Chapter 1

Introduction

1.1 Serendipity

The story goes like this: late in the summer of 1967, Jocelyn Bell and a team of other Cambridge graduate students had just finished building a new radio telescope, a 4.5 acre collection of dipole antennas. Bell was saddled with the task of analyzing, by hand, the pen-and-paper chart recordings that the telescope produced. The data rate—expressed in the most natural units—was 96 feet of paper per day.

Bell was looking for quasars, which, as point sources, should be highly modulated by interplanetary scintillation. But close to two months after beginning the search for these modulated extragalactic sources, Bell noticed something else: a "bit of scruff" that appeared on the recordings at the same right ascension each day. Months of poring over these recordings had sharpened Bell's eye, and she could tell by sight that the signal was neither a scintillating quasar nor terrestrial interference.

Bell's discovery turned out to be a big one: the "scruff" was a pulsar, and it was the first observational evidence for Baade & Zwicky (1934)'s suggestion that supernovae leave behind small, dense neutron stars. The discovery of PSR 1919+21 (Hewish et al., 1968) changed the face of radio astronomy (or, at least, gave it a prominent new feature) and has allowed astronomers to probe extreme physics, general relativity, and the interstellar medium as well as stellar evolution.

Bell's discovery wasn't all serendipity. Her keen eye and analytical fastidious-

ness deserve credit, too. But it is likely that modern analysis algorithms, which churn through hundred of gigabytes of data a day, would miss a signal that, like PSR 1919+21, is different in character and timescale from those for which they were designed to search.

A sensitive telescope and a little luck may produce some paradigm-shifting discoveries, but nimble algorithms and intelligent survey design can tip the odds in serendipity's favor. Applied to the multi-petabyte datasets expected from the new 7-pixel Arecibo L-band Feed Array (ALFA) and from future telescopes like the Square Kilometer Array (SKA) and the Low Frequency Array (LOFAR), these techniques are likely to reveal entirely new classes of objects.

1.1.1 A Brief History of Serendipity

Historically, what factors have conspired to yield the most surprising new discoveries in radio astronomy? New instruments, for one. Order-of-magnitude improvements in sensitivity, observing bandwidth, and resolution (spatial, spectral, or temporal) naturally lead to the discovery of new classes of objects. It has been said that the discoveries for which telescopes are best remembered are rarely the discoveries they were designed to make–unexplored regions of parameter space hold the best surprises.

The story of radio astronomy is all variations on this theme, going back to the birth of the field in 1933 (Jansky, 1933; Kellermann & Sheets, 1983). Karl Jansky, working for Bell Telephone Laboratories in Holmdel, New Jersey, designed and built a new, steerable instrument to study the static that Bell feared could complicate radio telephone service across the Atlantic.

Most of the static, Jansky found, came from thunderstorms. But beneath the

shifting storm static, Jansky detected a steady radio "hiss." He tracked the hiss for a year before determining that it was of astronomical origin: specifically, it came from the Milky Way. "Star noise," he called it.

In Jansky's case and in general, new discoveries require more than a new instrument. (Indeed, since new instruments so reliably return new discoveries, it may not be fair to call such discoveries serendipitous.) In the Preface to the workshop proceedings *Serendipitous Discoveries in Radio Astronomy* (Kellermann & Sheets, 1983), Kellerman writes:

> "...Major discoveries require the 'right person, in the right place, doing the right thing, at the right time...' It sometimes helped 'not to know too much.' "

In addition to new instruments, the real surprise discoveries rely on the intangibles Kellerman describes: the keen eye, the scientific naiveté (intentional or otherwise), and the lucky coincidence. So it was with Jansky's galactic radio emission and with Bell's pulsars. Add to the list of serendipitous discoveries the cosmic microwave background radiation (Penzias & Wilson, 1965), Jupiter's bursting radio emission (Burke & Franklin, 1955), radio novae (Hjellming & Wade, 1970), and quasars (Schmidt, 1963), and you have a respectable inventory of Radio Astronomy's Greatest Hits. Indeed, in his book *Cosmic Discovery: The Search, Scope and Heritage of Astronomy*, Martin Harwit estimates that half of all known cosmic phenomena were discovered unexpectedly.

1.1.2 Radio Transients: Probing New Parameter Space

Here is another way to look at it: new discoveries in astronomy come from probing new regions of parameter space, the space whose mix-and-match menu of di-

mensions includes source flux density, observing wavelength, emission duration, and angular resolution (Harwit, 2003). Jansky expanded the parameter space of observational astronomy to include radio as well as optical wavelengths. Bell's contribution was on the emission duration axis: she demonstrated that the radio sky is variable on very short timescales.

Yet the transient radio sky remains a frontier of observational astronomy, not for lack of scientific promise, but for practical obstacles: terrestrial radio frequency interference (RFI) and limited computing power and solid angle coverage.

Using modern RFI mitigation procedures and taking advantage of improvements in computing power, a transient survey could uncover, in addition to entirely new classes of sources, any of a long list of objects that are predicted to be transient radio emitters but that have not yet been observed. Rees (1977), for instance, suggested that pulses from the explosions of primordial black holes should be observable, but a subsequent search by Phinney & Taylor (1979) produced no detections. Gamma ray bursts, which share fireball physics with primordial black hole explosions, are also predicted to be sources of coherent radio emission coincident in time with the gamma emission, but no positive detections have been made (Benz & Paesold, 1998).

A census of the transient radio sky will also tell us more about objects that have already been observed in the radio band. Single-pulse searches may uncover very long- or short-period pulsars missed by traditional periodic searches (Nice, 1999). It may also be possible to detect extragalactic pulsars via giant pulses, occassional pulses that can reach 1000 times the mean pulse intensity (McLaughlin & Cordes, 2003).

1.2 Why Now?

Today, a confluence of instrumental capability and computing power make a thorough search for transient radio emission a realistic goal. In fact, such a survey is already being planned for ALFA. With ALFA, it will be possible to use cross-correlation rejection techniques, either in hardware or in software, to remove RFI. Because each of ALFA's seven beams points at a different location on the sky, any signal common to more than one beam can be identified as originating from some artificial, non-celestial source.

In addition, ALFA offers the obvious advantage of efficiency: it can cover the sky seven times faster than Arecibo's current single-pixel system. This is particularly important for a large survey using an already over-subscribed telescope. The SKA promises a hundred-fold boost in sensitivity, with even more robust interference mitigation properties and faster sky coverage.

1.3 Results Presented in this Thesis

In this work, we discuss some of the "nimble algorithms" suitable for processing the large data sets expected from ALFA and, in the future, from the SKA and the LOFAR. In Chapter 2, specific transient search algorithms (Friends of Friends and matched filtering) are discussed and applied to both simulated and real data. Chapter 3 presents results of and motivation for a search for transient radio emission from four known extrasolar planets.

REFERENCES

Baade, W. & Zwicky, F. 1934, Proceedings of the National Academy of Science, 20, 259

Benz, A. O. & Paesold, G. 1998, A&A, 329, 61

Burke, B. F. & Franklin, K. L. 1955, J. Geophys. Res., 60, 213

Harwit, M. 1981, Cosmic Discovery: The Search, Scope and Heritage of Astronomy

Harwit, M. 2003, Physics Today, 56, 38

Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F. & Collins, R. A. 1968, Nature, 217, 709

Hjellming, R. M. & Wade, C. M. 1970, ApJ, 162, L1

Jansky, K. 1933, Nature, 132, 66

Kellermann, K. I. & Sheets, B. 1983, Serendipitous Discoveries in Radio Astronomy

McLaughlin, M. A. & Cordes, J. M. 2003, ApJ, 596, 982

Nice, D. J. 1999, ApJ, 513, 927

Penzias, A. A. & Wilson, R. W. 1965, ApJ, 142, 419

Phinney, S. & Taylor, J. H. 1979, Nature, 277, 117

Rees, M. J. 1977, Nature, 266, 333

Schmidt, M. 1963, Nature, 197, 1040

Chapter 2

Detection Algorithms

2.1 Introduction

In this chapter, we address the problem of designing algorithms flexible enough to detect a wide variety of signals (e.g. broadband, narrowband, weak, strong, chirped), yet robust enough to return a low number of false positives. It is also critical that the algorithm return a manageably-sized event list since a human must follow up on the algorithm's output.

We discuss three possible detection algorithms: thresholding, the simplest approach; Friends of Friends, which builds on the thresholding algorithm; and matched filtering.

2.2 Pre-detection Processing

Before applying transient-seeking algorithms, we normalize data to correct for non-uniform instrumental frequency response.

Ignoring scintillation, the measured intensity of a source depends on the source's intrinsic intensity as well instrumental response. For either the case of linear polarization or the case of circular polarization, the measured intensity in each polarization (labeled 1,2) can be represented as

$$I_{obs}(\nu, t)_{1,2} = G_{1,2} B_{1,2}(\nu) I_{in}(\nu, t)_{1,2} \tag{2.1}$$

where $G_{1,2}$ represents the telescope's polarization-dependent gain, $B_{1,2}(\nu)$ represents the spectrometer's frequency response (bandpass) in each polarization, and

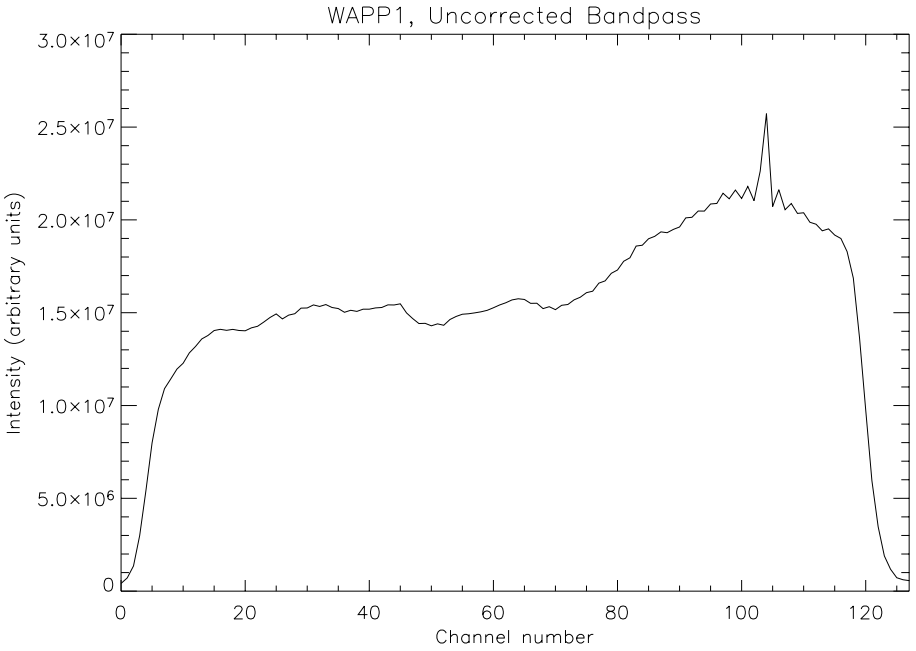

Figure 2.1: A typical 128-channel WAPP bandpass. A narrowband RFI spike appears in channels 103 and 104.

$I_{in}(\nu, t)_{1,2}$ is the source's intrinsic intensity.

To correct for the spectrometer's non-uniform frequency response for each dynamic spectrum, we integrate over time to calculate an average bandpass. Strong narrowband spikes are removed from the bandpass using an iterative algorithm which identifies these spikes and replaces them with the average value of the neighboring bandpass channels. The user selects the number of iterations to perform as well as the channel deviation limit. (We found 3 iterations and a 1.5σ deviation limit effective.)

Figures 2.1 and 2.2 show a typical WAPP bandpass before and after the removal of narrowband spikes.

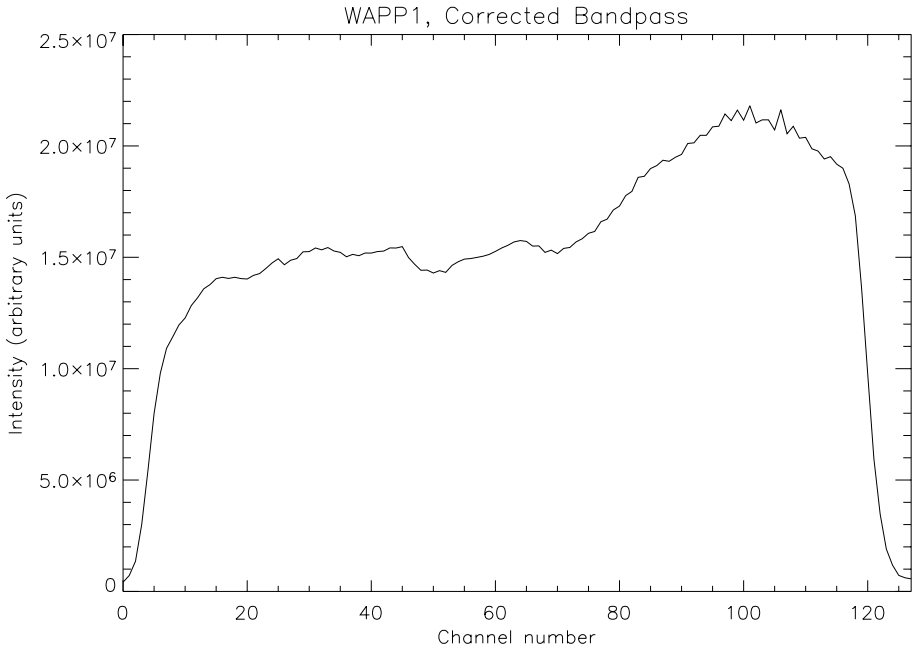

Figure 2.2: The same bandpass with the spikes removed using the bandpass correction algorithm. The channels are replaced with the average value of the six surrounding "good" channels.

2.3 Thresholding

The simplest way to parse large data sets is the threshold test: the spectrum the is first flattened, as described in the previous section, and a threshold limit (say, 5σ) is set. Any samples more than that threshold above the data mean are flagged for follow-up. The threshold choice depends on the size of the data set and on the acceptable number of false positives. Assuming Gaussian statistics, the number of samples above threshold due to radiometer noise is given by:

$$N = n_\nu n_f (\frac{1}{2} - P) \tag{2.2}$$

where n_ν is the number of frequency channels, n_f the number of time samples, and P the Gaussian probability density integrated from zero to the threshold $m\sigma$ (assuming that we are only interesting in points *above* threshold):

$$P = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{m\sigma} e^{\frac{x^2}{2\sigma^2}} dx \tag{2.3}$$

In most cases, the data contain strong RFI which biases the rms high. To get the most accurate rms, we iteratively mask strong pulses in the data and re-calculate the rms without them.

2.4 Friends of Friends

2.4.1 The Algorithm

The Friends of Friends (FOF) algorithm, which has roots in the percolation theory branch of statistical physics, was first borrowed by astronomers for the study of galaxy clustering (Huchra & Geller, 1982). The Friends of Friends (FOF) algorithm is suited to finding structure of all kinds; it has also, for example, been applied

to genome research. In the case of galaxy clustering, the FOF algorithm identifies clusters by the spatial distance between their component galaxies. First two nearby galaxies are identified, then a third within some specified distance of one of the previous pair, and so on, until no galaxies are near enough any member of the group to be related to it. This technique can identify both "social" and "singleton" galaxies. The basic anatomy of the algorithm is outlined in Figure 2.3.

We apply an FOF algorithm to dynamic spectra, two-dimensional data sets with observing frequency on one axis and observing time on the other. (Measured intensity occupies the z-dimension or greyscale level.) Instead of identifying spatially related galaxies, we aim to identify data points related in time and frequency.

To begin, the entire dynamic spectrum is normalized (see Section 2.2) and thresholded to produce a list of samples with measured intensity exceeding some user-defined threshold above the mean of the dynamic spectrum. This list (in our pipeline, an ASCII file containing the sample's frequency and time coordinates and measured intensity) is then processed by the FOF algorithm. There is a great deal of flexibility in this approach; the stringency of the thresholding (m, where the threshold is set to $m\sigma$) and the maximum time and frequency distance between "friendly" points (a radius of δs, or $\delta\nu$, δt if the user desires unequal time and frequency separation allowances) are all user-defined. Since the algorithm can produce groups of all sizes (one-sample groups are common; conversely, setting the previously described parameters too loosely can cause the routine to return a single group containing all the thresholded samples), the user also has control over the minimum population (p) at which he or she deems groups worthy of further attention.

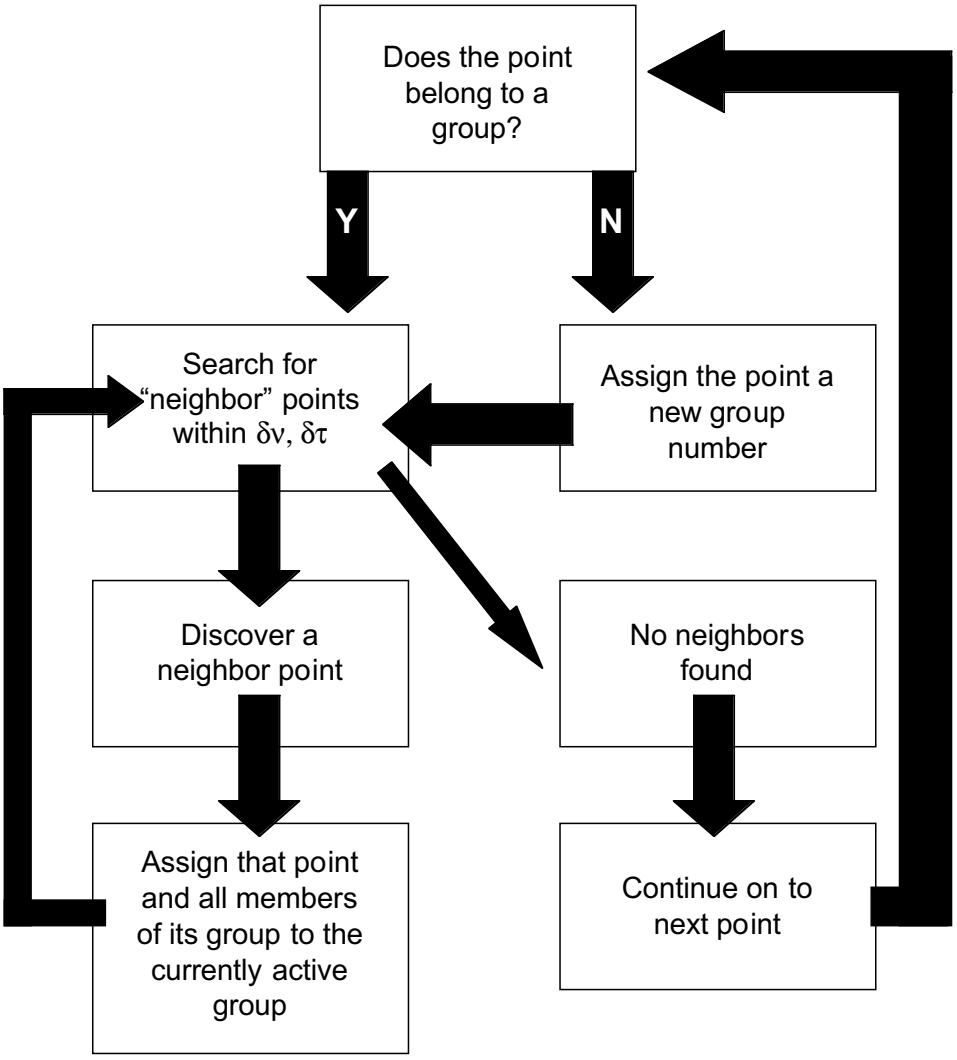

Figure 2.3: The Friends of Friends algorithm.

2.4.2 Simulations

We applied the FOF routine to a 1024×1024 grid of computer generated, normally distributed random numbers (noise). Figures 2.4 and 2.5 show how the input parameters affect the routine's output. In Figure 2.4, we compare 9 sets of input parameters for threshold levels ranging from $m = 2$ to $m = 4$. (Too few samples survive thresholds greater than $m = 4$.) The three leftmost (dotted) curves show the number of groups with populations greater than or equal to $p = 1$, 5, and 10 samples for $\delta s = 10$. The corresponding set of curves is shown for $\delta s = 15$ in dashed lines. Curves for $\delta s = 20$ are shown in thin solid lines. The total number of above-threshold points is plotted as a thick solid line.

Each of the nine curves has three major features: a dip at low thresholds, a central peak, and a high-threshold fall-off. At very low thresholds, the high density of above-threshold points leads to a small number of very populous groups. This is is responsible for the low-threshold dip. For high thresholds and low values of p, on the other hand, the density of samples above threshold is too low to support the formation of groups; the number of groups is therefore nearly equal to the number of above-threshold samples. At high thresholds, there is essentially no deviation between the thick solid line (total number of above-threshold samples) and the FOF output. Another way to see this is that, at high thresholds, each group contains a single sample. For this reason, the high-threshold fall-off is approximately Gaussian, mirroring the samples' normal distribution.

For very "tight" groups (that is, for those with a small maximum separation parameter), the low threshold dip disappears, as shown in Figure 2.5. This figure shows curves analogous to those in Figure 2.4 for a maximum separation of 5 samples. Here, the small group size compensates for the high density of over-

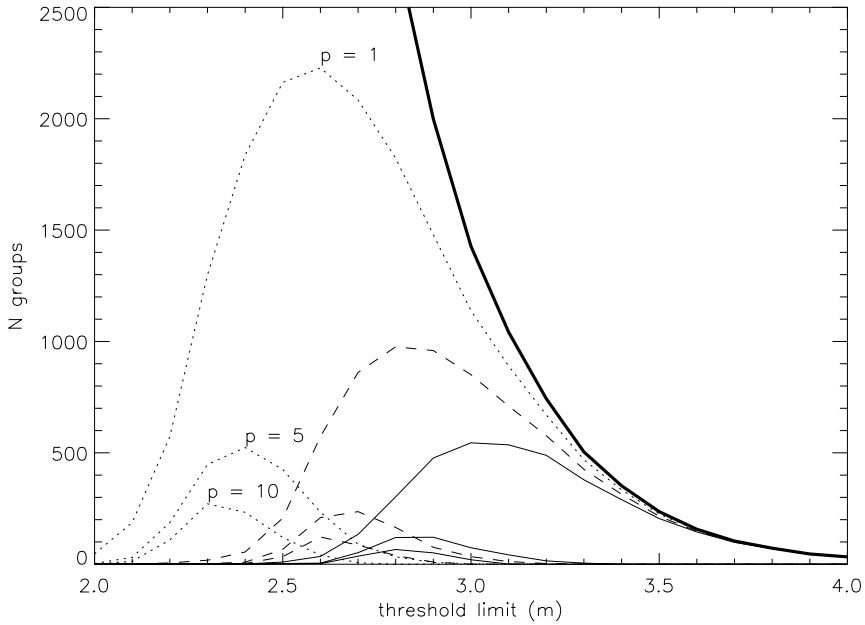

Figure 2.4: Number of groups returned by the FOF algorithm as a function of m is plotted for three values of minimum group population (p) and three values of δs. Dotted lines: $\delta s = 10$. Dashed lines: $\delta s = 15$. Thin solid lines: $\delta s = 20$. The total number of points above the threshold $m\sigma$ is plotted by the thick solid line. Data were a 1024×1024 grid of normally distributed random numbers.

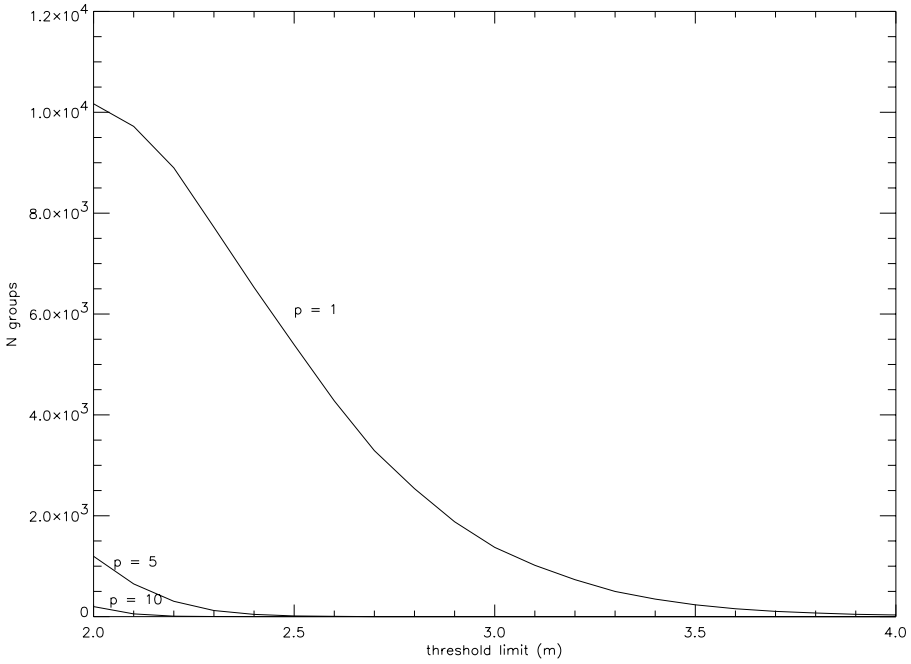

Figure 2.5: Number of groups returned by the FOF algorithm as a function of m for $\delta s = 5$. Such a tight clustering requirement eliminates the low threshold dip seen in Figure 2.4.

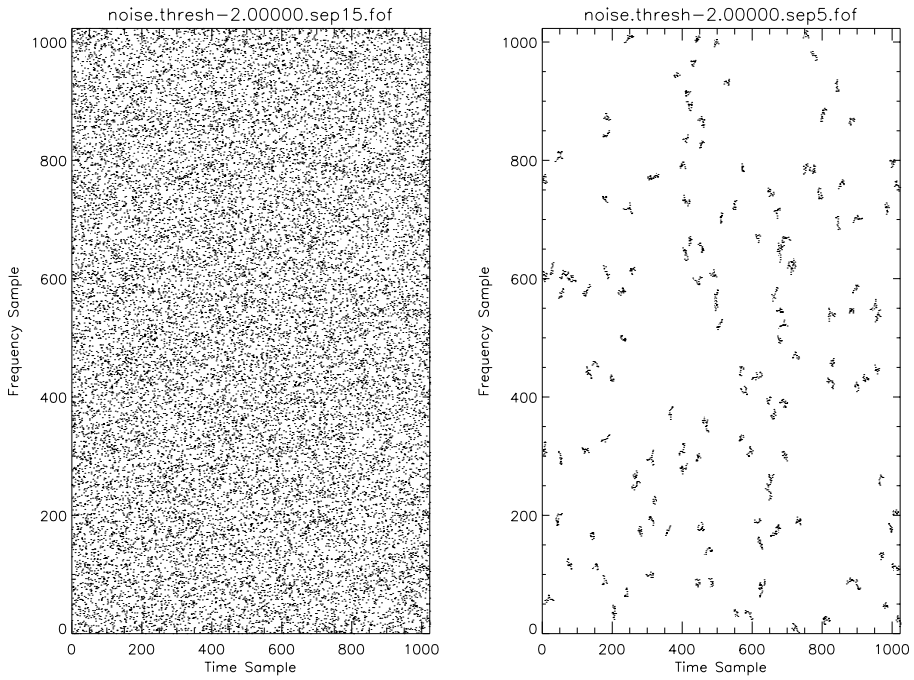

Figure 2.6: The FOF algorithm with $m = 2$ and $\delta s = 15$, left, and $\delta s = 5$, right, applied to Gaussian noise. Only groups containing more than 10 samples are plotted.

B

threshold samples, preventing the aglommeration of samples into a few very large groups. Instead, we see a large number of smaller groups, many with only one or two members.

Figure 2.6 demonstrates this effect. Members of groups with more than ten members are plotted. The left panel shows the result for a threshold of $m = 2$ and $\delta s = 15$. The result is a single group containing all of the above-threshold samples. On the left, a smaller maximum separation ($\delta s = 5$) is used. In this case, 138 groups contain more than 10 samples.

The simulations suggest three ways in which to reduce the number of false positives returned by the FOF routine. First, the individual sample threshold can

be increased. The number of false positives can also be reduced by increasing minimum group membership. A third approach is to increase $\delta\nu$ and $\delta\tau$.

2.4.3 Sample Application: UGC 2339

In Figures 2.9 and 2.10, the FOF routine is applied to data from the galaxy UGC 2339, collected by Bhat et al. at the Arecibo Observatory. Figure 2.9 shows the FOF output for $\delta\nu = \delta t = 10$ and for four different values of m. Each group is shown in an arbitrarily assigned color and is labeled with an identifying number.

In Figure 2.10, we compare the FOF output for a single threshold level ($m = 2.5$) and four different sets of $\delta\nu, \delta t$. The original thresholded data, as input to the FOF routine, is shown for reference in Figure 2.8.

Comparison with the unthresholded dynamic spectrum, shown in Figure 2.7, demonstrates that the FOF routine consistently identifies the strong narrowband RFI near channels 103-104. The exception occurs at $\delta\nu = \delta t = 2$, where the event is identified as seven distinct groups rather than a single one. The same problem affects the routine's performance in the first and last channels; the above-threshold points are too sparse to be identified as a single group.

We suggest two remedies that might be employed in future implementations of the algorithm. First, the algorithm could easily be modified to allow $\delta\nu \neq \delta t$. To seek out narrowband signals, $\delta t \gg \delta\nu$. Similarly, an "OR" statement could be added to identify both narrowband and broadband, short duration signals.

Once a group is identified, an additional routine can be called in to characterize the group by, for instance, its average power, aspect ratio (is it broadband? narrowband?), slope (useful to determining if the signal is a dispersed pulse), fill factor, or any other descriptors the user can dream up. We chose to display the

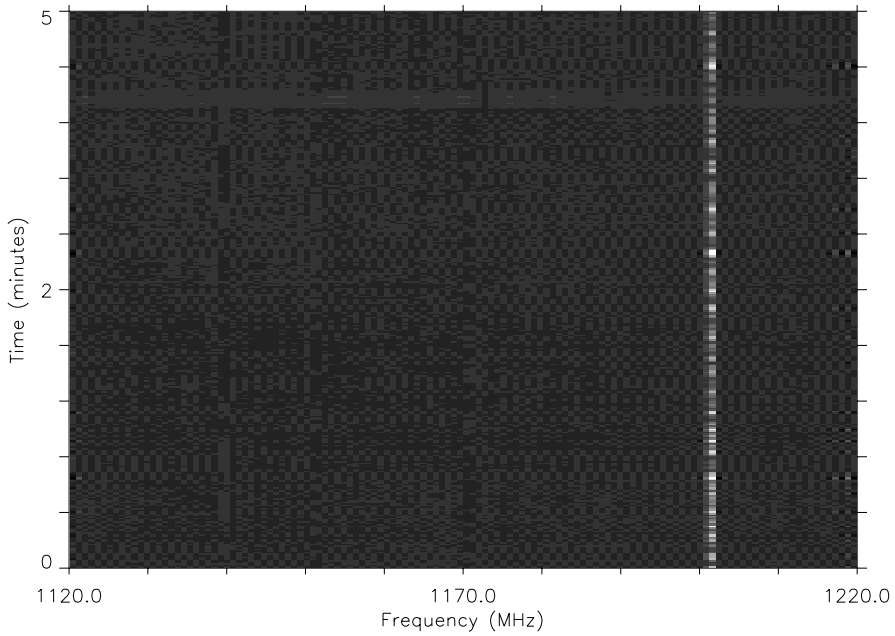

Figure 2.7: The dynamic spectrum of the galaxy UGC 2339, as observed at the Arecibo Observatory.

FOF output using an IDL GUI that allows the user to zoom in on and retrieve basic information about individual groups in a point-and-click environment. Figure 2.11 shows a typical screen shot from this display.

2.4.4 Sample Application: Giant Pulses from the Crab Pulsar

The Crab Pulsar (PSR B0531+21), a young pulsar (P = 33.4 ms, age $\approx$ 950 years) associated with the Crab Nebula supernova remnant, is distinguished by occassional extremely bright pulses. These "giant pulses" are detectable before pulse folding and dedispersion, and have flux densities at 800 MHz least 20 times those of "regular" pulses. They have been observed to exceed 1000 mJy. Approx-

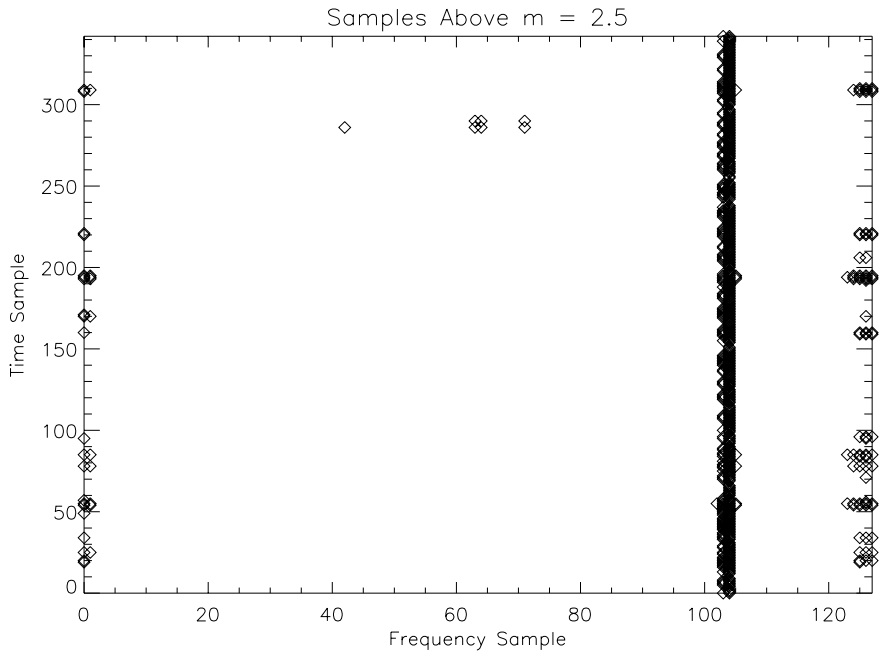

Figure 2.8: The thresholded dynamic spectrum of UGC 2339. Only samples above $m = 2.5$ are displayed.

imately 2.5% of pulses from the Crab are giant pulses, meaning that giant pulses occur, on average, every 10 seconds (Lundgren et al., 1995).

We apply the threshold test and the FOF algorithm to a 1 hour observation of the Crab Pulsar. Observations were made at the Arecibo Observatory 305-m telescope by Cordes et al. in 2002 and were recorded with the Wideband Arecibo Pulsar Processor (WAPP), using 512 frequency channels over a 12.5 MHz bandwidth centered on 430 MHz.

Giant pulses from the Crab are a good test of transient-seeking algorithms' abilities to detect narrowband, dispersed signals, especially those that might come from giant-pulsing pulsars outside our galaxy.

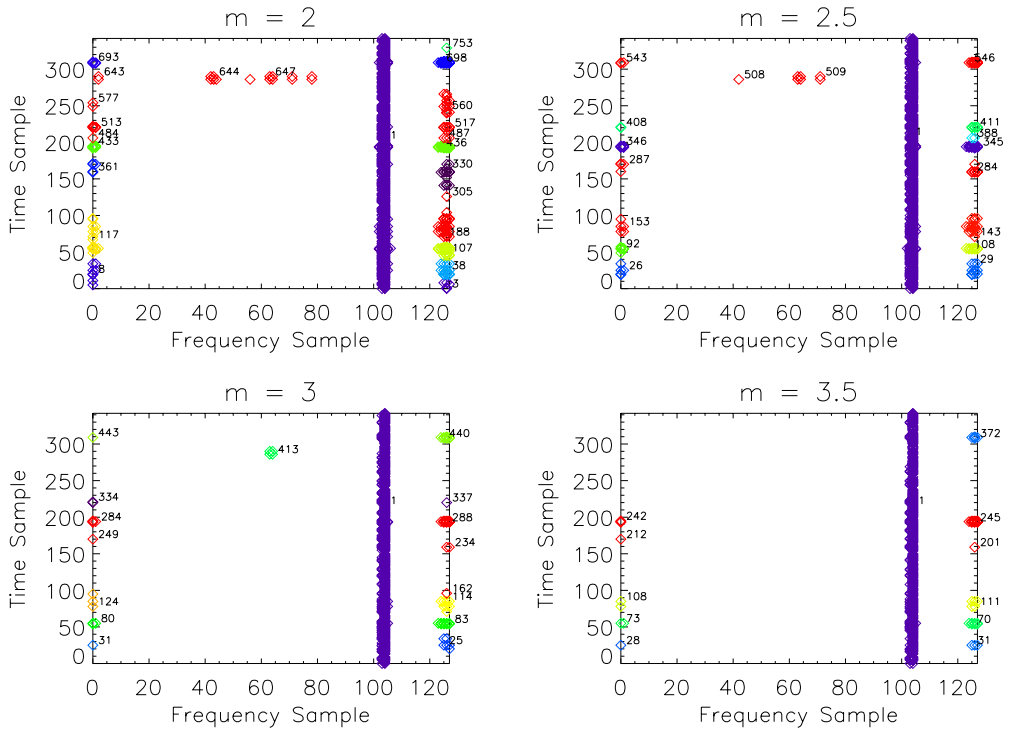

Figure 2.9: The FOF routine output for data from the galaxy UGC 2339. Four different values of m are compared. As expected, the routine returns the largest number of groups for low values of m.

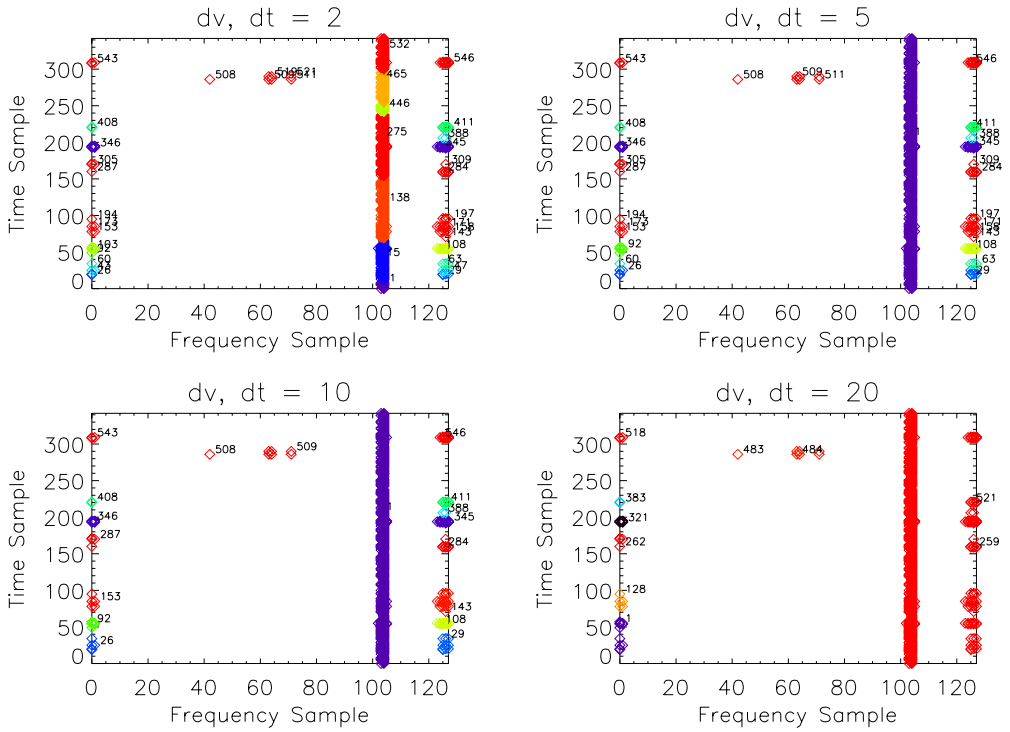

Figure 2.10: The FOF routine output for data from the galaxy UGC 2339. Four different values of $\delta\nu, \delta t$ are compared. Here, the routine performs best for large values of $\delta\nu, \delta t$.

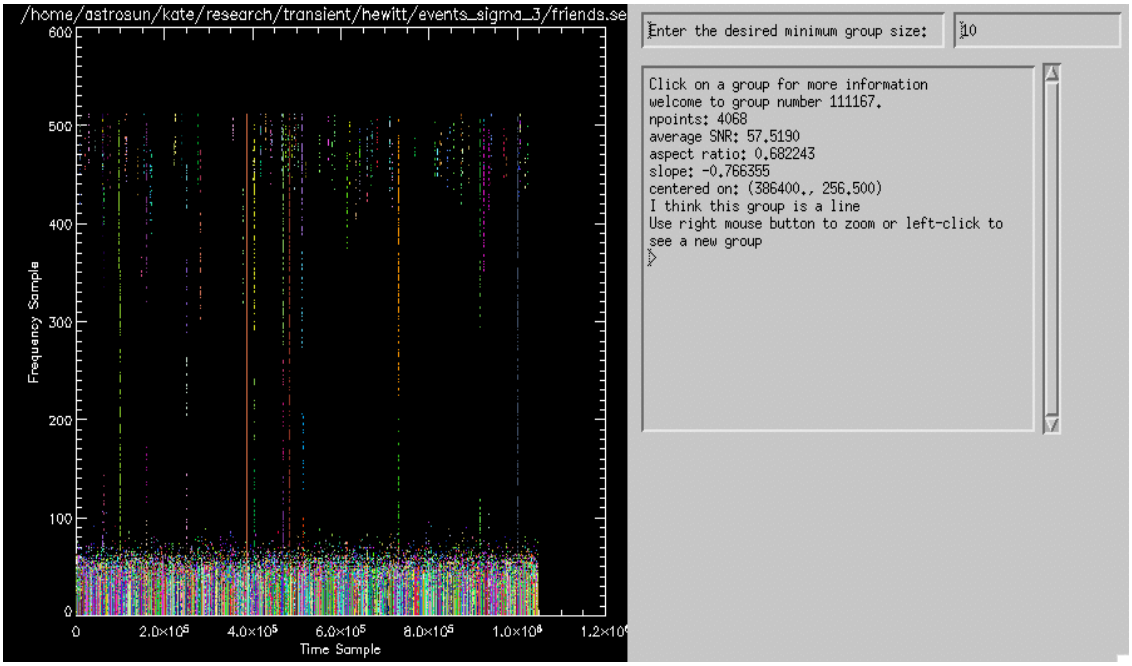

Figure 2.11: Friends of Friends GUI screenshot. The FOF output is shown in the left panel. Once the user clicks on a group, information about the group appears in the right panel. The user can also zoom in on a selected area of the FOF output and interactively set the minimum group size to display.

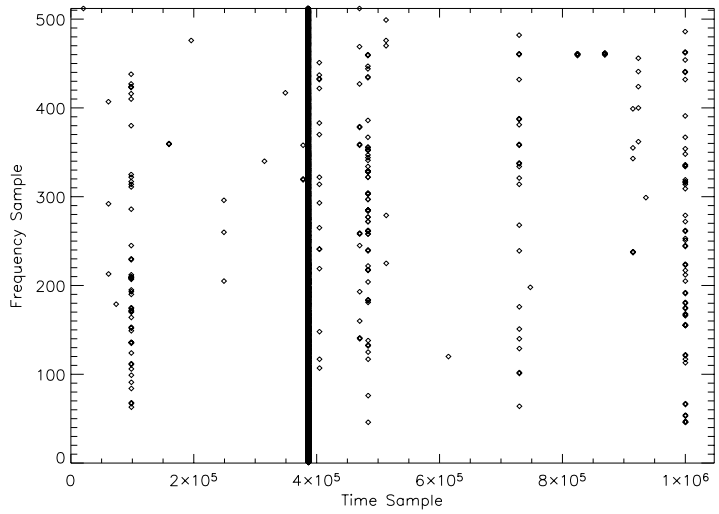

Figure 2.12: The thresholded dynamic spectrum of the Crab Pulsar, before dedispersion, reveals giant pulses as short, broadband, nearly vertical lines. The threshold is set to 10σ.

The Crab Pulsar: Threshold Test

We applied the threshold test to approximately 1 hour of data. Figure 2.12 shows the full dynamic spectrum with a threshold of $m = 10$, where m is the rms determined using the iterative approach described in Section 2.3. In Figure 2.13, we compare three different threshold values ($m = 3, 5, 10$) for a subset of the data spanning 0.5×10^5 128 μs samples (6.4 s).

The Crab Pulsar: Friends of Friends

We applied the Friends of Friends algorithm to the same Crab Pulsar data set described above using a threshold of $m = 5$ and a maximum separation $\delta s = 10$. Figure 2.14 shows all groups containing more than 30 samples. Different colors or greyscale levels are used to distinguish between the groups.

In Figure 2.15, we show a subset of the Friends of Friends output shown in

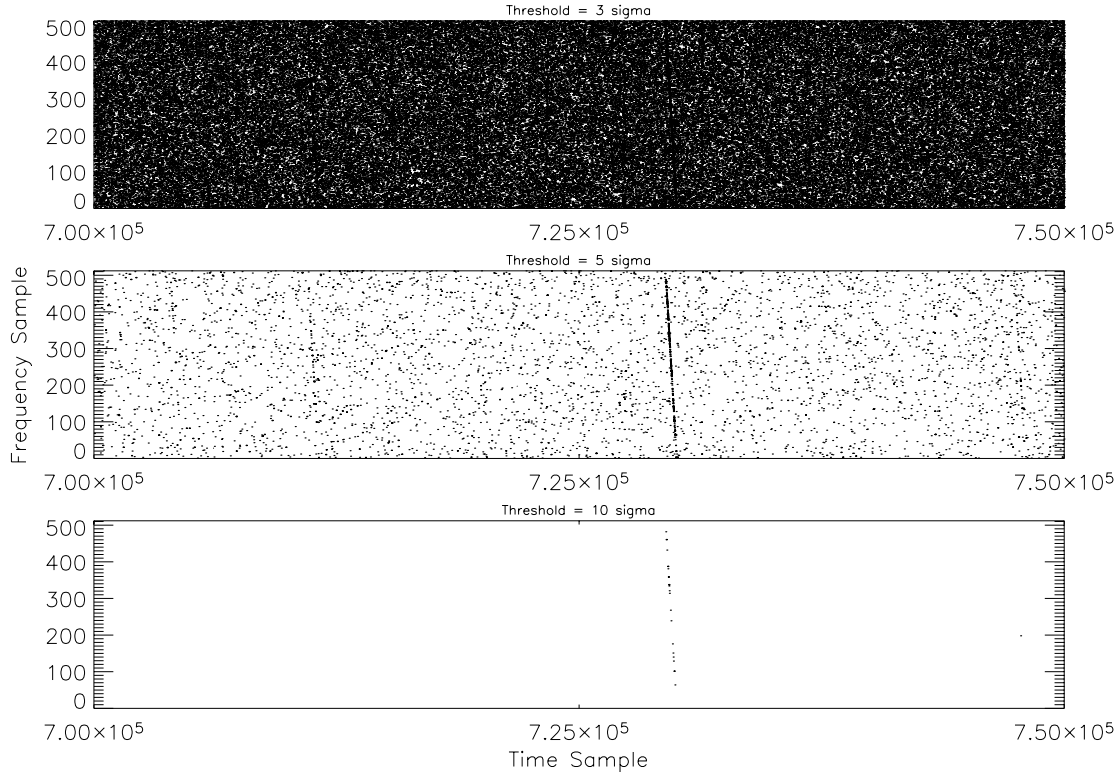

Figure 2.13: A 6.4 s subset of the dynamic spectrum in Figure 2.12 is subjected to (from top) 3, 5, and 10σ threshold tests. At 3σ, giant pulses cannot be discerned. At 5σ, a pulse is clearly visible. The pulse appears in only a fraction of frequency channels in the 10σ case.

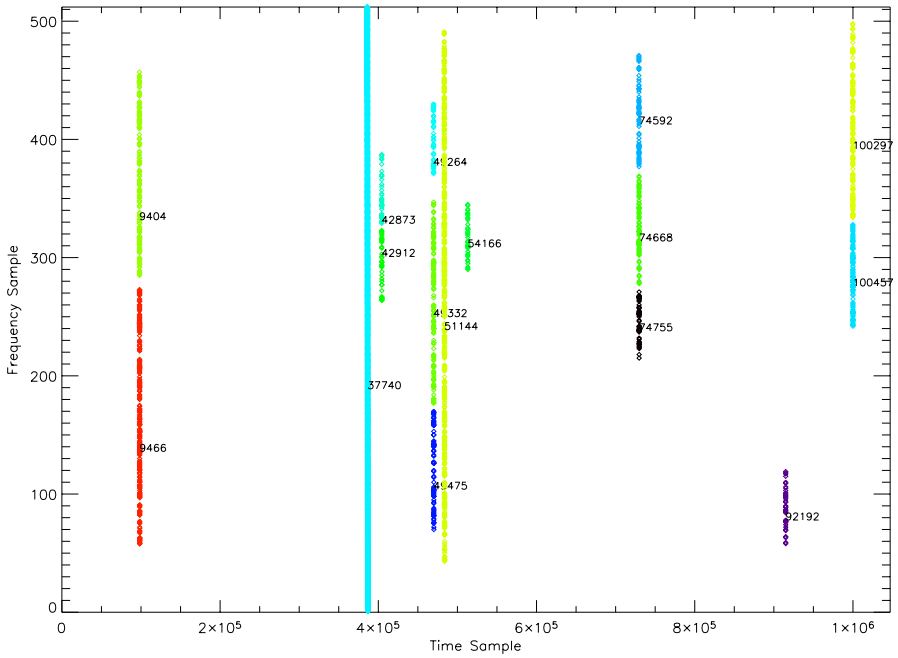

Figure 2.14: The Friends of Friends algorithm applied to an observation of the Crab Pulsar. Each group is annotated with an arbitrary numerical identifier.

Figure 2.14 spanning 3×10^4 samples ($\sim$ 4 seconds). Each group is labeled with a numerical identifier, its slope, and the uncertainty on the slope. The algorithm identifies four groups in this time span; three are actually part of one pulse, and the fourth group is the second pulse. As discussed in Section 2.4.3, the algorithm has a tendency to divide into multiple groups what is, to the eye, a single pulse. Future incarnations of the algorithm can address this problem by allowing a larger value of $\delta\nu$.

2.5 Matched Filtering

A third approach is matched filtering, in which a series of template filters is applied to data to identify regions of interest. The process is illustrated in Figure 2.16. The left panel a sample set of filter shapes and sizes that a routine might employ: filter length, width and slope are all variable. In principle, a filter could have any shape at all—it could be curved, discontinuous, etc. However, we restrict our discussion to filters of the kinds shown in the figure. Each filter moves throughout the data set until all possible regions have been sampled. A rough sketch of the motion of the "roving" filter is shown in the right panel of Figure 2.16

We use two different matched filtering routines; one for slopeless filters, and one for "chirped" filters. In the former routine, the filter size increases exponentially; that is, it begins with x and y dimenions both equal to 2 samples, then increases the y dimension to 4, 8, 16, etc. before varying the x dimension in an identical way.

We applied the zero-slope, matched-filtering algorithm to a 512×512 grid of computer-generated, normally distributed "noise." For each rectangular region, the mean within the filtered area, the mean outside the filtered area, and the

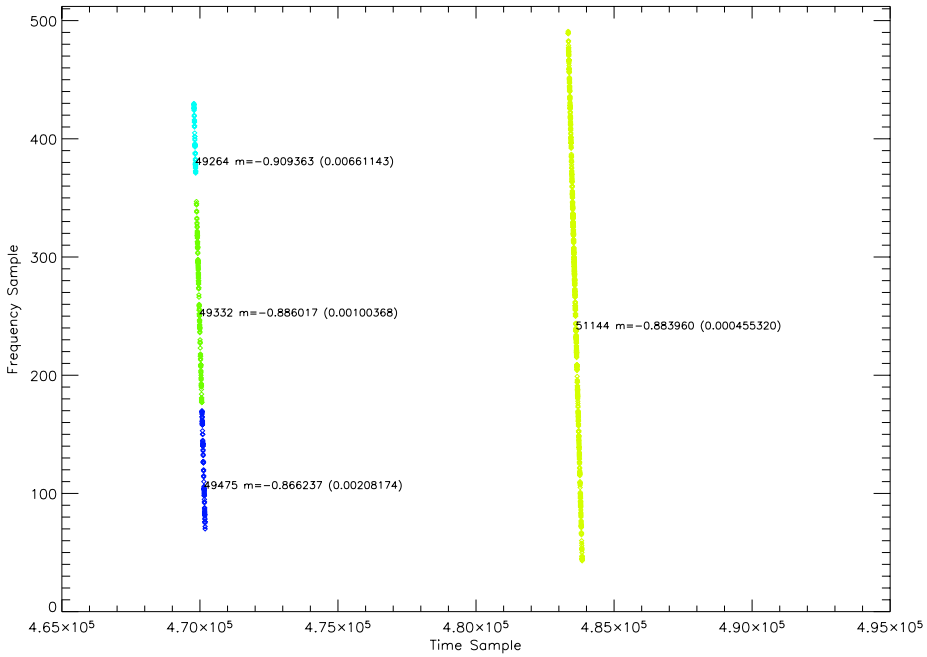

Figure 2.15: The Friends of Friends algorithm applied to a 4 second subset of the data in Figure 2.14. Each group is annotated with an arbitrary numerical identifier. The slope of each group and the uncertainty in the slope are also indicated.

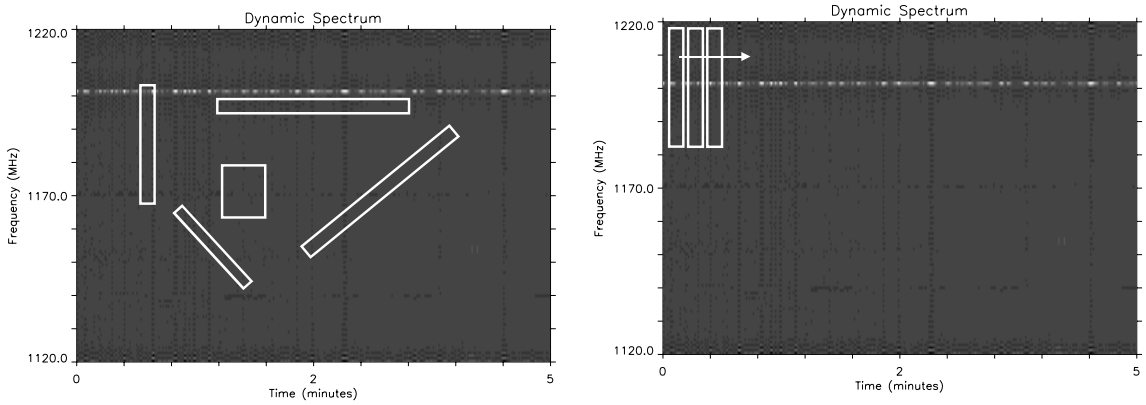

Figure 2.16: The matched filtering technique uses a series of filters, like those on the left, to find regions of interest within a data set. Each filter is moved across the data, as shown on the right.

fraction of samples within the filtered area above 3σ (fill factor) were calculated. The filter areas ranged from a minimum of four points to a maximum of 65,536 points.

Figure 2.17 compares the number of regions exceeding threshold ranges between 0 and 1.5σ for four filter regimes. Curve `a` includes all groups. Curves `b, c` and `d` include only regions containing fewer than 50, 10 and 5 samples, respectively. Note than an insignificant number of regions have a mean exceeding 1.5σ, where σ is defined as the standard deviation of the entire data set; no groups exceed a 3σ mean, suggesting that a 3σ threshold is appropriate for weeding out false positives due to radiometer noise for data sets of this size.

We then applied the same algorithm to identical noise overlaid with a 2×200 area of 10σ "signal."

Figure 2.18 shows the noise plus signal data set. The algorithm identifies the delineated area as the best matched filter on the basis of its fill factor, mean, and

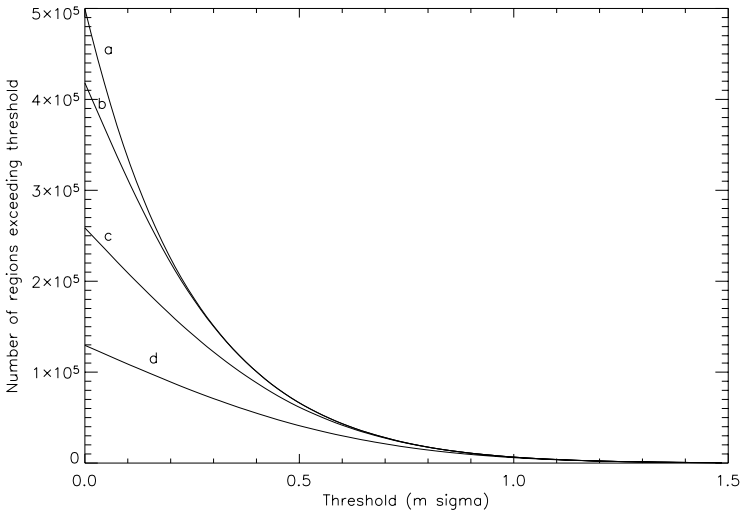

Figure 2.17: Number of filter regions exceeding thresholds from 0 to 1.5σ, of 998,988 regions total. Curve `a` includes all groups. Curves `b, c` and `d` include only regions covering fewer than 50, 10 and 5 samples, respectively.

size. When multiple regions have identical fill factors and means, as is the case here, we assume that the largest region is the best match. Statistics for the region are noted on the plot.

In the following plots, we compare the general statistics for the zero-slope algorithm applied to pure noise and noise plus signal.

In Figure 2.19, scatter plots show the relationship between the within-filter mean and the filter size for pure noise (left) and noise plus signal (right). The filter size (number of samples contained in the filter) is shown on the x-axis, and the mean for each region of that size is shown on the y-axis. On the righthand plot, as expected for normally distributed data, the most variation from zero is seen in small regions (the smallest regions contain four samples). The means of the largest regions show little deviation from zero. In the plot on the left, however, all but the largest filter sizes return groups with means $> 4\sigma$. The maximum mean