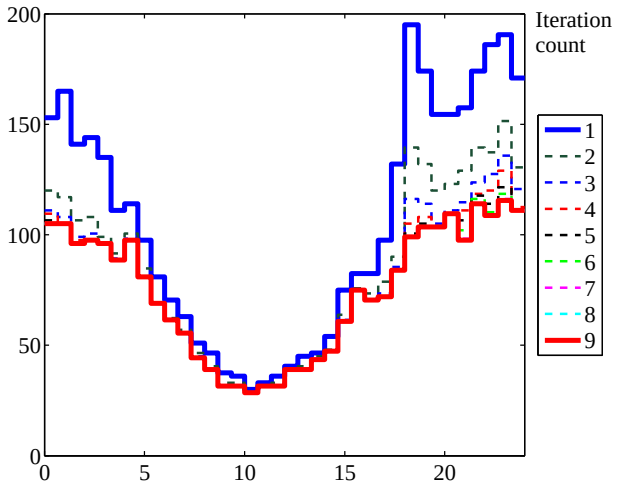

Figure 5.6: Iterative refinement of the estimate of $B(t)$ (expressed in #tweets per hour) by Algorithm 2 until convergence (`South Park`).

outliers than the empirical mean. When the fraction of outliers δ is very small, this is a good estimate of ϕ.

When this fraction is not very small, the empirical median $\widehat{\phi}\Delta$ overestimates the mean parameter $\phi\Delta$ of the Poisson distribution. We must therefore refine our estimate $\widehat{\phi}$ further. We do this by identifying tweet counts in the set $\mathcal{N}$ that are inconsistent with the current estimate of the baseline model Poisson($\widehat{\phi}\Delta$). These are the tweet counts $n \in \mathcal{N}$, that exceed the threshold $\widehat{\phi}\Delta \times \vartheta(\widehat{\phi}\Delta, p_f)$ for activity. Such counts are pooled into a set $\mathcal{R}$. As these observations are inconsistent with the current baseline estimate, we treat them as outliers and discard them when we refine our estimate of ϕ via $\widehat{\phi} = (1/\Delta)\text{median}(\mathcal{N} \setminus \mathcal{R})$. We repeat the preceding refinement step until convergence. It is easy to show that the sequence of sets

$\mathcal{R}$ (across iterations) is non-decreasing and convergence is therefore guaranteed. We summarize this procedure in Algorithm 2 and plot the associated refinement procedure in Figure 5.6. Notice that the estimates of $B(t)$-s across iterations (for the same time-of-day t) is non-increasing; this follows from the non-decreasing nature of the sequence of sets $\mathcal{R}$.

Note that, during this baseline estimation phase, we are lenient towards false alarms in setting our thresholds. For example, we set $p_f = 10^{-4}$ now, in contrast to the event times detection Algorithm 1 where we set p_f to 10^{-14}.

Summary: Given the set of tweet times E, we first estimate the baseline rate $B(t)$ using Algorithm 2. Using this estimate of $B(t)$, we transform to virtual time $x = f(t)$ via (5.9) in which the baseline rate is a constant λ tweets per v-hour ($B(t)$ is piecewise constant which results in a piecewise linear, easy to evaluate map $f(t)$). The set of virtual tweet times is the map H of the set E. On this set H, we employ Algorithm 1 (for the estimated value of $\lambda = \text{mean}(\widehat{\phi}(k))$) to detect event times across time scales. Having identified such topic specific times, we revert back to native time using the inverse map $t = f^{-1}(x)$, which as we noted in section 5.3.3 exists when $B(t) > 0$ (or $\widehat{\phi}(k) > 0$).

Algorithm 2 Estimate baseline profile

1: **inputs** Coherence time Δ (should divide 24 hours; we use 40 minutes) for the piecewise constant approximation, observed tweet count set $\mathcal{N}(k)$, $k = 0, \ldots, 24/\Delta - 1$ for the time-of-day $[k\Delta, (k+1)\Delta)$ and the false alarm rate parameter p_f (we use 10^{-4}).

2: **initialize** $\widehat{\phi}(k) = (1/\Delta) \times \text{median}(\mathcal{N}(k))$ for $k = 0, 1, \ldots, (24/\Delta) - 1$.

3: **repeat**

4: **for** $k = 0$ to $k = (24/\Delta) - 1$ **do**

5: $\mathcal{R}(k) = \left\{ y \in \mathcal{N}(k) \ : \ y \geq \widehat{\phi}(k)\Delta \times \vartheta(\widehat{\phi}(k)\Delta, p_f) \right\}$

6: $\widehat{\phi}(k) = (1/\Delta) \times \text{median}(\mathcal{N}(k) \setminus \mathcal{R}(k))$

7: **end for**

8: **until** All $\widehat{\phi}(k)$-s have converged

9: **return** Estimate of the baseline profile

$$B(t) = \sum_{k=0}^{k=(24/\Delta)-1} \widehat{\phi}(k) I_{[k\Delta,(k+1)\Delta)}(t).$$

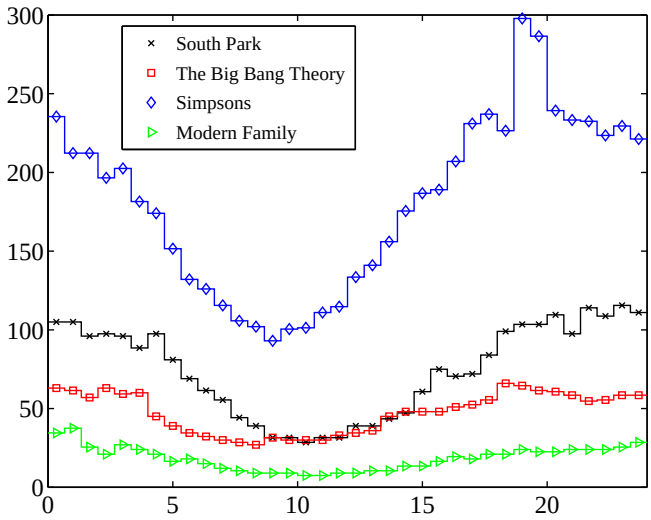

Figure 5.7: Baseline profile $B(t)$ (expressed in #tweets per hour) estimated by Algorithm 2 plotted against the time-of-day expressed in hours ($t = 0$ hours corresponds to 0:00:00 UTC).

5.3.6 Results

We queried the Twitter streaming API (using Tweepy[1]) for mentions of four television shows (from the first week of November 2013 up until the first week of May 2014). We plot the estimated baseline profiles in Figure 5.7. They indicate that the time-of-day effects are pronounced.

We plot event detection results in Figure 5.8 and summarize them in Table 5.1. We see that we identified topic specific activity that spanned about $10 - 20$ days in this 6 month window. The number of tweets that correspond to background chatter is more than 50% for the four shows, indicating that such simple temporal

Topic	**Total trending time (in days)**	**Max. event span (hours)**	**Total #tweets**	**#tweets in events**
South Park	19.3	9.8	641113	282130
TBBT	10.4	9	324962	104467
Simpsons	12.6	5.6	1237271	360878
Modern Family	16.7	20.6	162483	78441

Topic	**#events**	**Total #users**	**#users in events**
South Park	910	328933	154715
TBBT	330	170863	62515
Simpsons	1117	837206	295430
Modern Family	211	92339	45023

Table 5.1: Summary of topic specific trending times identified in the 171 day observation window for 4 TV shows

analysis can have value in helping us to focus on what matters for each topic when we wish to do more detailed analysis such as text analysis. We collapse contiguous activity windows into a single event. From Table 5.1, we see that the maximum event duration for the show `Mordern Family` is the longest. This stems from the fact that the baseline rates $B(t)$ for this show are small (the average rate λ is about 18 tweets per hour; Figure 5.7). Therefore, when we test for activity we use large test windows for this show. We chose the smallest time window Δ_0 to be $5/\lambda$ v-hours. This is results in poor time resolution for shows that have small average rates λ.

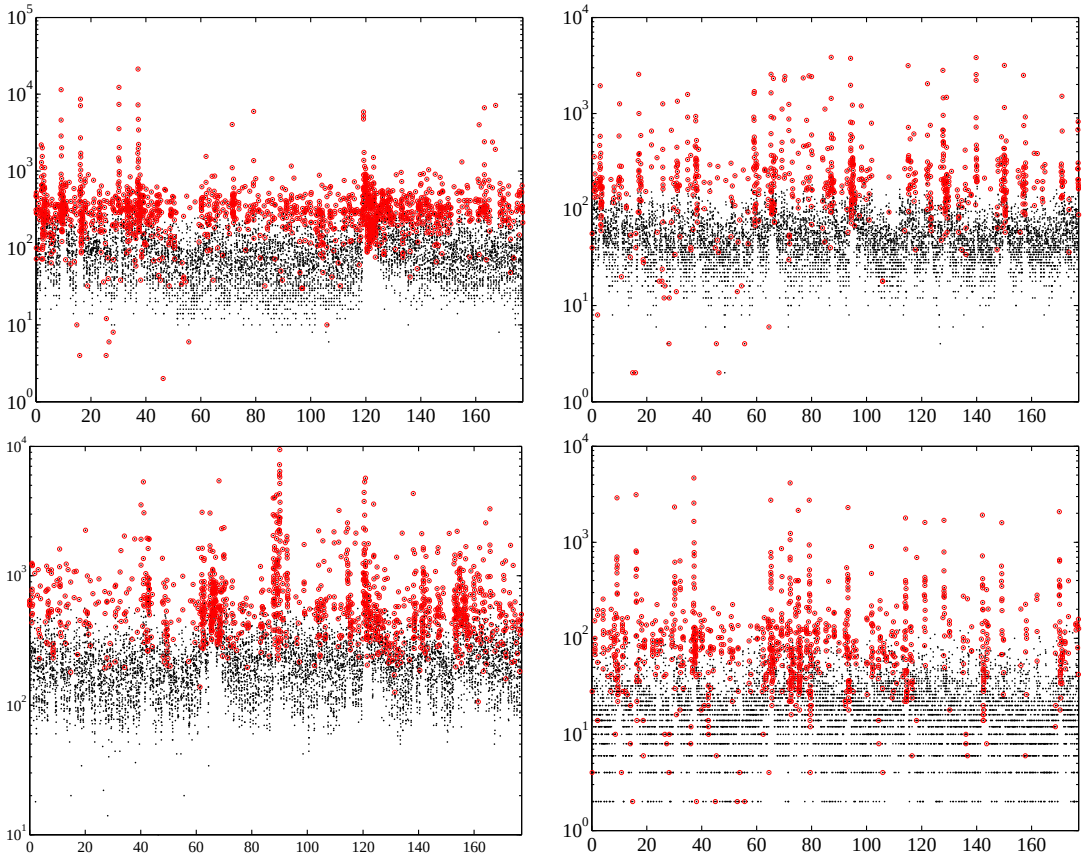

Figure 5.8: The observed frequency of tweets profile (in black) expressed in #tweets per hour (estimated using 30 minute windows) plotted against the time elapsed in days. Trending times are in marked in red. Top (left to right): `South Park` and `The Big Bang Theory`, Bottom (left to right): `Simpsons` and `Modern Family`

5.4 Discovering event relationships

Now that we have a systematic framework (via algorithms 1 & 2) for identifying events associated with a topic, we are interested in whether there are sub-topics (where this term is used quite loosely to denote phenomena associated with different causes) that emerge from a closer examination of these events. While text analysis may be a useful tool for this purpose, we focus here on whether it is possible to identify sub-topics from *crowd behavior* alone. Our basic premise is that when "many" users who tweet during an event also tweet during another event, it is likely the underlying reason behind the two events is the same. Based on this premise, we develop an algorithm which partitions the set of event times into sets that correspond to different reasons, by trying to ensure that two events which share a significant user pool are placed in the same partition. We only use tweets (matching our query) which fall within any of the identified event times, and use the tweet times and authorship (user IDs) for event clustering. Thus, starting with the observed tweet profile over time (Figure 5.9 top row), we extract event times (Figure 5.9 middle row) and would now like to present these event times with their relationships revealed (Figure 5.9 bottom row).

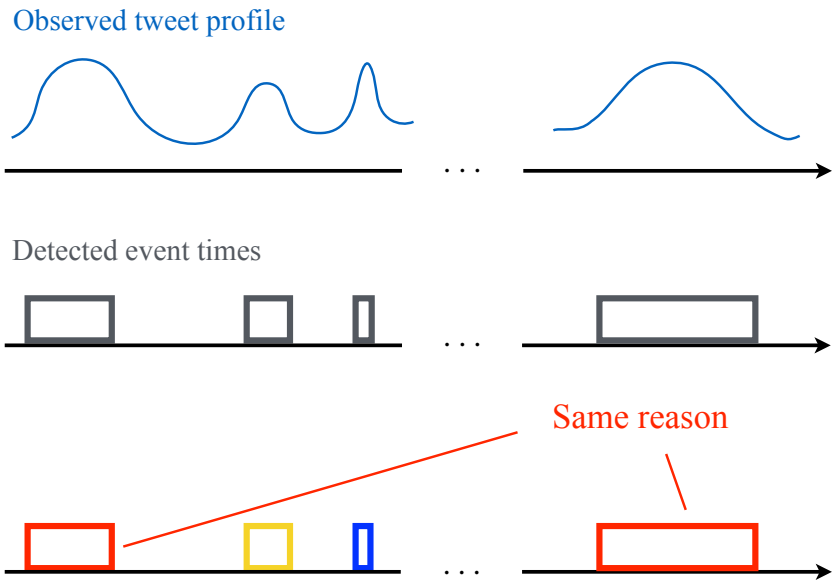

Figure 5.9: Dissecting detected activity

We first present the method employed to partition identified event times, then discuss its computational cost, and conclude this section by presenting the results for the TV show data described in Section 5.3.6.

5.4.1 "Distance" between events

In order to associate users with events, we build an undirected weighted bipartite graph B, as shown in Figure 5.10, with the n events on one side and m users with at least two search query matching tweets during these n event times on the other (note that $n \ll m$). The strength of the link $c(e, u) \geq 0$ between event e and user u could be the number of search query matching tweets by user u during

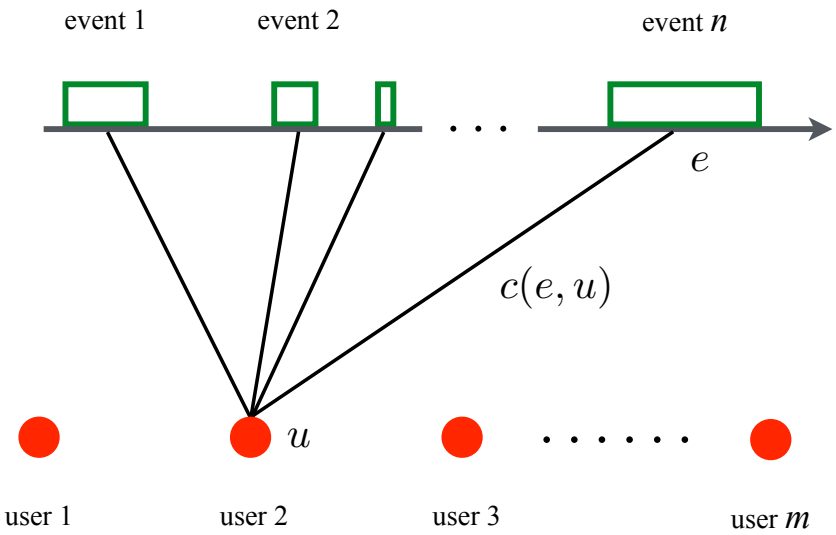

Figure 5.10: User-event bipartite graph B used to cluster events

event e. In order to cluster events, we need to choose a suitable definition for the distance between two events which ensures that any two events which share a large pool of common users are "close" to one another. We choose the (average) commute time between the two events for the random walk on B as our distance metric.

Commute times: The random walk on a weighted undirected graph $G(V, E)$, with edge weights $w(u, v)$ (the convention is $w(u, v) = 0$ if $(u, v) \notin E$) is a Markov chain on the state space V with the state transition function

$$\Pr\left[\text{Location}(l) = v | \text{Location}(l-1) = u\right] = \frac{w(u,v)}{d(u)}, \tag{5.12}$$

where $d(u)$ is the (weighted) degree of the vertex u given by $d(u) = \sum_{k \in V} w(u, k)$ and l is the number of hops taken by random walk so far. The average number

of hops taken by a random walk starting at vertex a to visit vertex b for the first time is the access time of b from a. The sum of the access times from a to b and from b back to a is the commute time of the random walk between vertices a and b.

For the user-event bipartite graph in Figure 5.10, when is the commute time between two events small? A random walk starting at event a is more likely to reach event b quickly & vice-versa when "many" users who tweet during event a also tweet during event b (the exact definition of "many" will depend on the choice of the link strength metric $c(e, u)$). The first hop of a random walk initiated at event a will take it to one of users who talked during event a. If this set of users overlaps heavily with the set of users who talk during event b, there is a "good chance" that the second hop of the random walk sees event b.

5.4.2 Commute time embedding of a graph

It is possible to give a map $\mathbf{m}(\cdot)$ from V to $\mathbb{R}^{|V|-1}$ such that the *square* of the Euclidean distance between $\mathbf{m}(u)$ and $\mathbf{m}(v)$ is equal to the commute time between the vertices u and v. This map $u \mapsto \mathbf{m}(u)$ is sometimes referred to as the "commute time embedding" of the graph G.

Denoting the (weighted) adjacency matrix with entries $w(u, v)$ by A and the diagonal matrix with diagonal entries equal to $d^l(u)$ by D^l, the *normalized adjacency matrix* N of the graph is the symmetric matrix $D^{-1/2}AD^{-1/2}$. The normalized adjacency matrix is closely related to the random walk on the weighted graph G. Let the eigenvalues of N be $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_{|V|}$ and the corresponding eigenvectors be $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{|V|}$ (normalized so that $\mathbf{z}_k^T \mathbf{z}_k = 1$). Let the u-th element of the eigenvector $\mathbf{z}_k$ be $z_k(u)$ and let

$$\pi(u) = d(u) \Big/ \sum_{v \in V} d(v)$$

denote the stationary distribution of the random walk on this weighted graph. The commute time embedding of the vertex u is given by [37]

$$\mathbf{m}(u) = \frac{1}{\sqrt{\pi(u)}} \left(\begin{array}{cccc} \frac{z_2(u)}{\sqrt{1-\nu_2}} & \frac{z_3(u)}{\sqrt{1-\nu_3}} & \cdots & \frac{z_{|V|}(u)}{\sqrt{1-\nu_{|V|}}} \end{array} \right). \tag{5.13}$$

Commute time embedding of events in the bipartite graph B: The adjacency matrix A of any bipartite graph with two independent sets of size n and m can be written in the following form:

$$A = \left(\begin{array}{cc} 0_{n \times n} & X \\ X^T & 0_{m \times m} \end{array} \right).$$

For the graph in Figure 5.10, the (e, u)-th element of the $n \times m$ matrix X is given by $c(e, u)$, the strength of the link between event e and user u. Since A takes the

above form, we can write the normalized adjacency matrix as:

$$N = \begin{pmatrix} 0_{n\times n} & D_1^{-1/2} X D_2^{-1/2} \\ D_2^{-1/2} X^T D_1^{-1/2} & 0_{m\times m} \end{pmatrix},$$

where D_1^l denotes the $n \times n$ diagonal matrix with the degrees of events $d_1(e) = \sum_{u=1}^{u=m} c(e,u)$ raised to power l on the diagonal and D_2^l denotes the $m \times m$ diagonal matrix with degrees of users $d_2(u) = \sum_{e=1}^{e=n} c(e,u)$ raised to power l on the diagonal. We denote $D_1^{-1/2} X D_2^{-1/2}$ by $\widetilde{X}$. Let

$$\widetilde{X} = \sum_{k=1}^{k=n} \mu_k \mathbf{p}_k \mathbf{q}_k^T$$

be the singular value decomposition of $\widetilde{X}$. We denote the column vector $\begin{bmatrix} x^T & y^T \end{bmatrix}^T$ by $[x; y]$. Consider the vector $[\mathbf{p}_k; \mathbf{q}_k]$: It is easy to verify that this is an eigenvector of N corresponding to eigenvalue μ_k. Similarly the vector $[\mathbf{p}_k; -\mathbf{q}_k]$ is also an eigenvector, with corresponding eigenvalue $-\mu_k$. All other eigenvalues of N are equal to zero; there are $m - n$ eigenvectors of N corresponding to the eigen value 0 (assuming that n singular values of the $n \times m$ matrix $\widetilde{X}$ are nonzero). These eigenvectors are of the form $\{[0_{n\times 1}; \mathbf{x}_l]\}$, where $\{\mathbf{x}_l\}$ can be chosen to be any orthonormal basis perpendicular to $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$.

While some of the preceding observations have been made in prior work on spectral partitioning of bipartite graphs [17], unlike [17], we wish to cluster vertices only on one "side" of the bipartite graph B (specifically, the events). From the

special structure of the eigenvectors of the normalized adjacency matrix N of the user-event bipartite graph, we can drastically reduce the dimensionality of the embedding and show that the commute time embedding of the event e is given by:

$$\mathbf{m}(e) = \frac{1}{\sqrt{\pi(e)}} \left(\begin{array}{cccc} \frac{p_2(e)}{\sqrt{1-\mu_2^2}} & \frac{p_3(e)}{\sqrt{1-\mu_3^2}} & \cdots & \frac{p_n(e)}{\sqrt{1-\mu_n^2}} \end{array} \right), \tag{5.14}$$

where $p_k(e)$ is the e-th row of the left singular vector $\mathbf{p}_k$ of $\widetilde{X}$ and $\pi(e) = d_1(e)/(2\sum d_1(k))$. Note that this is an $n-1$ dimensional embedding and not a $|V|-1 = m+n-1$ dimensional embedding, which is the case in general. This reduction in dimensionality follows from (i) the bipartite nature of the graph and (ii) the fact that only commute times between events (which form one of the two "sides" of B) interests us. It is important to note that the eigenvalues μ_k and $-\mu_k$ of the normalized adjacency matrix N appear via a single term $1/\sqrt{1-\mu_k^2}$ in (5.14) rather than via $1/\sqrt{1-\mu_k}$ and $1/\sqrt{1+\mu_k}$ as one would expect from the expression for the commute time embedding in (5.13).

Typically, the first "few" singular values are close to 1 (in fact $\mu_1 = 1$) and the later μ_k-s fall off to zero. The $1/\sqrt{1-x^2}$ non-linearity that appears in the expression for the commute time embedding further amplifies this fall off. This motivates us to approximate the commute time embedding of events using leading singular vectors $\mathbf{p}_k,\ k = 2, \ldots, L+1$ of $\widetilde{X}$ to arrive at an $L \ll n$ dimensional

embedding in which Euclidean distance squares approximate commute times on the bipartite graph in Figure 5.10. We refer to these L-dimensional embeddings as the "approximate commute time embeddings" and denote them by $\widehat{\mathbf{m}}(e)$.

We can use the square of the Euclidean distance between these L-dimensional embeddings of events, which roughly approximates commute times, as the distance metric when we cluster events. In practice, we notice better clustering performance when we normalize the approximate commute time embedding $\widehat{\mathbf{m}}(e)$ to unit norm vectors $\widetilde{\mathbf{m}}(e)$ by setting $\widetilde{\mathbf{m}}(e) = \widehat{\mathbf{m}}(e)/\|\widehat{\mathbf{m}}(e)\|$ (similar behavior has been observed in [42], where spectral methods are used to cluster points in K-dimensional space; they however do not scale the eigenvectors $\mathbf{z}_l$ by $1/\sqrt{1-\nu_l}$ as we do here). We refer to $\widetilde{\mathbf{m}}(e)$ as the normalized (approximate) commute time embedding of the event e. In section 5.4.3, we briefly describe the algorithm we employ to cluster the embeddings $\{\widetilde{\mathbf{m}}(e)\}$ of events.

The need for normalization: Using a simple connection between commute times and recurrence times, we can see why the preceding normalization is required to facilitate clustering. The recurrence time $r(u)$ of a vertex u is defined as the average number of hops taken by the random walk starting off at vertex u to return back to u for the first time. Recurrence times (for a connected graph with $|V| < \infty$) are inversely proportional to the stationary distribution of the random walk i.e., $r(u) \propto 1/\pi(u) \propto 1/d(u)$. The commute time between two

vertices a and b must be at least as large as the larger of the two recurrence times $r(a)$ and $r(b)$. As a result, a low degree vertex u is "far away" from all vertices in the commute time embedding of the graph even when this vertex is tightly connected to only one of the "dominant clusters" of the graph, making its cluster membership unambiguous. This is also fairly evident in the expression for the commute time embedding of u in (5.13), where we see that if the vertex u has small degree $d(u) \propto \pi(u)$, then the $1/\sqrt{\pi(u)}$ scaling in (5.13) pushes its commute time embedding away from the origin, whereas the embeddings of high degree vertices are pushed toward the origin. When a vertex is densely connected to only one of the dominant clusters of the graph (its "home" cluster), it is likely that most traversals of the random walk from any vertex which lies outside its home cluster to the tagged vertex u will pass via other vertices in its home cluster. Therefore, the commute times between the tagged vertex u and vertices in its home cluster must be smaller than the commute times between u and a vertex outside its home cluster. Thus, the *ordering* of the commute times to the tagged vertex u still retains the "neighborhood" structure of the graph even if this low degree vertex u is far away from all vertices. Since the magnitude of the commute time embedding (roughly proportional to $1/\sqrt{d(u)}$) does not capture this neighborhood information, its direction $\mathbf{m}(u)/\|\mathbf{m}(u)\|$ must have captured it. Therefore, we normalize the approximate commute time embedding of events $\widehat{\mathbf{m}}(e)$ to arrive at

$\widetilde{\mathbf{m}}(e)$ and use these for the subsequent clustering phase. This explicitly reveals to standard Euclidean clustering algorithms such as k-means (a variant of which is applied to this embedding) that it is only the directions of the commute time embeddings that matter when we cluster them.

5.4.3 Clustering events using normalized commute time embeddings

We map events to a collection of n points $\{\widetilde{\mathbf{m}}(e)\}$ on the unit hypersphere in $\mathbb{R}^L$, so that the Euclidean distance between points roughly captures the *ordering* of commute times of the random walk on the user-event graph. We use deterministic annealing [49] (which can be roughly viewed as a version of k-means which adapts k to the data) to cluster these embeddings $\{\widetilde{\mathbf{m}}(e)\}$. We will now quickly walk through the mechanics of this clustering algorithm.

Deterministic Annealing (DA) for squared Euclidean distance: DA is a soft partitioning algorithm that attempts to best approximate the collection of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ using a mixture of k circularly symmetric Gaussian random variables centered at $\mathbf{y}_j,\ j = 1, \ldots, k$ of equal variance $T/2$ and mixture weights p_j. The parameter T is referred to as the "temperature" (analogy to physical annealing) and each of $\mathbf{y}_j$-s are called cluster centers/centroids. Each iteration

of DA involves a cooling step $T = \alpha T,\ \alpha < 1$, followed by a few EM iterates to optimize cluster center positions $\mathbf{y}_j$ and mixture weights p_j so as to minimize the average squared Euclidean distance of the data points $\{\mathbf{x}_i\}$ from the cluster centers $\mathbf{y}_j$ (as per the mixture of k Gaussians model). The algorithm starts off at a high temperature T. Since the variance of each cluster $T/2$ is high, all cluster centers $\mathbf{y}_j$ collapse to a single point and we have effectively one cluster (i.e., $k = 1$). As the temperature is progressively lowered, distinct cluster centers $\mathbf{y}_j$ emerge (typically one at a time), thereby increasing k. The temperature below which the next cluster will emerge can be predicted on the fly [49]. It is also possible to simply divide each cluster into two sufficiently close clusters and check whether or not the two collapse into a single cluster after a few EM iterates. At any temperature, we can pause and ask for the maximum of $p(\mathbf{y}_j|\,\mathbf{x}_i)$ for each point $\mathbf{x}_i$ from among the $j = 1, \ldots, k$ clusters. This gives us a k-partition of the n points $\{\mathbf{x}_i\}$. If this partition is "good", we can stop the annealing procedure. We now elaborate on our metric for "partition goodness" when we run DA with $\mathbf{x}_i = \widetilde{\mathbf{m}}(i)$.

Partition quality: The two-step random walk on events in the user-event bipartite graph B (when we treat 2 hops given by event–user–event as a single hop) can be shown to be the same as a random walk on the induced "event similarity

graph" (on n vertices) whose adjacency matrix is given by

$$A_s = D_1^{1/2} \widetilde{X} \widetilde{X}^T D_1^{1/2} = X D_2^{-1} X^T. \tag{5.15}$$

Indeed, if we had started off with the corresponding similarity metric between two events events a and b given by the (a, b)-th element of A_s,

$$w(a, b) = \sum_{u=1}^{u=m} \frac{c(a, u)c(b, u)}{d_2(u)}, \ 1 \leq a, b \leq n, \tag{5.16}$$

the resulting commute time embedding would have been the same as (5.14) (except for the scale factor $\sqrt{2}$ which accounts for the fact that two hops on B correspond to one effective hop on the event similarity graph). Therefore, we compute the modularity score of the partition $\{S_l, \ l = 1, \ldots, k\}$ induced by the association probabilities $p\left(\mathbf{y}_j | \, \mathbf{x}_i\right)$ of DA on *this* event similarity graph when judging its quality (the bipartite graph B is a graph on $m + n$ vertices and its adjacency matrix A cannot be used directly to test the quality of a partition of the n events; therefore, we use the above similarity graph's adjacency matrix A_s (a $n \times n$ matrix) to compute modularity scores). The modularity score of the partition $\{S_l\}$ on this weighted graph is given by the expression [25]:

$$Q(\{S_l\}) = \sum_{l=1}^{l=k} \left(\sum_{a,b \in S_l} \frac{w(a, b)}{M} - \left(\sum_{a \in S_l} \frac{d_1(a)}{M} \right)^2 \right),$$

where $M = \sum_e d_1(e)$, the total edge weight on the user-event bipartite graph (which is also equal to twice the total edge weight on the "event similarity" graph).

Summary: We store the best partition $\{S_l^{\text{opt}}\}$ induced by DA as we progressively decrease temperature T (using modularity on the event similarity graph as the measure of goodness). We run DA until the number of clusters k exceeds the maximum number $k_{\max}$ (we set $k_{\max} = n/10$ for our results) and return the best partition encountered so far.

5.4.4 Complexity

Typically, the number of events n is small, whereas the number of users m can potentially grow large over time (for the 4 TV shows we followed for 6 months, we identified $n < 1200$ events, while the number of users, m, who tweet during events is roughly 200 times bigger; see table 5.1). In order to compute the normalized commute time embeddings, we need the first $L + 1$ eigenvalues and eigenvectors (normalized to unit norm and chosen to be orthogonal in the case of repeated eigenvalues) of $\widetilde{X}\widetilde{X}^T$. Constructing this dense $n \times n$ matrix from observations $\{c(e, u)\}$ requires $O(n^2)$ memory and $O(mn^2)$ operations. We are interested in only the dominant eigenvalues & vectors of the positive semi-definite matrix $\widetilde{X}\widetilde{X}^T$ and iterative methods such as Lanczos and subspace iteration can be used to solve this problem efficiently (roughly $O(Ln^2)$ operations for the dense matrix $\widetilde{X}\widetilde{X}^T$). For the clustering phase, each DA iterate is of complexity $O(Ln)$, while evaluating

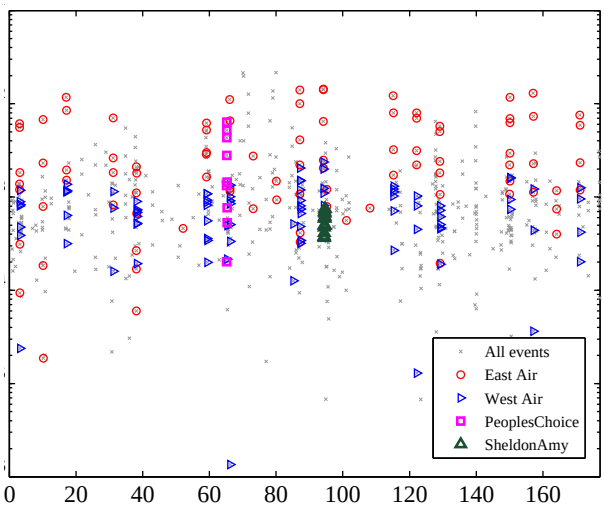

Figure 5.11: Identified events and their groupings (time in days) for the TV show `The Big Bang Theory`. y-axis is #tweets during each event in log scale

the modularity score of each partition takes $O(n^2)$ operations on the dense event similarity graph A_s.

5.4.5 Implementation details

We specify implementation considerations before presenting our results. We use Algorithm 1 to identify times of increased activity. We refer to each continuous segment of activity as an event. While an event identified using Algorithm 1 may span a long time duration, we wish to give our clustering algorithm the opportunity to dissect it into sub-events associated with different "reasons," and therefore slice events into segments of at most 30 minutes. This enabled us, for

example, to dissect east and west coast US air times, which were typically returned as a single event for the TV show `The Big Bang Theory` (Figure 5.11).

On the other hand, we do recognize that events that are nearby in time are likely to share the same reason, and we build this intuition into our choice of weight, $c(e, u)$, between event e and user u. A natural choice is to set $c(e, u)$ equal to the number of tweets made by the user u during the event e. However, better clustering performance is obtained by a softer notion of tweet count obtained by spreading the contribution of a tweet to multiple events in its vicinity. Let $\mathcal{S}(e)$ denote the time window corresponding to the event $e = 1, \ldots, n$. For a tweet authored by user u at time q during any one of the events, we first compute the probability mass under a two-sided exponential p.d.f. whose mean is q and standard deviation is σ (we choose $\sigma = 30$ minutes) in the interval $\mathcal{S}(e)$, as follows:

$$z(e, q) = \frac{1}{\sqrt{2}\sigma} \int_{\mathcal{S}(e)} \exp\left(-\sqrt{2}|t - q|/\sigma\right) \, \mathrm{d}t.$$

Letting $T(u)$ denote the set of tweet times of a user that fall within any one of the events $\mathcal{S}(e),\ e = 1, \ldots, n$, we define $c(e, u)$ as $\sum_{q \in T(u)} \left(z(e, q) / \sum_{k=1}^{k=n} z(k, q) \right)$. This operation spreads the membership of a tweet at time q across events in the neighborhood of q.

5.4.6 Results

We now summarize results (using a $L = 30$ dimensional embedding) for this phase of the temporal inference system, which dissects events into clusters. When we examine the text of tweets in the identified grouping of event times, we see these partitions are consistent with the underlying cause for the topic to trend. These results suggest that massive online feeds can be well understood with minimal processing by restricting attention to just coarse authorship information (user ID & time).

We present word clouds of the identified clusters for `The Simpsons` and manual summaries for the other TV shows. In Figure 5.12, we plot nine clusters of the detected events of `The Simpsons` sorted in descending order of modularity scores. Cluster 4 matches US air times, while cluster 7 matches UK air times. Cluster 6 corresponds to reaction to a sketch referring to the Oscars 2014 ceremony that was posted on `The Simpsons` official Twitter feed. Cluster 2 matches periodic resurgence of news regarding a `Simpsons`-themed LEGO set on Twitter. Cluster 3 mainly consisted of retweets of a post made by a YouTube celebrity that involved the `Simpsons` characters. Most of the tweets in cluster 1 were reactions to a tweet posted on `The Simpsons` official feed referring to a 2005 episode of the show that happened to depict a NFL Superbowl final with the same two teams

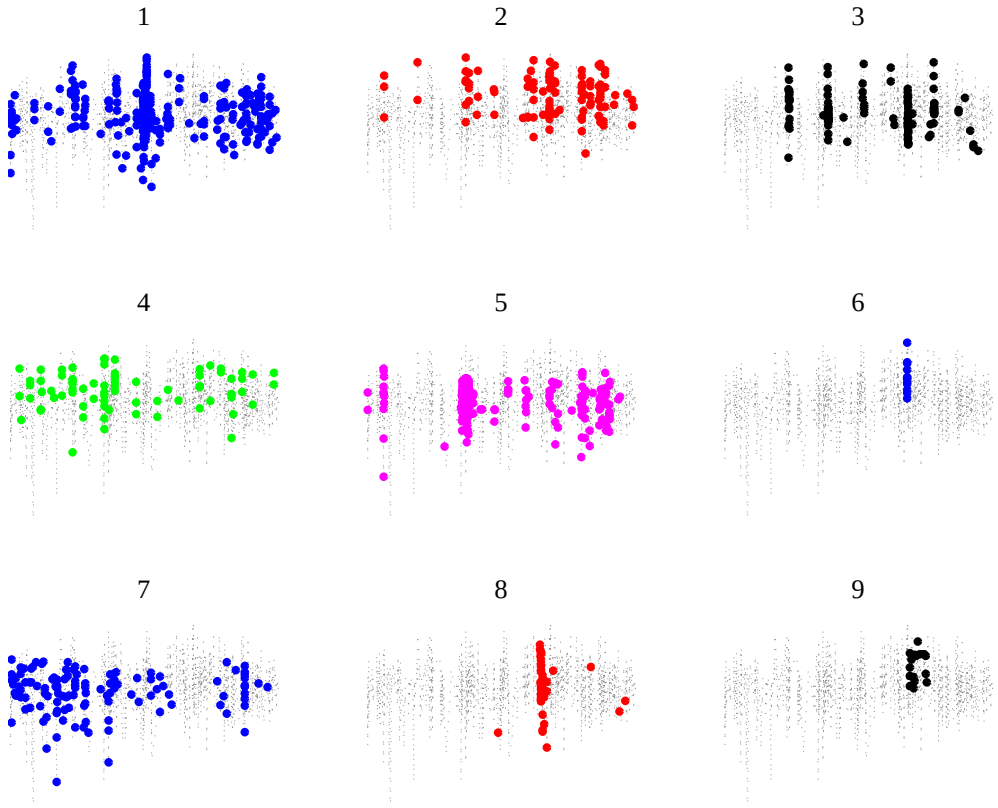

Figure 5.12: Partition of detected events for the show `Simpsons`

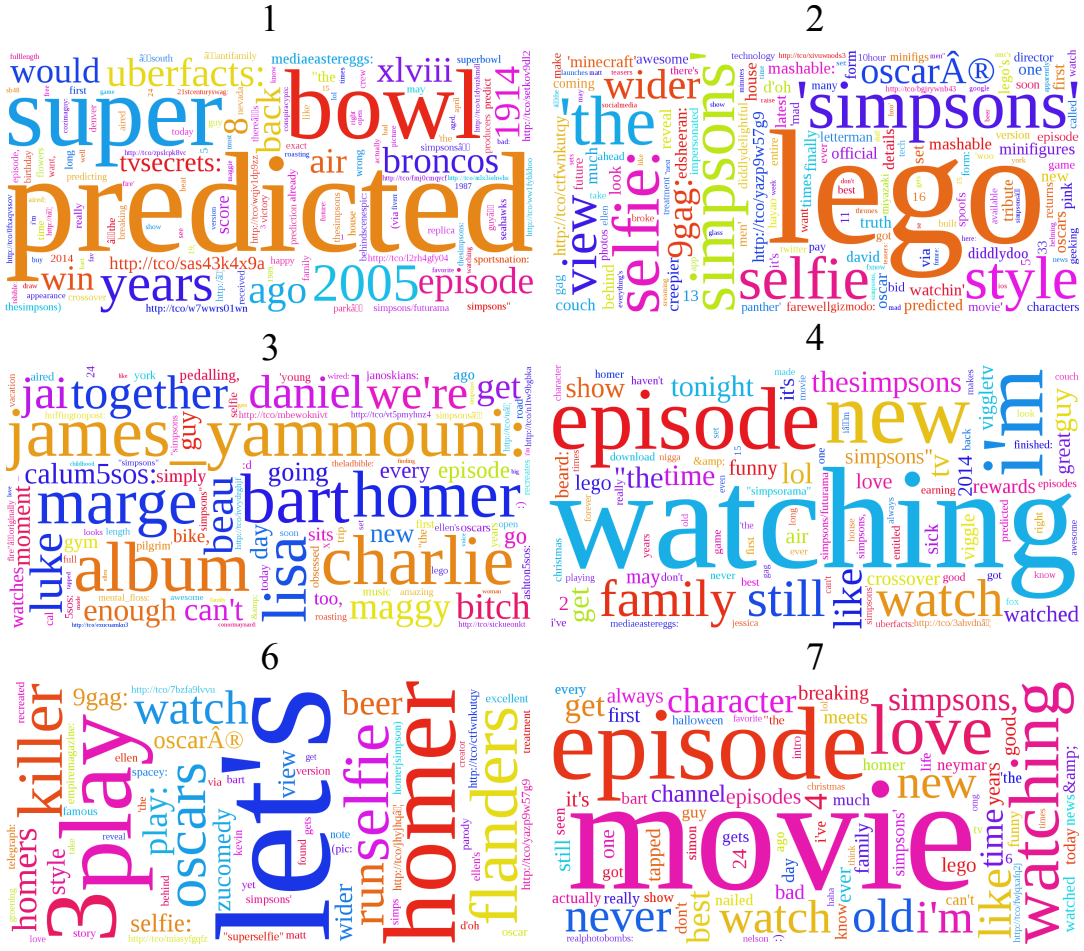

Figure 5.13: Word clouds for the show `Simpsons`

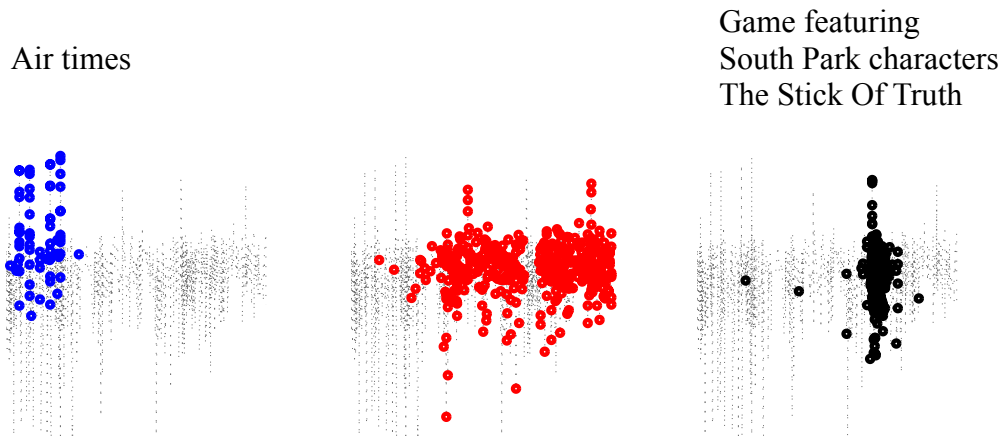

Figure 5.14: Partition of detected events for the show `South Park`

that made it to the final this year. However, other seemingly unrelated events have also been placed in cluster 1. The word clouds for these clusters are shown in Figure 5.13. Cluster 8 corresponds to news of a deal between "Chelsea FC" soccer team franchise and the makers of `The Simpsons`.

We display clusters along with short summaries for the three other shows in Figures 5.14, 5.15 and 5.16. The clusters are sorted in the descending order of modularity scores. We give summaries only for clusters for which the correspondence with real world events is unambiguous.

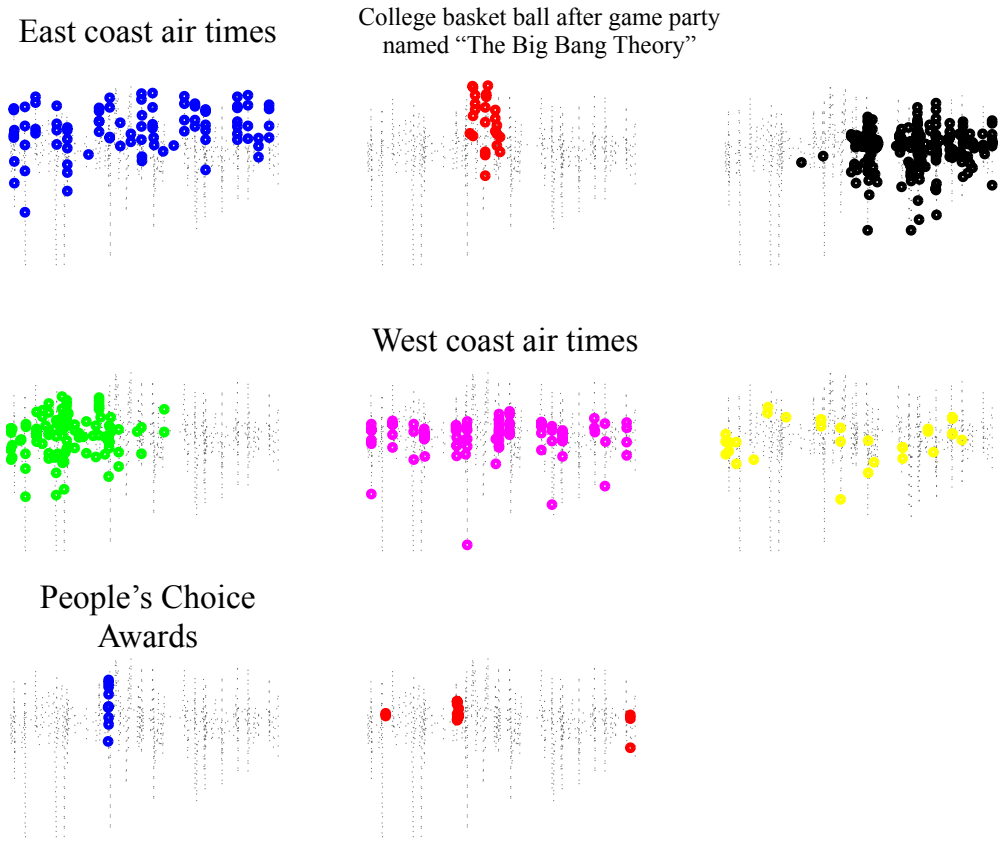

Figure 5.15: Partition of detected events for the show `The Big Bang Theory`

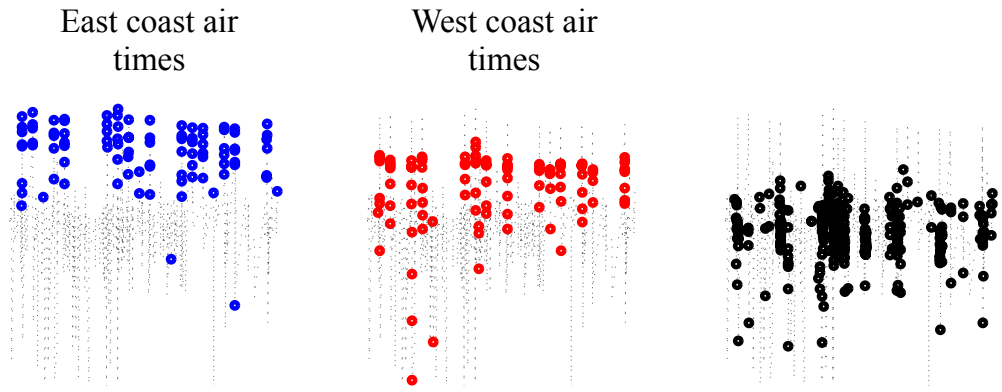

Figure 5.16: Partition of detected events for the show `Modern Family`

Chapter 6

Conclusions

In this dissertation, we give three examples which demonstrate that by carefully restricting the volume of measurements/coordination signals, we can arrive at scalable solutions.

6.1 Compressive parameter estimation in AWGN

For parameter estimation in AWGN, we have identified isometry conditions under which the only effect of making compressive measurements is an SNR penalty equal to the dimensionality reduction factor. We prove this by establishing a connection between the isometry conditions and the CRB/ZZB. Based on the threshold behavior of the ZZB, we also observe that, in order to avoid large estimation errors, the compressive measurements must not only preserve the geometry, but the SNR after the dimension reduction penalty must also be above a threshold.

Motivated by channel estimation for 60GHz systems, we investigate the problem of compressive frequency estimation. Specifically, for a mixture of K sinusoids of length N, we show that $O(K \log NK\delta^{-1})$ measurements suffice to provide isometries, where δ is the smallest singular value of appropriate matrices (a complete characterization is given for $K = 1$). We verify that the convergence of the ZZB to the CRB can be used to tightly predict the number of measurements needed to avoid error floors for frequency estimation of a single sinusoid.

Open Issues

While we identified geometry preservation conditions for the CRB and the ZZB in AWGN, a topic for further investigation is whether any additional geometry preservation conditions are needed for other error bounds like the Weiss-Weinstein bound for the same AWGN measurement model. Random projections are known to preserve ℓ_2 norms (Euclidean geometries). Since it is only ℓ_2 norms that matter for Gaussian-perturbed measurements, random projections are successful in such scenarios. It is, however, not clear whether such random projections will be effective in reducing dimensionality in non-Gaussian settings and this a interesting problem that requires further study.

We leave open the issue of establishing the relationship between the smallest singular value δ and the minimum spacing between sinusoids, and characterizing the regime in terms of the minimum separation of the frequencies for which the stronger isometry results established for a single sinusoid can be extended to $K > 1$. Another interesting topic for future work is the development of an analytical understanding of multi-dimensional sinusoid estimation, motivated by practical applications such as large 2D arrays for mm-wave communication [46] and imaging.

6.2 Geographic Routing for Mobile Ad Hoc Networks

To the best of our knowledge, this is the first work that provides a *provably* scalable position-publish protocol while providing guarantees on route stretch for the accompanying geographic routing protocol. Key to scalability is a probabilistic approach to updating a *subset* of nodes, and to geographic forwarding with imperfect information. We give new sufficiency results on the scaling of the communication radius needed to ensure that a greedy routing forwarding decision does not deviate too much in angle from the estimate used.

Open Issues

Our emphasis here was on providing analytical insight and design criteria, verified by simulations. Mapping our ideas to practice require detailed protocol specifications at the level of packet level format and processing, and more extensive simulations for a wide variety of mobility models. In addition, while we focus on distant nodes in proving scalability, it may be possible to significantly optimize our protocol as the distance to the destination decreases. Another direction for future investigation is the design of position-publish strategies that account for large holes in the deployment region. It is interesting to note that, while we have assumed a uniform set of protocol parameters for all destination nodes to prove scalability, in practice, each potential destination can choose its parameters differently, depending on the tradeoffs between routing overhead, reliability and stretch that it desires to obtain. While we specify routing efficiency in terms of the length of routing trajectories, an important problem for future investigation is whether provable guarantees can be given for efficiency when we define stretch using number of hops. Such a characterization is useful since wireless resources like energy and spectrum usage can be directly related to the number of hops.

6.3 Inference from time on Twitter

We have demonstrated that significant information about a user's interest can be mined from his/her tweet times alone, by correlating these with the timing of appropriately chosen events in the external world. The Bayesian framework that we develop for extracting this information is shown to be effective in detecting baseball fandom from the tweet times of users over a one month period. Measurements from "neighbors" (in the sense of Twitter mentions) provides additional performance gains, with improvements of about 50% in detection accuracy for a false alarm rate of 1%.

Our results also show that significant information on topic evolution in Twitter can be gleaned from minimalistic methods that use only the time and authorship of tweets. We present a principled approach, grounded in detection and estimation theory, to baseline modeling and activity detection over multiple time scales. By examining tweet time patterns across the pool of users who appear in the feed, we have shown that we can identify meaningful subtopics or event classes. In particular, for the four TV shows we tracked over a period of six months, not only did our analysis identify expected activity during episode air times, but it also uncovered other phenomena such as commercial tie-ins and awards.

Open Issues

Our current implementation of event detection and clustering is based on batch algorithms, and it is of interest to explore streaming implementations. While our event detection algorithm is naturally matched to streaming implementation, an important topic for further investigation is a streaming version of the event embedding and clustering algorithm for dissecting a topic into subtopics. Another topic for future study is whether the topic-level information obtained from aggregate feeds can be employed for detailed user-level inference and whether ideas similar to the ones presented here for event detection can be used to identify the emergence of interesting motifs on social networks.

Appendices

Appendix A

A.1 (Extended) Ziv-Zakai Bound Review[9]

Consider the problem of estimating a parameter $\boldsymbol{\theta}$ from measurements

$$\mathbf{y} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{z},\ \boldsymbol{\theta} \in \Theta, \mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbb{I}). \tag{A.1}$$

For an estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, let $\boldsymbol{\epsilon} = \widehat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}$ denote the estimation error. The ZZB lower bounds the error $\mathbb{E}|\mathbf{a}^T\boldsymbol{\epsilon}|^2$ for any $\mathbf{a} \in \mathbb{R}^K$ by relating it to the probabilities of error in a sequence of detection problems. We begin by describing one of the detection problems.

Consider a simplified version of the preceding model, in which the parameter $\boldsymbol{\theta}$ takes only two values $\boldsymbol{\phi}$ and $\boldsymbol{\phi}+\boldsymbol{\delta}$, occurring with probabilities $p(\boldsymbol{\phi})/(p(\boldsymbol{\phi}) + p(\boldsymbol{\phi}+\boldsymbol{\delta}))$ and $p(\boldsymbol{\phi}+\boldsymbol{\delta})/(p(\boldsymbol{\phi}) + p(\boldsymbol{\phi}+\boldsymbol{\delta}))$, respectively. There are two possible ways to estimate $\boldsymbol{\theta}$:

• *Optimal detection-theoretic approach:* Compute the Bayesian posterior probabilities $p(\boldsymbol{\phi}|\mathbf{y})$ and $p(\boldsymbol{\phi}+\boldsymbol{\delta}|\mathbf{y})$. Choose $\boldsymbol{\phi}$ if $p(\boldsymbol{\phi}|\mathbf{y}) > p(\boldsymbol{\phi}+\boldsymbol{\delta}|\mathbf{y})$ and $\boldsymbol{\phi}+\boldsymbol{\delta}$ otherwise. Denote the probability of error with this approach by $P_{\text{opt}}(\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta})$.

• *Heuristic approach using the estimate* $\widehat{\boldsymbol{\theta}}(\mathbf{y})$*:* Form the estimate $\widehat{\boldsymbol{\theta}}(\mathbf{y})$; this could take any value in Θ, and is not restricted to $\{\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta}\}$. Classify based on the following rule: if $\mathbf{a}^T\widehat{\boldsymbol{\theta}}(\mathbf{y}) < \mathbf{a}^T\boldsymbol{\phi} + (h/2)$, where $h = \mathbf{a}^T\boldsymbol{\delta}$, choose $\boldsymbol{\phi}$ to have occurred; else, choose $\boldsymbol{\phi}+\boldsymbol{\delta}$. Denote the probability of error with this scheme by $P_{\text{nn}}^{\widehat{\boldsymbol{\theta}}}(\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta})$.

Since the Bayesian detection rule is optimal, we have $P_{\text{opt}}(\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta}) \leq P_{\text{nn}}^{\widehat{\boldsymbol{\theta}}}(\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta})$. In order to use this observation to bound $\mathbb{E}|\mathbf{a}^T\boldsymbol{\epsilon}|^2$, we begin with the identity

$$\mathbb{E}\left|\mathbf{a}^T\boldsymbol{\epsilon}\right|^2 = \frac{1}{2}\int_0^\infty \Pr\left(\left|\mathbf{a}^T\boldsymbol{\epsilon}\right| \geq h/2\right) h \; dh, \tag{A.2}$$

and relate $\Pr\left(\left|\mathbf{a}^T\boldsymbol{\epsilon}\right| \geq h/2\right)$ to the probability of error with the heuristic rule $P_{\text{nn}}^{\widehat{\boldsymbol{\theta}}}(\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta})$ as follows:

$$\Pr\left(\left|\mathbf{a}^T\boldsymbol{\epsilon}\right| \geq h/2\right) = \int_{\mathbb{R}^K} \left(p(\boldsymbol{\phi}) + p(\boldsymbol{\phi}+\boldsymbol{\delta})\right) P_{\text{nn}}^{\widehat{\boldsymbol{\theta}}}(\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta}) \; d\boldsymbol{\phi}, \tag{A.3}$$

where $\boldsymbol{\delta}$ is *any* vector satisfying $\mathbf{a}^T\boldsymbol{\delta} = h$. We now use the lower bound $P_{\text{nn}}^{\widehat{\boldsymbol{\theta}}}(\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta}) \geq P_{\text{opt}}(\boldsymbol{\phi}, \boldsymbol{\phi}+\boldsymbol{\delta})$ in (A.3) and substituting back in (A.2), we get the basic version of the ZZB.

We can further tighten the bound in two ways: (a) by choosing $\boldsymbol{\delta}$ appropriately and (b) by exploiting the fact that $\Pr\left(\left|\mathbf{a}^T\boldsymbol{\epsilon}\right| \geq h/2\right)$ is non-increasing using the

valley filling operation $\mathcal{V}\{\ \}$, defined as $\mathcal{V}\{q(h)\} = \max_{r\geq 0} q(h+r)$ (refer [9] for details). This gives us the ZZB in (2.16).

Appendix B

B.1 Proof of Theorem 3.2

Let $\boldsymbol{\omega} = [\omega_1 \ \cdots \ \omega_K]^T$, $\mathbf{g} = [g_1 \ \cdots \ g_K]^T$ and $\mathbf{X}(\boldsymbol{\omega}) = [\mathbf{x}(\omega_1) \ \cdots \ \mathbf{x}(\omega_K)]$. We note that an ϵ-isometry for all vectors of the form $\mathbf{X}(\boldsymbol{\omega})\mathbf{g}$ such that $\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\| > \delta$ and $\|\mathbf{g}\| = 1$ is equivalent to (3.7). We discretize the frequencies $[0, 2\pi]$ uniformly into R points (R is specified later) and obtain the set F. We first prove a $2\epsilon_0$ isometry for all vectors in the span of $\mathbf{X}(\mathbf{q})$ for all frequency tuplets $\mathbf{q} \in F^K$ (i.e., $\mathbf{q} = [q_1 \ \cdots \ q_K]^T$ with $q_l \in F$). We then extend this to a $3.5\epsilon_0$ isometry for vectors $\mathbf{X}(\boldsymbol{\omega})\mathbf{g}$ such that $\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\| > \delta$ and $\|\mathbf{g}\| = 1$ by: (a) approximating them to nearby points in the span of $\mathbf{X}(\mathbf{q})$, (b) choosing $R = O(N^{1.5}K^{0.5}\delta^{-1}\epsilon_0^{-1})$ so that the approximation is good.

Sampling: For any tuplet of sampled frequencies $\mathbf{q} \in F^K$, if $\boldsymbol{A}$ preserves the norm of $\left(6\epsilon_0^{-1}\right)^{2K}$ well-chosen samples in the span of $\mathbf{X}(\mathbf{q})$ up to $\epsilon_0 < 2/5$, it can be shown that $\boldsymbol{A}$ will preserve the norms of all vectors in the span of $\mathbf{X}(\mathbf{q})$ up to

$2\epsilon_0$ [5] (since we are concerned only with ℓ_2 distances from sampled points, we map the unit ball in $\mathbb{C}^K$ to the unit ball in $\mathbb{R}^{2K}$ using the map $f(\mathbf{z}) = [\Re\{\mathbf{z}^T\}\ \Im\{\mathbf{z}^T\}]^T$ and use corresponding covering arguments in [5]. The other argument used in [5], which is closure w.r.t. to addition is satisfied in $\mathbb{C}^K$ as well). Since there are R^K sampled frequency tuplets $\mathbf{q} \in F^K$, by demanding that $\boldsymbol{A}$ preserves the norm of $R^K \left(6\epsilon_0^{-1}\right)^{2K}$ samples, we can provide a $2\epsilon_0$ isometry for the span of $\mathbf{X}(\mathbf{q})\ \forall \mathbf{q} \in F^K$.

Isometry for mixtures of arbitrary frequencies: We now extend this to an $3.5\epsilon_0$ isometry result for vectors of the form $\mathbf{X}(\boldsymbol{\omega})\mathbf{g}$ such that $\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\| > \delta$ and $\|\mathbf{g}\| = 1$ by choosing R appropriately.

Let $\mathbf{q}$ be a tuplet in F^K that is close to $\boldsymbol{\omega}$ satisfying $\max_l |q_l - \omega_l| \leq \pi/R$. We let $\mathbf{e}_l = \mathbf{x}(\omega_l) - \mathbf{x}(q_l)$ and bound the absolute value of each term of $\mathbf{e}_l$ using the mean value theorem to get $\|\mathbf{e}_l\| \leq \pi N/(\sqrt{2}R)$. We use this to calculate a bound on the difference between a vector $\mathbf{X}(\boldsymbol{\omega})\mathbf{g}$ and its approximated version $\mathbf{X}(\mathbf{q})\mathbf{g}$. Using the definition of $\mathbf{e}_l$, we obtain

$$\mathbf{X}(\boldsymbol{\omega})\mathbf{g} = \mathbf{X}(\mathbf{q})\mathbf{g} + \sum_{l=1}^{l=K} g_l \mathbf{e}_l. \tag{B.1}$$

Using the triangle inequality and the fact that $\sum_l |g_l| \leq \sqrt{2K}$ (since $\|\mathbf{g}\| = 1$), we have

$$\frac{\|\mathbf{X}(\mathbf{q})\mathbf{g}\|}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|} \lessdot 1 \pm \frac{\pi\sqrt{K}NR^{-1}}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}. \tag{B.2}$$

where $x \lhd y \pm z$ denotes $y - z \leq x \leq y + z$.

Next, we bound the difference between the vectors $\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}$ and $\boldsymbol{A}\mathbf{X}(\mathbf{q})\mathbf{g}$. We see that $\|\boldsymbol{A}\|_F = \sqrt{M}$ and, therefore, have $\|\boldsymbol{A}\mathbf{e}_k\| \leq \sqrt{M}\,\|\mathbf{e}_k\|$. Furthermore, since $\boldsymbol{A}$ preserves the norms of all vectors of the form $\mathbf{X}(\mathbf{q})\mathbf{g}$, where $\mathbf{q} \in F^K$ up to an isometry constant $2\epsilon_0$ (and scale factor of $\sqrt{M/N}$), we get

$$\sqrt{\frac{N}{M}}\frac{\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}{\|\mathbf{X}(\mathbf{q})\mathbf{g}\|} \lhd 1 \pm \left(2\epsilon_0 + \frac{\pi N\sqrt{NK}R^{-1}}{\|\mathbf{X}(\mathbf{q})\mathbf{g}\|}\right). \tag{B.3}$$

Before we proceed to give the isometry result, we need to characterize how small $\|\mathbf{X}(\mathbf{q})\mathbf{g}\|$ can be in (B.3). Since $\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\| > \delta$, from (B.2) we have the following:

$$\frac{\|\mathbf{X}(\mathbf{q})\mathbf{g}\|}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|} \lhd 1 \pm \pi\sqrt{K}N(R\delta)^{-1}. \tag{B.4}$$

Choosing $R = (4\pi)N\sqrt{NK}\epsilon_0^{-1}\delta^{-1}$, we have that

$$\frac{\|\mathbf{X}(\mathbf{q})\mathbf{g}\|}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|} \lhd 1 \pm 0.25\epsilon_0. \tag{B.5}$$

For this choice of R, from (B.3), we see that

$$\sqrt{\frac{N}{M}}\frac{\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}{\|\mathbf{X}(\mathbf{q})\mathbf{g}\|} \lhd 1 \pm \left(2\epsilon_0 + \frac{0.25\epsilon_0\delta}{\|\mathbf{X}(\mathbf{q})\mathbf{g}\|}\right). \tag{B.6}$$

Using the lower bound from (B.5), $\|\mathbf{X}(\mathbf{q})\mathbf{g}\| \geq (1 - 0.25\epsilon_0)\,\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\| \geq \delta(1 - 0.25\epsilon_0)$,

$$\sqrt{\frac{N}{M}}\frac{\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}{\|\mathbf{X}(\mathbf{q})\mathbf{g}\|} \lhd 1 \pm 2.5\epsilon_0. \tag{B.7}$$

Substituting the bounds for $\|\mathbf{X}(\mathbf{q})\mathbf{g}\|$ in terms of $\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|$ from (B.5), we have that

$$\sqrt{\frac{N}{M}}\frac{\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|} \lhd 1 \pm 3.5\epsilon_0. \tag{B.8}$$

Number of measurements: It only remains to specify the number of measurements M required to preserve the norms of the $R^K \left(6\epsilon_0^{-1}\right)^{2K}$ samples up to ϵ_0. Using the value for R just obtained, and setting $\epsilon = 3.5\epsilon_0$, we see that we must preserve the norms of $(18 \times 7^3 \pi N^{1.5} K^{0.5} \epsilon^{-3} \delta^{-1})^K$ vectors (samples) up to $2\epsilon/7$ w.h.p. We relate the probability of preserving these norms to the number of measurements M via the concentration results (2.8) for Uniform$\{\pm 1/\sqrt{N}, \pm j/\sqrt{N}\}$ (setting δ in (2.8) and (2.9) to $32\epsilon/49$ – here we have used the fact that when $\epsilon < 1$, $\max\{(1+2\epsilon/7)^2 - 1, 1-(1-2\epsilon/7)^2\} < 32\epsilon/49$). We employ the union bound and (2.8) to compute the probability that the norm of at least one sample is not preserved. This probability becomes vanishingly small for $M = O\left(\epsilon^{-2} K \log\left(NK\epsilon^{-1}\right.\right.$ $\left.\left.\delta^{-1}\right)\right)$ measurements, which concludes the proof.

B.2 Proof of Theorem 3.1

For the matrix $\mathbf{T}(\boldsymbol{\omega})$, K of the columns are of the form $\tau d\mathbf{x}(\omega)/d\omega$, while the remaining K are of the form $\mathbf{x}(\omega)$. When $\tau d\mathbf{x}(\omega)/d\omega$ is approximated by $\tau d\mathbf{x}(q)/d\omega$, where q is the frequency on an uniformly spaced frequency grid with

R points that is the closest to ω, the norm of the approximation error is upper bounded by $\pi N/(\sqrt{2}R)$. The upper bound on the norm of the error in approximating $\mathbf{x}(\omega)$ by $\mathbf{x}(q)$ used in theorem 3.2 is also $\pi N/(\sqrt{2}R)$. Therefore, by following the proof of theorem 3.2 with K set to $2K$ (because number of columns of $\mathbf{X}(\boldsymbol{\omega})$ is only K), we obtain the proof for theorem 3.1.

B.3 Proof of Theorem 3.3

We present the results for closely spaced frequencies first (tangent plane isometries), and then move to the well-separated setting.

Tangent plane isometry: For a single sinusoid, the tangent plane matrix at ω is given by $\mathbf{T}(\omega) = [\mathbf{x}(\omega)\ \tau d\mathbf{x}(\omega)/d\omega]$ where $\tau = 1/\|d\mathbf{x}(\omega)/d\omega\|$. The smallest singular value of $\mathbf{T}(\omega)$, denoted by σ_{tangent}, satisfies

$$\begin{aligned} \sigma^2_{\text{tangent}} &= 1 - \tau\,|\langle \mathbf{x}(\omega), d\mathbf{x}(\omega)/d\omega\rangle| && \text{(B.9)} \\ &= 1 - \tau \left| \sum_{n=1}^{n=N} |h_n|^2 j\omega\,(n - (N+1/2)) \right| && \text{(B.10)} \end{aligned}$$

where the second equality is obtained from the definition of the sinusoid (3.1) by noting that the nth entry of $\mathbf{x}(\omega)$ is $h_n e^{j\omega(n-(N+1)/2)}$. From the definition of $H(\omega)$, we see that,

$$\sigma^2_{\text{tangent}} = 1 - \tau\,|dH(0)/d\omega|\,, \qquad \text{(B.11)}$$

and therefore $\sigma_{\text{tangent}} = \sqrt{1-\tau\chi}$ where $\chi = |dH(0)/d\omega|$. By Jensen's inequality, we see that $\chi^2 < 1/\tau^2$ when the weight sequence $\{h_n\}$ has more than one non-zero tap. Thus, $\tau\chi < 1$ and therefore σ_{tangent} is strictly positive. Setting $\delta = \sqrt{1-\tau\chi}$ in Theorem 3.1, we can provide tangent plane ϵ-isometries for a single sinusoid with $M = O\left(\epsilon^{-2}\log\left(N\epsilon^{-1}(1-\tau\chi)^{-1}\right)\right)$ measurements.

Extending tangent plane isometry to pairwise isometry for frequencies separated by at most $1/N^{1.5}$: We now extend ϵ-isometry of the tangent planes to a pairwise 2ϵ-isometry for any two frequencies ω_1, ω_2 whose separation $\Delta = \omega_2 - \omega_1$ is "small" (we quantify how small later) by exploiting continuity. Let $q = (\omega_1 + \omega_2)/2$ be the average of the two frequencies. For small values of $|\Delta|$, a first-order Taylor series expansion for $\mathbf{x}(\omega_1)$ and $\mathbf{x}(\omega_2)$ around $\mathbf{x}(q)$ will have small errors. Such an expansion gives us

$$\mathbf{x}(\omega_1) = \mathbf{x}(q) - (\Delta/2)(d\mathbf{x}(q)/d\omega) + \mathbf{e}_1, \tag{B.12}$$

$$\mathbf{x}(\omega_2) = \mathbf{x}(q) + (\Delta/2)(d\mathbf{x}(q)/d\omega) + \mathbf{e}_2, \tag{B.13}$$

where $\mathbf{e}_1, \mathbf{e}_2$ are the approximation errors. Consider a linear combination $\mathbf{X}(\boldsymbol{\omega})\mathbf{g}$ where $\mathbf{X}(\omega) = [\mathbf{x}(\omega_1)\ \mathbf{x}(\omega_2)]$ and $\mathbf{g} = [g_1\ g_2]$. This can be written as

$$\mathbf{X}(\boldsymbol{\omega})\mathbf{g} = \mathbf{v} + \mathbf{e} \tag{B.14}$$

where $\mathbf{e} = g_1\mathbf{e}_1 + g_2\mathbf{e}_2$ and

$$\mathbf{v} = (g_1 + g_2)\,\mathbf{x}(q) + (\Delta/2)\,(g_2 - g_1)\,(d\mathbf{x}(q)/d\omega), \tag{B.15}$$

lies in the span of $\mathbf{T}(q) = [\mathbf{x}(q)\ \tau(d\mathbf{x}(q)/d\omega)]$, the tangent plane at $\omega = q$.

Since $\boldsymbol{A}$ guarantees ϵ-isometries for tangent planes at all frequencies, for any vector $\mathbf{T}(q)\mathbf{h}$ in the tangent plane at q, the quantity $\|\boldsymbol{A}\mathbf{T}(q)\mathbf{h}\|$ is bounded within $(1 \pm \epsilon)\sqrt{M/N}\|\mathbf{T}(q)\mathbf{h}\|$. Expanding out $\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|/\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|$ in terms of $\mathbf{v}$ and $\mathbf{e}$ and applying the tangent plane isometry condition to $\|\boldsymbol{A}\mathbf{v}\|/\|\mathbf{v}\|$, we can show that

$$\sqrt{\frac{N}{M}}\frac{\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|} \lhd 1 \ \pm \ \left(\epsilon + \frac{5\sqrt{N}\|\mathbf{e}\|}{\|\mathbf{v}\|}\right). \tag{B.16}$$

where $x \lhd y \pm z$ denotes $y - z \leq x \leq y + z$. Next, we get bounds on $\|\mathbf{e}\|$ and $\|\mathbf{v}\|$ as follows. First, we use the mean value theorem to show that the error is bounded as $\|\mathbf{e}\| \leq N^2\Delta^2/(4\sqrt{2})$. Next, since $\mathbf{v}$ lies in the span of $\mathbf{T}(q)$, we can use the bound on the minimum singular value of $\mathbf{T}(q)$ to get $\|\mathbf{v}\| \geq \sqrt{1 - \tau\chi}|\Delta|/(\sqrt{2}\tau)$. The details are given in Appendix B.4. Substituting these bounds in the above equation, we obtain

$$\sqrt{\frac{N}{M}}\frac{\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|} \lhd 1 \ \pm \ \left(\epsilon + \frac{5\tau|\Delta|N^{2.5}}{4\sqrt{1 - \tau\chi}}\right). \tag{B.17}$$

We note that $\tau = 1/\|d\mathbf{x}(\omega)/d\omega\|$ scales as $1/N$. Therefore, defining a scale-invariant constant $\alpha = 1/(N\tau)$, we see that, as long as the frequency separation $|\Delta| \leq (4\alpha\epsilon\sqrt{(1 - \tau\chi)}/5)/N^{1.5}$, we can get a 2ϵ isometry

$$\sqrt{\frac{N}{M}}\frac{\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|}{\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|} \lhd 1 \ \pm \ 2\epsilon. \tag{B.18}$$

Thus, if $\boldsymbol{A}$ provides an $\epsilon/2$ tangent plane isometry for all frequencies (which can be achieved with $M = O\left(\epsilon^{-2}\log\left(N\epsilon^{-1}(1-\tau\chi)^{-1}\right)\right)$ measurements), we can extend it to a pairwise ϵ-isometry for the set of frequencies ω_1, ω_2 whose separation $|\omega_1 - \omega_2| \leq (4\alpha(\epsilon/2)\sqrt{(1-\tau\chi)}/5)/N^{1.5}$.

Pairwise isometry for frequencies separated by more than $1/N^{1.5}$: We now use Theorem 3.2 to quantify the number of measurements necessary to guarantee pairwise ϵ-isometry for two frequencies that are separated by more than $\mu/N^{1.5}$, where $\mu = (4\alpha(\epsilon/2)\sqrt{(1-\tau\chi)}/5)$.

First, we obtain a bound on the smallest singular value of $\mathbf{X}(\omega_1, \omega_2) = [\mathbf{x}(\omega_1)\ \mathbf{x}(\omega_2)]$. Denoting the smallest singular value by σ^2_{signal}, we can show that it satisfies

$$\sigma^2_{\text{signal}} = 1 - |\langle \mathbf{x}(\omega_1), \mathbf{x}(\omega_2)\rangle|\,. \tag{B.19}$$

Furthermore, we can show that $|\langle \mathbf{x}(\omega_1), \mathbf{x}(\omega_2)\rangle| = |H(\omega_1 - \omega_2)|$, where $H(\omega) = \sum_{n=1}^{n=N} |h_n|^2 e^{j\omega(n-(N+1)/2)}$. Thus, we have $\sigma^2_{\text{signal}} = 1 - |H(\omega_1 - \omega_2)|$.

Suppose now that $|\omega_1 - \omega_2| > \mu/N^{1.5}$. For large values of N, the smallest singular value of $\mathbf{X}(\omega_1, \omega_2)$ is bounded as

$$\sigma_{\text{signal}} > \sqrt{\frac{0.4\zeta\mu^2}{N}},\ \text{where}\ \zeta = -\frac{N^{-2}}{2!}\left.\frac{d^2|H(\omega)|^2}{d\omega^2}\right|_{\omega=0}. \tag{B.20}$$

The details are given in Appendix B.5.

We now apply Theorem 3.2 with $\delta = \sqrt{0.4\zeta\mu^2/N}$. The set of all frequencies $|\omega_1 - \omega_2| > \mu/N^{1.5}$ is contained in $\Lambda_p(\sqrt{0.4\zeta\mu^2/N})$ and thus, we can guarantee

pairwise ϵ-isometry for this set with $M = O\left(\epsilon^{-2}\log\left(N\epsilon^{-1}(1-\tau\chi)^{-1}\zeta^{-1}\alpha^{-1}\right)\right)$ measurements.

Combining the isometries in the regimes $|\omega_1 - \omega_2| \leq \mu/N^{1.5}$ and $|\omega_1 - \omega_2| \geq \mu/N^{1.5}$ completes the proof of Theorem 3.3.

B.4 Extending tangent plane isometry

We first derive a bound on $\|\mathbf{e}\|$. Applying the triangle inequality to $\mathbf{e}$, we obtain $\|\mathbf{e}\| \leq |g_1|\|\mathbf{e}_1\| + |g_2|\|\mathbf{e}_2\|$. Since the quantity we wish to bound $\|\boldsymbol{A}\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|/\|\mathbf{X}(\boldsymbol{\omega})\mathbf{g}\|$ does not depend on $\|\mathbf{g}\|$, we can, without loss of generality, restrict attention to $\|\mathbf{g}\| = 1$. Thus, we have $|g_i| \leq 1$. We use the mean value theorem to obtain bounds on $\|\mathbf{e}_i\|$ $i = 1, 2$ (the mean value theorem relates $\mathbf{e}_i$ to $d^2\mathbf{x}(\omega_i')/d\omega^2$ for some $\omega_i' \in [\omega_1, \omega_2]$) and ultimately get $\|\mathbf{e}\| \leq N^2\Delta^2/(4\sqrt{2})$.

In order to obtain a lower bound for $\|\mathbf{v}\|$, we rewrite $\mathbf{v}$ as

$$\mathbf{v} = \mathbf{T}(q) \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \frac{\Delta}{\sqrt{2}\tau} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}. \tag{B.21}$$

We now recall that the minimum singular value of the product of two matrices is at least as large as the product of their minimum singular values. The minimum singular value of $\mathbf{T}(q)$ is $\sigma_{\text{tangent}} = \sqrt{1-\tau\chi}$ and the corresponding value for the other two matrices are $|\Delta|/\sqrt{2}\tau$ and 1 respectively. Thus, the minimum singular

value of the product of the three matrices is greater than $\sqrt{1-\tau\chi} \times \frac{|\Delta|}{\sqrt{2}\tau}$. Since $\|\mathbf{g}\|= 1$, we immediately get the desired bound on $\|\mathbf{v}\|$.

B.5 Smallest singular value for well-separated frequencies

We wish to obtain a lower bound for the smallest singular value σ^2_{signal} of the matrix $[\mathbf{X}(\omega_1)\ \mathbf{X}(\omega_2)]$ when the frequencies satisfy $|\omega_1 - \omega_2| > \mu/N^{1.5}$. First, we note that this is equivalent to upper-bounding $|H(\omega_1 - \omega_2)|$ since $\sigma^2_{\text{signal}} = 1 - |H(\omega_1 - \omega_2)|$. Since $|H(\omega)|$ is not necessarily monotonic (imagine that $|h_n|^2$ is the Hamming window; $|H(\omega)|$, being the magnitude of the Fourier transform of $|h_n|^2$, has sidelobes), it is not true in general that that the maximum of $|H(\omega)|, |\omega| > \mu/N^{1.5}$ occurs at $\omega = \mu/N^{1.5}$. However, we now make two observations that allow us to analyze the behavior of $|H(\omega)|$ only at the minimum separation $\mu/N^{1.5}$.

First, if there were no restrictions on the frequencies (ω_1, ω_2), $|H(\omega_1 - \omega_2)|$ has a maximum $(= 1)$ when $\omega_1 = \omega_2$. Second, because (i) the set of the frequencies we are excluding $|\omega_1 - \omega_2| < \mu/N^{1.5}$ is very small (it is smaller than $\pi/(2N)$ for large enough N) and (ii) we restrict ourselves to sequences $\{h_n\}$ such that $|H(\omega)|$

is monotone in $(0, \pi/(2N))$, the maximum of $|H(\omega_1 - \omega_2)|, \pi/(2N) > |\omega_1 - \omega_2| > \mu/N^{1.5}$ is guaranteed to occur when $\omega_2 = \omega_1 \pm \mu/N^{1.5}$.

For small values of $|\omega_1 - \omega_2|$, we can expand $|H(\omega)|^2$ around $\omega = 0$ to get

$$|H(\omega_1 - \omega_2)|^2 = 1 - \zeta N^2 (\omega_1 - \omega_2)^2 \pm O\left(N^4 (\omega_1 - \omega_2)^4\right), \tag{B.22}$$

where $\zeta = -(N^{-2}/2!)\, d^2|H(\omega)|^2/d\omega^2|_{\omega=0}$. For $|\omega_1 - \omega_2| = \mu/N^{1.5}$, we have

$$\left|H\left(\mu/N^{1.5}\right)\right|^2 = 1 - \zeta\mu^2/N \pm O\left(1/N^2\right). \tag{B.23}$$

we see that $|H(\mu/N^{1.5})|$ approaches 1 with increasing N. Since we assume that all side-lobes are smaller than $D < 1$, there exists some N beyond which the maximum of $|H(\omega_1 - \omega_2)|, |\omega_1 - \omega_2| > \pi/(2N)$ is guaranteed to be smaller than $|H(\mu/N^{1.5})|$. Therefore, for sufficiently large N, the maximum of $|H(\omega_1 - \omega_2)|$ for all $|\omega_1 - \omega_2| > \mu/N^{1.5}$ occurs at $|\omega_1 - \omega_2| = \mu/N^{1.5}$. Plugging the expression for $|H(\omega_1 - \omega_2)|$ in σ^2_{signal} for this frequency separation, we have that,

$$N\sigma^2_{\text{signal}} \geq 0.5\zeta\mu^2 \pm O\left(1/N\right). \tag{B.24}$$

To arrive at the above expression we have used the following: $|H(\omega)| \leq 1\ \forall\omega$ (since $\sum|h_n|^2 = 1$). This gives us

$$1 - |H(\omega)| \geq (1 + |H(\omega)|)(1 - |H(\omega)|)/2\ \forall\omega. \tag{B.25}$$

Therefore, $\sigma^2_{\text{signal}} \geq (1 - |H(\mu/N^{1.5})|^2)/2$, which yields (B.24). The first term in (B.24) given by $0.5\zeta\mu^2$, is a constant and the second term decays to zero (as $1/N$).

Therefore, for large values of N, the second term is much smaller than the first and σ^2_{signal} is bounded away from zero. In particular, for large enough N (how large it needs to be depends on μ and the behavior of $|H(\omega)|^2$ at $\omega = 0$), we have $\sigma_{\text{signal}} > \sqrt{0.4\zeta\mu^2/N}$.

Appendix C

C.1 Proof of Theorem 4.1

We start by laying out the notation used in this proof. Every node in the network can forward packets to any node that is within the communication radius $r(n)$. We refer to the circle of radius z centered around a point $\mathbf{u}$ by $C(\mathbf{u}, z)$. We reserve the symbol $r = r(n)$ for the communication radius and $\mathbf{v}$ for the location of the network node (relay) under consideration. Therefore, the node $\mathbf{v}$ can forward a packet to any node inside the circle $C(\mathbf{v}, r)$.

Around each node $\mathbf{v}$, we choose $\lceil 2\pi/\delta \rceil$ "anchor" regions $A_{\mathbf{v}}^k$ of the following form: (i) $A_{\mathbf{v}}^k = C(\mathbf{v}, r) \cap C(\mathbf{u}_k, r)$, where $\mathbf{u}_k$ is a point on the perimeter of the circle $C(\mathbf{v}, 2r\cos(\delta/2))$, where δ is the constant in the statement of Theorem 4.1 (satisfying $0 < \delta \leq \pi/3$) (ii) The $\lceil 2\pi/\delta \rceil$ points $\{\mathbf{u}_k\}$ are carefully chosen so that any ray drawn outward from this node $\mathbf{v}$ intersects at least one anchor region $A_{\mathbf{v}}^k$.

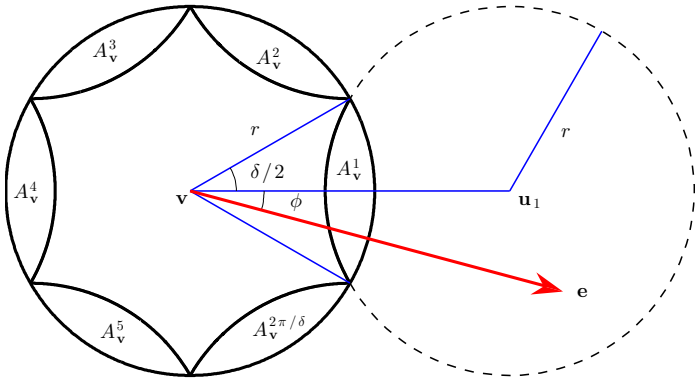

Figure C.1: Anchor regions $A^k_{\mathbf{v}}$ numbering $\lceil 2\pi/\delta \rceil$ in the neighborhood (distances smaller than the communication radius r) around a typical node $\mathbf{v}$. ϕ is the direction of the estimate $\mathbf{e}$ towards which a packet is being greedily routed by $\mathbf{v}$.

Figure C.1 illustrates one such choice of anchor regions around the node $\mathbf{v}$ for $\delta = \pi/3$ ($2\pi/\delta$ chosen to be an integer for the sake of convenience).

We now provide a summary of the proof strategy used. Using an union bound, we first show that when we choose a large enough communication radius ($r^2 \geq c \log n$, for an appropriate choice of constant c), then w.h.p., every anchor region in the network is occupied by at least one node. We follow this up with the implications of the occupancy of the $\lceil 2\pi/\delta \rceil$ anchor regions $\{A^k_{\mathbf{v}}\}$ around a node $\mathbf{v}$ for greedy forwarding decisions made by this node.

Asymptotic occupancy guarantee for anchor regions

Let $E_{\mathbf{v}}^k$ denote the "error event" that all $n-1$ network nodes (other than $\mathbf{v}$) reside outside the anchor region $A_{\mathbf{v}}^k$. In other words, $E_{\mathbf{v}}^k$ denotes the event that the anchor region $A_{\mathbf{v}}^k$ is unoccupied. All nodes are uniformly and independently distributed over the deployment region of area n. Therefore, denoting the area of the region P by $|P|$, we have that

$$\Pr\left[E_{\mathbf{v}}^k\right] = \left(1 - \left|A_{\mathbf{v}}^k\right|/n\right)^{n-1}.$$

We refer to the event that one of the $n\lceil 2\pi/\delta\rceil$ anchor regions being empty as the "cumulative error event" and denote it by E_{all}. Since $E_{\text{all}} = \bigcup_{\mathbf{v},k} E_{\mathbf{v}}^k$, we use the union bound to arrive at:

$$\Pr\left[E_{\text{all}}\right] \leq \sum_{\mathbf{v},k} \Pr\left[E_{\mathbf{v}}^k\right] = \left\lceil \frac{2\pi}{\delta} \right\rceil n \left(1 - \frac{(\delta - \sin\delta)\, r^2}{n}\right)^{n-1},$$

where we have used the fact that $\left|A_{\mathbf{v}}^k\right| = (\delta - \sin\delta)\, r^2$. Let $\nu > 0$ denote a constant. We note that when

$$r^2 = (1+\nu)\log n/(\delta - \sin\delta)\,, \tag{C.1}$$

$\Pr\left[E_{\text{all}}\right] \to 0$ as n grows. Thus, for large n, if r satisfies (C.1), all anchor regions are occupied w.h.p.

Implications of anchor region occupancy for greedy forwarding

Consider a packet at the node $\mathbf{v}$, which is being forwarded towards an estimate $\mathbf{e}$ (need not correspond to the location of any of the n nodes in the network) which is no closer to $\mathbf{v}$ than $2r$. All $\lceil 2\pi/\delta \rceil$ anchor regions $\{A_{\mathbf{v}}^k\}$ around $\mathbf{v}$ are occupied by at least one node. We denote the line segment joining the points $\mathbf{u}$ and $\mathbf{v}$ by $L(\mathbf{u}, \mathbf{v})$. w.l.o.g., we assume that $L(\mathbf{v}, \mathbf{e})$ intersects the anchor region $A_{\mathbf{v}}^1$. Let ϕ denote the angle between $L(\mathbf{v}, \mathbf{e})$ and $L(\mathbf{v}, \mathbf{u}_1)$, where $\mathbf{u}_1$ is the point used in the construction of the anchor region $A_{\mathbf{v}}^1$ (this is depicted in Figure C.1). We now provide the answer to the question: What does the occupancy of the region $A_{\mathbf{v}}^1$ mean for the quality of the greedy forwarding decision taken by $\mathbf{v}$ corresponding to this estimate $\mathbf{e}$? More specifically, denoting the next hop (the node to which this packet is forwarded to) by $\mathbf{x}$, we ask, what is largest absolute value that the angle between $L(\mathbf{v}, \mathbf{e})$ and $L(\mathbf{v}, \mathbf{x})$ can take?

Consider the tangent to the inner boundary of the anchor region $A_{\mathbf{v}}^1$ which is perpendicular to the line segment $L(\mathbf{v}, \mathbf{e})$ (tangent to $C(\mathbf{u}_1, r)$; marked in Figure C.2). Let 2Φ denote the angle subtended by the acute sector S of $C(\mathbf{v}, r)$ associated with this tangent (we mark this sector in Figure C.2 using the $\square$ sym-

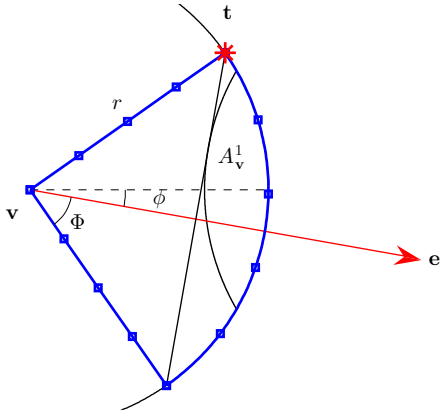

Figure C.2: Sector S is marked using the $\square$ symbol. The point $\mathbf{t}$ in $C(\mathbf{v}, r) \setminus S$ closest to the estimate $\mathbf{e}$ is marked using the $*$ symbol.

bol). It can be shown that $\cos\Phi = 2\cos(\delta/2)\cos\phi - 1$. Since the line $L(\mathbf{v}, \mathbf{e})$ intersects this anchor region $A^1_{\mathbf{v}}$, we have that $\phi \leq \delta/2$, and as a result $\Phi \leq \delta$.

Consider any point $\mathbf{u} \in C(\mathbf{v}, r) \setminus S$. When $\delta \leq \pi/3$ and $\ell(\mathbf{v}, \mathbf{e}) > 2r$, it can be shown that: (i) $\ell(\mathbf{u}, \mathbf{e}) \geq \ell(\mathbf{t}, \mathbf{e})$ (where $\mathbf{t} \in C(\mathbf{v}, r) \setminus S$ is the point marked in Figure C.2 using the $*$ symbol) (ii) $C(\mathbf{e}, \ell(\mathbf{e}, \mathbf{t}))$ encompasses the anchor region $A^1_{\mathbf{v}}$ completely. Therefore, every point inside the anchor region $A^1_{\mathbf{v}}$ is closer to $\mathbf{e}$ than any point in $C(\mathbf{v}, r) \setminus S$ (i.e., $A^1_{\mathbf{v}} \subset C(\mathbf{e}, \ell(\mathbf{e}, \mathbf{t})) \subseteq C(\mathbf{e}, \ell(\mathbf{e}, \mathbf{u}))$).

We now show (by contradiction) that when $\delta < \pi/3$, $\ell(\mathbf{v}, \mathbf{e}) > 2r$, the next hop $\mathbf{x}$ lies within this sector S of width 2Φ. Now suppose that the next hop $\mathbf{x} \in C(\mathbf{v}, r) \setminus S$. Our preceding discussions imply that $A^1_{\mathbf{v}} \subset C(\mathbf{e}, \ell(\mathbf{e}, \mathbf{x}))$. Since greedy geographic forwarding always chooses the neighbor closest to the destination, we infer that the anchor region $A^1_{\mathbf{v}}$ is not occupied. But this contradicts the

assumption that $A^1_{\mathbf{v}}$ is occupied by at least one node. Thus, the next hop $\mathbf{x}$ must lie inside the sector S. The absolute angle between direction along which the packet is forwarded and the direction of the estimate using which it is forwarded is therefore bounded by $\Phi \leq \delta$.

Summary

We have shown that: (i) when the communication radius $r(n)$ scales as (C.1), then w.h.p., all anchor regions around every node are occupied. This corresponds to choosing ϵ in (4.1) so that (4.2) holds (ii) When the $\lceil 2\pi/\delta \rceil$ anchor regions around a node $\mathbf{v}$ are occupied, the absolute angle between the direction along which the packet is forwarded $L(\mathbf{v}, \mathbf{x})$ and the direction of the position estimate $L(\mathbf{v}, \mathbf{e})$ using which this forwarding decision is made can be no larger than δ (for any $\delta < \pi/3$ and $\mathbf{e}$ such that $\ell(\mathbf{v}, \mathbf{e}) > 2r$), thus proving Theorem 4.1. □

C.1.1 Proof of Corollary 4.2

Consider the scenario where the uncertainty of all position estimates is bounded by $U_{\max}$. This ensures that the angle between the estimated direction and the true direction is no greater than $\arcsin U_{\max}$. If the communication radius r is large enough, we infer from Theorem 4.1 that greedy forwarding decisions re-

sult in angular distortions smaller than δ (as long as the relay-estimate distance is no smaller than $2r$). Therefore, the angle between the direction along which the packet is forwarded and the true direction is bounded by $\arcsin(U_{\max}) + \delta$ and progress towards the destination per unit distance travelled is no lesser than $\cos(\arcsin(U_{\max}) + \delta)$. Therefore, every hop reduces the packet-destination distance until the relay-estimate distance is smaller than $2r$. From the definition of uncertainty, we have that when the relay-estimate distance is smaller than $2r$, the relay-destination distance is smaller than $2(1 - U_{\max})^{-1} r$. Therefore, location information is exact and routing is guaranteed to be successful thereafter (when ϵ satisfies (4.2) for some $\delta \leq \pi/3$, $\epsilon > \epsilon_0 \approx 1.6$ needed for successful routing with exact location information [62]).

C.2 Probability of missing an update ring

The density of nodes throughout the network is a uniform 1 node per unit area. This remains invariant under our Brownian motion model. For $r_i \leq a \leq r_i + d_i$, let us denote by $\Lambda_U(a, t)$ the "update density", or the density of the subset of nodes in the ring with active updates, where t is the time elapsed since update issue and a is the distance of from the center of the update ring (the update density is circularly symmetric and thus is a function of only the distance from

the center of the ring). At $t = 0$, all nodes in the ring have active updates, so that $\Lambda_U(a, 0) = I_{[r_i, r_i + d_i]}(a)$, where I_B denotes the indicator function of a set B. As time proceeds, the positions of the nodes with active updates is smeared out by the Gaussian kernel induced by 2D Brownian motion, so that

$$\Lambda_U(a, t) = \Lambda_U(a, 0) \otimes \mathcal{N}\left(\mathbf{0}, \sigma^2 t\, \mathbb{I}_2\right) \tag{C.2}$$

where $\otimes$ stands for 2D convolution. Let $\Lambda^\star(a) = \Lambda_U(a, T_i)$ be the worst case update density (just before the update expires). When a packet meets a node at a distance a from the center of the i-th update ring, the probability that it does not get an active update is therefore at most $1 - \Lambda^\star(a)$.

The worst case packet traversal for missing an update ring is given by a radial cut through, and for this trajectory the packet meets at least d_i/r nodes, and a miss occurs if none of these have an active update (we wish to reiterate that meeting fewer nodes inside an update region increases the chance that the packet misses this update and that the number of nodes via which a packet is relayed can be no lesser than d_i/r). We therefore obtain that the miss probability for the i-th ring satisfies

$$P_{\text{miss}}(i) \leq \prod_{\ell=1}^{\ell = d_i/r} \left(1 - \Lambda^\star\left(r_i + \ell r\right)\right). \tag{C.3}$$

Taking logarithms, the product becomes a sum which we then approximate as an integral using $r/d_i \to 0$ for i large.

$$\begin{aligned}\log P_{\text{miss}}(i) &\leq \sum_{l=1}^{l=d_i/r} \log\left(1-\Lambda^{\star}\left(r_i+\ell r\right)\right)\\ &\leq -\sum_{l=1}^{l=d_i/r} \Lambda^{\star}\left(r_i+lr\right)\\ &\approx -\frac{1}{r}\int_{r_i}^{r_i+d_i} \Lambda^{\star}\left(a\right)\,\mathrm{d}a. \end{aligned} \tag{C.4}$$

The worst case update density for the i-th ring $\Lambda^{\star}(a)$ (a is the distance from the center of the update ring) is the density just before the timer T_i corresponding to the update elapses and from (C.2), we have that

$$\Lambda^{\star}\left(a\right) = \frac{1}{\sigma^2 T_i}\int_{r_i}^{r_i+d_i} \rho \exp\left(\frac{-a^2-\rho^2}{2\sigma^2 T_i}\right) I_0\left(\frac{a\rho}{\sigma^2 T_i}\right)\mathrm{d}\rho, \tag{C.5}$$

where $I_0\left(\cdot\right)$ denotes the zeroth order modified Bessel function of the first kind. The probability of missing the the i-th update ring satisfies (C.4). Using (C.5) in (C.4),

$$\log P_{\text{miss}}(i) \lessapprox \frac{-1}{\sigma^2 r T_i}\iint_{\mathcal{A}_i} \rho \exp\left(\frac{-a^2-\rho^2}{2\sigma^2 T_i}\right) I_0\left(\frac{a\rho}{\sigma^2 T_i}\right)\mathrm{d}a\;\mathrm{d}\rho,$$

where $\mathcal{A}_i$ is given by $r_i \leq a, \rho \leq (r_i + d_i)$. For outer rings $\min_{\mathcal{A}_i}(a\rho) = r_i^2 \gg \sigma^2 T_i$ (since $\gamma < 2$; required to limit abnormal updates). Therefore, the argument of $I_0(\cdot)$ is large and we can use the approximation $I_0(t) \geq \exp(t)/\sqrt{2\pi t}$ (which is

valid for large t) to arrive at

$$\log P_{\text{miss}}(i) \lessapprox \frac{-\sigma\sqrt{T_i}}{r\sqrt{2\pi}} \iint_{\mathcal{B}_i} \sqrt{y/x} \exp\left(-\frac{(x-y)^2}{2}\right) \mathrm{d}x \ \mathrm{d}y,$$

where we have set $x = a/\sqrt{\sigma^2 T_i}$, $y = \rho/\sqrt{\sigma^2 T_i}$ and $\mathcal{B}_i$ is given by $r_i/\sqrt{\sigma^2 T_i} \leq x, y \leq (r_i + d_i)/\sqrt{\sigma^2 T_i}$. We note that $\min_{\mathcal{B}_i} \sqrt{y/x} = \sqrt{r_i/(r_i + d_i)}$. Since r_i scales faster than d_i, this is well-approximated by 1 for outer rings. Using this approximation, we arrive at

$$\begin{aligned}
\log P_{\text{miss}}(i) &\lessapprox \frac{-\sigma\sqrt{T_i}}{r\sqrt{2\pi}} \iint_{\mathcal{B}_i} \exp\left(-(x-y)^2/2\right) \mathrm{d}x \ \mathrm{d}y \\
&= \frac{-\sigma\sqrt{T_i}}{r} \times \int_0^{d_i^{\text{eff}}} \left(1 - 2Q\left(x\right)\right) \ \mathrm{d}x \\
&\approx \frac{-\sigma\sqrt{T_i}}{r\sqrt{2\pi}} \times \left(d_i^{\text{eff}}\right)^2.
\end{aligned}$$

where $Q(x) = \left(1/\sqrt{2\pi}\right) \int_x^\infty e^{-t^2/2} \ \mathrm{d}t$ is the CCDF of the standard normal distribution and d_i^{eff} denotes $d_i/\left(\sigma\sqrt{T_i}\right)$. The above approximation is accurate when d_i^{eff} is small enough so that for $0 \leq x \leq d_i^{\text{eff}}$, $Q(x)$ is well approximated by $0.5 - \left(x/\sqrt{2\pi}\right)$. As we choose $\gamma > 1 + \mu$ to accommodate scalability constraints, this is equivalent to the ring index i being large enough. Thus, for outer rings, $\log P_{\text{miss}}(i) \lessapprox -d_i^2/\left(r\sigma\sqrt{2\pi T_i}\right)$.

C.3 Bound on route stretch

To provide bounds on the worst case route stretch, we need to understand the geometry of the update rings surrounding a destination node that executes the position-publish algorithm. The publish algorithm ensures that at all times, there exists *exactly one set* of normal updates with valid confidence region guarantees corresponding to each ring index. All other updates with valid confidence region guarantees are abnormal updates issued in order to prevent older unexpired updates (made stale by atypically large movements of the destination node) from misdirecting packets. When a packet is relayed through an update ring to which updates that were issued earlier have become stale, abnormal updates issued to the same region ensure that the packet latches on to the newer estimate that they possess rather than the stale estimates. This is because the routing algorithm prefers newer updates (more recent) when it is presented with a tie in terms of the ring indices. The analysis in the derivation of $P_{\text{miss}}(i)$ holds here for the probability of missing the *newer* updates (in regions with stale updates that also have the same spatial validity). The role of abnormal updates is therefore to prevent routing failures as a result of misdirection from stale updates. Therefore, we can assume that abnormal updates and stale updates which were compensated for

using abnormal updates are absent in a discussion of worst case routing stretch guarantees.

Consider a destination node which is surrounded by one normal update ring of each ring index, all of which satisfy their corresponding confidence region guarantees. We denote the center and the position estimate (both of which coincide for normal updates) of the i-th update ring by $\mathbf{c}_i$. Since these position updates satisfy their corresponding confidence region guarantees, $\ell(\mathbf{c}_i, \mathbf{d}) \leq \beta r_i$.

Inner and outer envelope of updates: In the following computations it is useful to define the concept of inner and outer envelope of an update of ring index i. The outer envelope of updates of ring index i is defined as the set of points farthest from the destination node that can possess a spatially active update of ring index i. Similarly the inner envelope of updates of ring index i is defined as the set of points closest to the destination node that can possess a spatially active update of ring index i. In each direction the farthest spatially valid update can be $(1+\beta) r_i$ away corresponding to $\ell(\mathbf{c}_i, \mathbf{d}) = \beta r_i$ and $\mathbf{c}_i$ in the same direction (we neglect d_i in this computation as $d_i/r_i \ll 1$). So the outer envelope of updates of ring index i is the circle of radius $(1+\beta) r_i$ centered around the destination node $\mathbf{d}$. Similarly the inner envelope of updates is the circle of radius $(1-\beta) r_i$ centered around the destination node.

If the packet source has an active estimate of the destination, then the packet proceeds towards this estimate. However, if it does not possess an active update, the source launches the packet in some arbitrarily chosen direction and the packet eventually *bootstraps* when it is relayed to a node which possesses an active estimate of the destination node's location. This process of bootstrapping contributes to route stretch in addition to the route stretch stemming from lazy updates.

Firstly, we examine the contribution of the lazy position updates to the route stretch after bootstrap. An upper bound on this is given via the worst case uncertainty seen by a packet that has bootstrapped. Then we provide an upper bound on the worst case contribution to route stretch due to the bootstrapping process which corresponds to a confluence of unfavorable geometric configurations of the immediate inner and outer rings surrounding the packet source and the launch direction chosen by it.

Stretch after bootstrap

What is the worst case uncertainty seen by a packet after it has bootstrapped? To answer this question consider a segment of any packet trajectory from just after it has acquired the update for the $(i+1)$-th ring till it acquires the update for the i-th ring. In this segment of the packet trajectory, the packet is routed using

the estimate from the $(i+1)$-th ring and so the location estimate $\mathbf{e} = \mathbf{c}_{i+1}$ can disagree from the true destination location by not more than $\beta r_{i+1} = \alpha\beta r_i$ (i.e., $\ell(\mathbf{d}, \mathbf{e}) \leq \beta r_{i+1} = \alpha\beta r_i$). The farthest from the destination node that an update of index $i+1$ can be obtained by the packet is given by $(1+\beta) r_{i+1}$, the radius of the outer envelope of the $(i+1)$-th ring and the closest the packet can get to the destination without latching on to the i-th ring update is $(1-\beta) r_i$, the radius of inner envelope of updates of the i-th ring. This is the region where a packet can use the $(i+1)$-th indexed update, and in this region, uncertainty $\ell(\mathbf{d}, \mathbf{e})/\ell(\mathbf{p}, \mathbf{d})$ satisfies $U \leq U_{\max}$, where $U_{\max} = \alpha\beta/(1-\beta)$. This bound is independent of the update index $i+1$ and thus as long as the packet does not "miss" any update ring that it is relayed through, the uncertainty seen by it is no greater than $U_{\max}$. Therefore, from Corollary 4.2 of Theorem 4.1, the route stretch after bootstrap is at most $1/\cos(\arcsin U_{\max} + \delta) \approx 1/\sqrt{1 - U_{\max}^2}$ (assuming that δ is small and ϵ in the choice of communication radius $\pi r^2 = (1+\epsilon)\log n$ is chosen to satisfy (4.2)).

From the preceding discussion, we see that when $U_{\max} < 1$ or $\beta r_{i+1} < (1-\beta) r_i$, before the packet reaches the estimate $\mathbf{c}_{i+1}$ corresponding to the $(i+1)$-th ring, it is guaranteed to be relayed via the i-th ring, thus acquiring the estimate $\mathbf{c}_i$ corresponding to the i-th update. Therefore, the packet is guaranteed to progressively obtain better estimates of the destination, and this imposes the extra constraint on routing reliability (Section 4.5.2) given by $\alpha\beta < 1 - \beta$ or $U_{\max} < 1$.

Bootstrapping cost

Consider a packet originating from the source node positioned at $\mathbf{s}$ for the destination at $\mathbf{d}$. Suppose the rings that surround the source are the i-th (inner) and $(i+1)$-th (outer) rings. A packet launched in any direction first bootstraps at one of these two rings. Since we are interested in route stretch, which is a ratio of distances, we henceforth scale all distances by r_i, the radius of the inner ring. The source can be surrounded by the i-th and $(i+1)$-th ring only if it is inside the region between the inner envelope of the i-th ring and the outer envelope of the $(i+1)$-th ring. Therefore $1-\beta \leq \ell(\mathbf{s},\mathbf{d}) \leq \alpha(1+\beta)$.

When the packet is launched from the source in an arbitrary direction, the scenario where the packet latches to the ring closer to the destination is a better case. So consider the case where the packet bootstraps at the outer ring. We note that for all distances of the source node from the destination node the packet travels the farthest distance before bootstrapping if it is launched tangential to the inner envelope of the inner ring and bootstraps at the outer envelope of the outer ring. This scenario is depicted in Figure C.3 (left). We denote the point where this tangential trajectory touches the inner envelope of the inner ring by $\mathbf{t}$ and the point on the outer envelope of the outer ring where the packet bootstraps by $\mathbf{b}$.

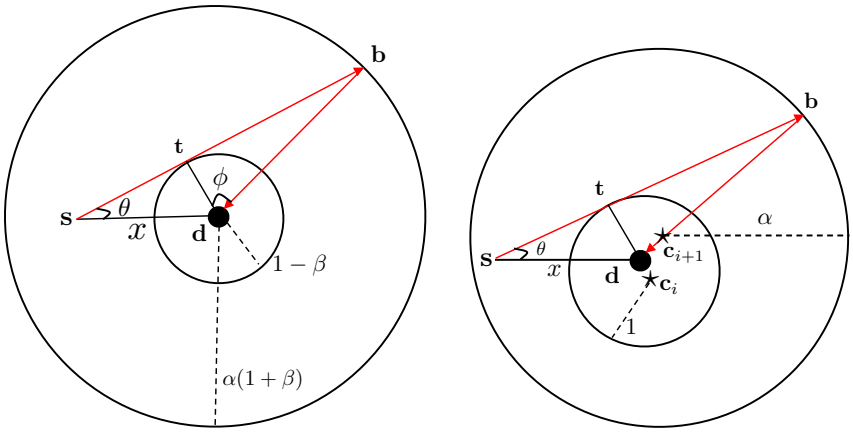

Figure C.3: Left: Contribution of bootstrapping to stretch via envelopes. Right: A configuration of ring centers $\mathbf{c}_i$ and $\mathbf{c}_{i+1}$ marked $\star$ which satisfies the confidence region guarantees and has the same bootstrapping cost as the worst case envelope based calculations

Let θ be the angle between the launching direction and the direction of the destination. Then $\sin\theta = (1-\beta)/x$. Let ϕ be the angle between the vectors $\mathbf{t}$ and $\mathbf{b}$ as shown in Figure C.3 (left). Then we have $\cos\phi = (1-\beta)/(\alpha\,(1+\beta))$. From the cosine formula, the worst case distance traveled before bootstrap $Z(x) = \ell\,(\mathbf{s},\mathbf{b})$ for a packet originating at a distance $x = \ell\,(\mathbf{s},\mathbf{d})$ from the destination satisfies (using Theorem 4.1 and assuming that ϵ is large enough so that δ in (4.2) is small; note that forwarding along a fixed direction is equivalent to forwarding greedily towards an estimate at ∞ and thus in the regime where Theorem 4.1 is

applicable)

$$Z(x) = \sqrt{x^2 + \alpha^2 (1+\beta)^2 - 2x\alpha (1+\beta) \cos\left(\frac{\pi}{2} - \theta + \phi\right)}$$
$$= \sqrt{x^2 - (1-\beta)^2} + \sqrt{\alpha^2 (1+\beta)^2 - (1-\beta)^2}.$$

Overall stretch

We are now ready to bound the overall stretch using the preceding two ingredients: A packet originating at a distance x from the destination travels a distance not exceeding $Z(x)$ before bootstrap at $\mathbf{b}$. Since the stretch after bootstrap is at most $1/\sqrt{1 - U_{\max}^2}$ ($U_{\max}$ is the upper bound on uncertainty given by (4.4)) and $\ell(\mathbf{b}, \mathbf{d}) = \alpha(1+\beta)$, the overall route stretch is at most $\mathtt{S}$:

$$\mathtt{S} = \max_{\ell(\mathbf{s},\mathbf{d})} \left(\ell(\mathbf{s}, \mathbf{b}) + \left(\ell(\mathbf{b}, \mathbf{d}) \Big/ \sqrt{1 - U_{\max}^2}\right)\right) \Big/ \ell(\mathbf{s}, \mathbf{d})$$
$$= \max_{1-\beta \le x \le \alpha(1+\beta)} \left(\sqrt{x^2 - (1-\beta)^2} + \frac{\alpha(1-\beta^2)}{\sqrt{(1-\beta)^2 - \alpha^2\beta^2}} + \sqrt{\alpha^2 (1+\beta)^2 - (1-\beta)^2} \right) \Big/ x$$

and it can be shown that this maximum is equal to the expression in (4.7).

While we consider worst case envelopes of rings for the above discussion, this scenario can be mapped to a feasible configuration of inner and outer update rings as is shown in Figure C.3 (right) because each point on the envelopes corresponds to a certain valid choice of ring center. The center of the inner ring is $\mathbf{c}_i =$

$-\beta\mathbf{t}/(1-\beta)$ and that of the outer ring $\mathbf{c}_{i+1} = \beta\mathbf{b}/(1+\beta)$. Note that the same launch trajectory is now a tangent to the inner ring with center as specified. Further $\ell(\mathbf{c}_i, \mathbf{d}) = \beta$ and $\ell(\mathbf{c}_{i+1}, \mathbf{d}) = \alpha\beta$, which satisfy their confidence region guarantees.

Bibliography

[1] Tweepy: An easy-to-use Python library for accessing the Twitter API. `https://github.com/tweepy/tweepy.git`.

[2] I. Abraham, D. Dolev, and D. Malkhi. LLS: a locality aware location service for mobile ad hoc networks. In *Proceedings of the 2004 joint workshop on Foundations of mobile computing*, DIALM-POMC '04, 2004.

[3] D. Achlioptas. Database-friendly random projections. PODS '01, page 274281, New York, NY, USA, 2001. ACM.

[4] E. Arias-Castro and Y. Eldar. Noise Folding in Compressed Sensing. *Signal Processing Letters, IEEE*, 18(8):478–481, 2011.

[5] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253263, 2008.

[6] R. Baraniuk and M. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):5177, 2009.

[7] S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward. A distance routing effect algorithm for mobility (DREAM). In *Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, 1998.

[8] S. Basu and Y. Bresler. A global lower bound on parameter estimation error with periodic distortion functions. *Information Theory, IEEE Transactions on*, 46(3):1145–1150, 2000.

[9] K. Bell, Y. Steinberg, Y. Ephraim, and H. Van Trees. Extended Ziv-Zakai lower bound for vector parameter estimation. *Information Theory, IEEE Transactions on*, 43(2):624 –637, mar 1997.

[10] I. Ben-Gal. Outlier detection. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 131–146. 2005.

[11] Z. Ben-Haim, T. Michaeli, and Y. Eldar. Performance Bounds and Design Criteria for Estimating Finite Rate of Innovation Signals. *Information Theory, IEEE Transactions on*, 58(8):4993–5015, 2012.

[12] B. N. Bhaskar, G. Tang, and B. Recht. Atomic norm denoising with applications to line spectral estimation. *CoRR*, abs/1204.0562, 2012.

[13] E. Candes and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *CoRR*, abs/1203.5871, 2012.

[14] E. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406 –5425, dec. 2006.

[15] Y. Chi, L. Scharf, A. Pezeshki, and A. Calderbank. Sensitivity to basis mismatch in compressed sensing. *Signal Processing, IEEE Transactions on*, 59(5):2182–2195, 2011.

[16] M. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the Political Alignment of Twitter Users. In *2011 IEEE third international conference on social computing (SOCIALCOM) and Privacy, Security, Risk and Trust (PASSAT)*, 2011.

[17] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, 2001.

[18] D. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289 –1306, april 2006.

[19] M. F. Duarte and R. G. Baraniuk. Spectral compressive sensing. *Applied and Computational Harmonic Analysis*, 2012.

[20] H. Dubois-Ferriere, M. Grossglauser, and M. Vetterli. Age matters: efficient route discovery in mobile ad hoc networks using encounter ages. In *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing*, 2003.

[21] A. Fannjiang and W. Liao. Coherence pattern-guided compressive sensing with unresolved grids. *SIAM Journal on Imaging Sciences*, 5(1):179–202, 2012.

[22] P. Ferreira. Super-resolution, the recovery of missing samples and vandermonde matrices on the unit circle. In *Proceedings of the Workshop on Sampling Theory and Applications, Loen, Norway*, 1999.

[23] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

[24] R. Flury and R. Wattenhofer. MLS: an efficient location service for mobile ad hoc networks. In *Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing*, 2006.

[25] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5), Feb. 2010.

[26] M. Grossglauser and D. N. C. Tse. Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Transactions on Networking*, August 2002.

[27] M. Grossglauser and M. Vetterli. Locating mobile nodes with EASE: learning efficient routes from encounter histories alone. *IEEE/ACM Transactions on Networking*, June 2006.

[28] P. Gupta and P. Kumar. Critical power for asymptotic connectivity. In *Decision and Control, 1998. Proceedings of the 37th IEEE Conference on*, 1998.

[29] P. Gupta and P. Kumar. The capacity of wireless networks. *Information Theory, IEEE Transactions on*, mar 2000.

[30] L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 11(9):1074–1085, Sep 1992.

[31] J. Haupt and R. Nowak. Compressive sampling for signal detection. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 3, pages III–1509 –III–1512, april 2007.

[32] H. Jeffreys. *Theory of probability*. Oxford University Press, 1998.

[33] D. Johnson and D. Maltz. Dynamic source routing in ad hoc wireless networks. In T. Imielinski and H. Korth, editors, *Mobile Computing*. 1996.

[34] B. Karp and H. T. Kung. GPSR: greedy perimeter stateless routing for wireless networks. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, 2000.

[35] J. Li, J. Jannotti, D. S. J. De Couto, D. R. Karger, and R. Morris. A scalable location service for geographic ad hoc routing. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, 2000.

[36] C. X. Lin, B. Zhao, Q. Mei, and J. Han. PET: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, 2010.

[37] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

[38] I. Maravic and M. Vetterli. Sampling and reconstruction of signals with finite rate of innovation in the presence of noise. *Signal Processing, IEEE Transactions on*, 53(8):2788–2805, 2005.

[39] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, 2011.

[40] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, 2010.

[41] M. Mishali and Y. C. Eldar. Xampling: Compressed Sensing for Analog Signals. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

[42] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press*, 14, 2001.

[43] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, 2012.

[44] A. Pak and P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, may 2010.

[45] D. Ramasamy, S. Venkateswaran, and U. Madhow. Compressive estimation in AWGN: General observations and a case study. In *Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on*, pages 953–957, 2012.

[46] D. Ramasamy, S. Venkateswaran, and U. Madhow. Compressive tracking with 1000-element arrays: A framework for multi-Gbps mm wave cellular downlinks. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 690–697, 2012.

[47] D. Ramasamy, S. Venkateswaran, and U. Madhow. Compressive parameter estimation in awgn. *Signal Processing, IEEE Transactions on*, 62(8):2012–2027, April 2014.

[48] D. Rife and R. Boorstyn. Single tone parameter estimation from discrete-time observations. *Information Theory, IEEE Transactions on*, 20(5):591–598, 1974.

[49] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11), Nov. 1998.

[50] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, 2010.

[51] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, 2009.

[52] C. A. Santiváñez, R. Ramanathan, and I. Stavrakakis. Making link-state routing scale for ad hoc networks. In *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing*, 2001.

[53] L. L. Scharf. *Statistical signal processing*, volume 98. Addison-Wesley Reading, MA, 1991.

[54] D. Shamma, L. Kennedy, and E. Churchill. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events. *CSCW Horizons*, 2010.

[55] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: Modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, 2011.

[56] G. Tang, B. N. Bhaskar, and B. Recht. Near Minimax Line Spectral Estimation. *CoRR*, abs/1303.4348, 2013.

[57] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed Sensing off the Grid. *CoRR*, abs/1207.6053, 2012.

[58] H. Van Trees and K. Bell. Bayesian bounds for parameter estimation and nonlinear filtering and tracking. 2007.

[59] H. L. Van Trees, K. L. Bell, and Z. Tian. *Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2013.

[60] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. arXiv e-print, Nov. 2010. Chapter 5 of: Compressed Sensing, Theory and Applications. Edited by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012.

[61] M. Wakin. Manifold-based signal recovery and parameter estimation from compressive measurements. *Arxiv preprint arxiv:1002.1247*, 2010.

[62] P.-J. Wan, C.-W. Yi, L. Wang, F. Yao, and X. Jia. Asymptotic critical transmission radii for greedy forward routing in wireless ad hoc networks. *Communications, IEEE Transactions on*, May 2009.

[63] A. Weiss and B. Friedlander. Preprocessing for direction finding with minimal variance degradation. *Signal Processing, IEEE Transactions on*, 42(6):1478–1485, 1994.

[64] J. Weng and B.-S. Lee. Event detection in twitter. In *ICWSM*, 2011.

[65] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 177–186. ACM, 2011.