6 Research Description

We now describe three research projects that would be enabled by the funding of this Research Infrastructure proposal. These projects are extremely data-intensive, and require large-scale data storage and data manipulation capabilities. They also require a tight coupling of the storage and computational infrastructure in order to perform complex operations on the data sets.

We propose to use the resources from this grant to develop the required storage, computation, and networking infrastructure. A significant fraction of this infrastructure would be shared among these projects so that we achieve significant cost benefits over developing the infrastructure separately for each project. For instance, since online disk storage is relatively expensive compared to near-line or offline tapes, up to 300 TB of disk storage will be shared among the projects (this is in addition to the 100-150 TB of dedicated disk space for each project). This arrangement works well because each project has a temporary need for large amounts of online storage when converting raw data into its processed form, but has a reduced (although still large) storage requirement when operating on processed data. Similarly, the computational and networking infrastructure are shared among the different projects. This storage and computation architecture allows the possibility of adding other data-intensive research programs in the future.

We now detail each of the three projects, and we explain how the need for a petabyte storage service arises from the underlying research.

6.1 Physically Accurate Imagery

Researchers: Kavita Bala, Steve Marschner.

An important goal in computer graphics is *physically accurate rendering*: starting with a model of a three-dimensional scene, we need the ability to compute images that predict with quantitative accuracy what would be seen if that same scene were viewed from a particular vantage point in reality. This kind of rendering enables a wide set of applications where only very accurate renderings are useful. One such application is a virtual museum. Rare artifacts from around the world could be digitized, each at its own site, into appearance representations accurate enough to produce highly realistic views from any reasonable distance under any lighting. These objects could be placed into a single environment, where new display algorithms would allow a user to navigate in real time, giving a fully realistic, immersive experience of a museum collection that could never be assembled in reality. Other applications range from architectural lighting design to automotive design to visual effects for feature films.

Despite the remarkable progress in physically accurate rendering algorithms over the last two decades, many open problems remain. In particular, it is still not possible to accurately simulate the appearance of complex objects with real materials. For further progress, it is now critical to gain a better understanding of how light reflects from complex objects and structures occurring in the real world, and how to accurately and efficiently represent and render such objects for computer graphics. To this end, two major ongoing research endeavors at Cornell are (1) to use measurements of real materials to develop better reflectance models and (2) to explore new rendering approaches that can use measured or precomputed data to produce physically accurate renderings. Both of these research thrusts require a very large and scalable storage infrastructure to achieve their full potential.

As a specific example, Cornell researchers are developing very accurate techniques for measuring the *scattering function* of an object. The scattering function describes how the distribution of light reflected from the object depends on the distribution of light incident on the object. Accurately measuring the scattering function for real objects is crucial to physically accurate rendering because it contains all the relevant information about an object's interaction with light. However, accurately measuring the scattering function is a challenging task due to the following issues:

- The effort required to build an apparatus capable of accurate measurements.
- The time required to make measurements of the desired accuracy.
- The storage and computational infrastructure for storing and processing resulting data.

To address these issues, Cornell researchers are acquiring an instrument called the Spherical Gantry. The Spherical Gantry is a versatile four-axis motion system designed for optical scattering measurement. It is sufficiently general to provide the motion required to measure any part of the scattering function, which makes it unnecessary to build new apparatus for each project; this addresses the first issue above. The motion of the Spherical Gantry is synchronized so that measurements can be streamed out of a scientific digital video camera while the gantry is in motion. This is in contrast to most existing measurement apparatus, which work in a move-stop-capture mode and are one to two orders of magnitude slower. This addresses the second difficulty above.

The remaining impediment to accurately measuring the scattering function is the difficulty of storing and processing the data. At the data rates of present cameras, we may expect a *single dataset* to contain over 10^7 one-megapixel 16-bit images, or about 50 TB of uncompressed data. Storing, processing, and archiving this data requires tightly coupled storage and computational infrastructure that goes well beyond what we currently have available. If this proposal is funded, we expect to acquire the desired resources and measure the scattering function on a scale much larger than has ever been attempted before. By making this data publicly available through a Web service, we expect that it will benefit the graphics community as a whole.

Cornell is the ideal university to undertake such projects. The Program of Computer Graphics (PCG) has been at the forefront of rendering and light reflection research. Influential work in reflection modeling and measurement ranges from Cook and Torrance's ground-breaking work [CT82] to more recent work by He et al. [HTSG91], Lafortune et al. [LFTG97] and Marschner et al. [MWL$^+$99, MWLT00] that produced widely-used reflectance models and measurement techniques. In rendering, Cornell has been a pioneer in global illumination techniques [GTGB84, LSG94, SAG94, BWG03], including the work on radiosity that initiated the field of global illumination.

6.1.1 Background and Prior Research

The physical process of light reflecting from an object can be seen as an operator that maps an incident light distribution to a reflected light distribution. Each of these two distributions is defined over a space of rays: the incident distribution comprises light moving along every ray that enters the neighborhood of the object, and the reflected distribution comprises light moving along every ray that exits that neighborhood. Since the operator is linear, we can represent it by a function S on pairs of rays; this function is the *scattering function.* The incident and reflected ray spaces are each four-dimensional (two dimensions for direction and two for position), so the domain of S is eight-dimensional. S contains all the information that could ever be required to render the object (in the absence of wave optics effects) because it can be integrated to compute the appearance of the object from any view under any illumination.

Nearly every research project involving higher dimensional radiance representations or measurement of appearance can be framed as a scattering function computation under particular assumptions. For example, representations of the light reflected from an object, such as the

Light Field or Lumigraph [LH96, GGSC96b, WAA$^+$00, SH99, BDT99a, SK02, GGHS03], measure the reflected distribution as a function of the outgoing ray for a fixed illumination condition. Work on linear relighting [DHT$^+$00, ZWCS99, SKS02a] generally stores information about light reflected from an object under a range of (usually distant) illumination conditions, then combines those images to produce renderings under any (distant) illumination field. Work on *object capture* [MPN$^+$02, MPDW03, LKG$^+$03, Mar98] builds a more complete representation of an object's appearance by using images from many viewpoints under varying illumination but generally depends on known surface geometry. Finally, scattering models for surfaces, including BRDF models [CT82, HTSG91, Lar92, LFTG97, MWLT00], fiber scattering models [MJC$^+$03], and BSSRDF models [JMLH01] can be thought of as higher-level representations of the scattering function. The personnel of this proposal have been active contributors to this area, particularly to scattering models [MWL$^+$99, MWLT00], object capture [Mar98], and multidimensional shading representations [BDT99a, BDT99b, Bal99].

Because the full scattering function has an 8D domain, any approach based on sampling the whole domain, or even most of the dimensions of the domain, leads to enormous numbers of samples. For this reason, all the above methods have dealt with low-dimensional slices (e.g., fixing the viewpoint) or projections (e.g., fixing the illumination) or have introduced assumptions (e.g., distant illumination, single-surface reflection) to limit the precomputation time or the time to capture based on what can be acquired given the available apparatus and measurement time.

From an acquisition point of view, no previous work has been able to include enough dimensions to get a global view of the scattering function. If enough dimensions can be included, and particularly if the set of incoming rays can be made the same as the set of outgoing rays, we can examine the higher level structure of the scattering function, which is not visible in many projections and slices. For example, the scattering function is expected to be invariant with respect to interchanging the viewing direction and light direction (Helmholz reciprocity). We would like to capture data in high enough dimensions and at high enough resolution to begin to look at the broad structure of the scattering function.

6.1.2 Proposed Research: Accurately Measuring the Scattering Function

In order to support broader investigations into the scattering functions of real objects, researchers at Cornell are acquiring an instrument called the Spherical Gantry. The Spherical Gantry, which is expected to be in operation in early 2004, will provide very flexible control over measurement geometry and will also support rapid data acquisition.

An overview of the Spherical Gantry's mechanical design is shown in Figure 1. It consists of a platform, on which an object or sample of material may be placed, and two arms. The first arm is used to position a light detector or digital camera, while the second arm is used to position a light source or digital projector. The basic mechanical design is the same as a similar device built by the same manufacturer for the Stanford Graphics lab, allowing us to take advantage of that proven design while re-engineering many details, including all-new control electronics, with the benefit of experience with the earlier device.

The camera arm moves with two degrees of freedom: a rotation of the whole arm about a vertical axis at the base (axis 1 in the figure) and a rotation of the upper segment of the arm about a horizontal axis at the elbow (axis 2 in the figure). This gives the ability to position the camera at any point on a sphere centered in the object volume (save for a small area at the bottom where the camera would collide with the base). The light source arm moves with a single degree of freedom, rotating about a horizontal axis (axis 3 in the figure), and the object platform rotates about a

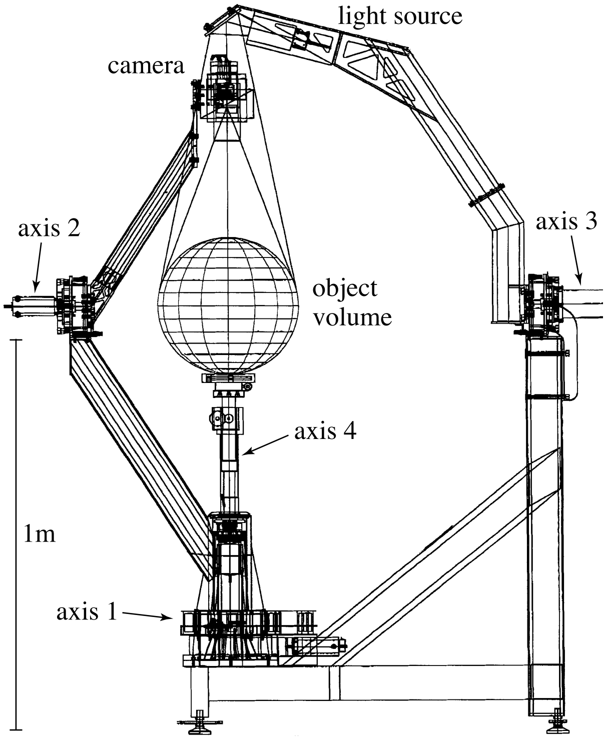

Figure 1: The mechanical design of the Spherical Gantry.

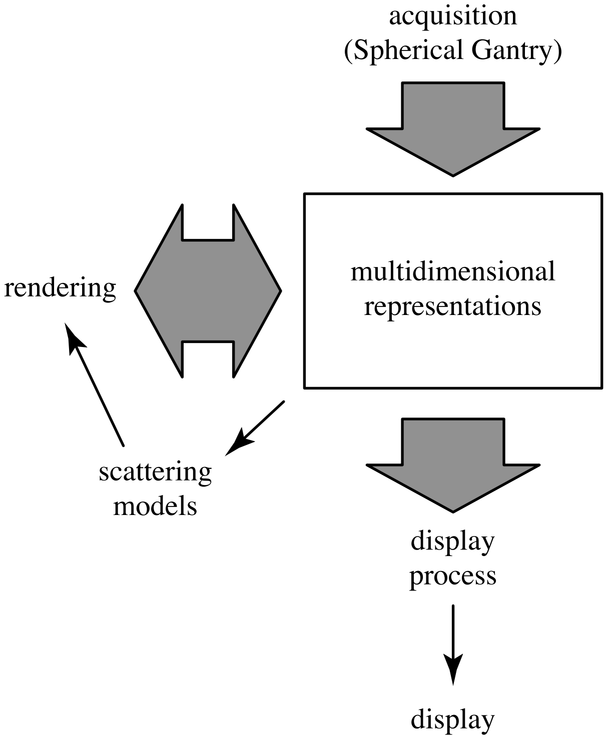

Figure 2: Multidimensional representations in rendering.

vertical axis (axis 4 in the figure). Together these two rotations effectively give the light source access to a sphere concentric with the camera sphere (save for a similar stay-out area around the base). Thus, the object or sample can be viewed from any direction while simultaneously being illuminated from any other direction. The arms of the Spherical Gantry are extremely rigid and under precise computer control, allowing images to be captured continuously by a high-resolution video camera while the gantry is in motion. If a digital projector is used as the light source, the gantry can illuminate and measure arbitrary rays, giving fully general ability to measure the scattering function.

We intend to capture and archive detailed scattering function datasets for dozens of objects with different kinds of optical properties. These datasets will make important contributions to several lines of work: development of scattering models, development of object capture methods for archival of unique objects, and development of multidimensional scattering representations and associated rendering methods.

Archaeologists, librarians, and others are interested in digitally archiving unique artifacts by recording their full three-dimensional appearance so that they can be examined without requiring access to the artifacts themselves. The proposed multidimensional datasets will be ideal for research into digital object capture methods that can be used in these applications, because most object capture methods work from a subset of the data we propose to capture. Thus our datasets will enable virtual testing of new capture methods (by taking subsets of the data) and rigorous validation of the results.

6.1.3 Proposed Research: Multidimensional Rendering Algorithms

Figure 2 illustrates the overall approach of our proposed research on multidimensional representations of the appearance of objects. Examples of multidimensional representations include the scattering function of an object, sampled radiance reflected from an object, and precomputed representations of energy transfer. We will create these representations either by acquiring them from real objects using the Spherical Gantry (as described above) or by computing them using offline rendering algorithms.

While there are existing multidimensional representations for rendering, these approaches have mostly focused on simpler problems due to the infeasible amount of storage required for the full high dimensional representation required for high-quality rendering. All these approaches consider single, stand-alone objects. Other constraints that are often imposed are pre-determined illumination [LH96, GGSC96a, BDT99a, WAA$^+$00] or lighting by distant environment maps [SKS02b, NRH03]. However, real scenes consist of multiple objects interacting with each other, for example by casting shadows on each other and reflecting light off each other. Our research will focus on novel representations and rendering algorithms for computing high-quality radiance including such global illumination effects.

We will explore several approaches to using multi-dimensional representations to compute global illumination in complex, dynamic scenes. One approach is to precompute the entire shading function for a scene, and display images at interactive rates by looking up the appropriate shading information for any given image. The shading function is a five dimensional function that varies at every location in space and in every direction. The precomputation of this function could be done using traditional rendering algorithms. However, we are also interested in developing algorithms that precompute 5D shading of complex scenes by using our previously acquired 8D scattering functions. Once the 5D shading function is precomputed, we will be able to generate high-quality imagery from arbitrary viewpoints of these complex scenes. Our research will explore unified representations that will not only be used for computing global illumination in scenes, but also used for displaying these rendering solutions at interactive rates.

Another important area of research is the support of dynamic scenes using multidimensional representations. Five-dimensional representations only capture the appearance of static scenes. When a user modifies the scene, for example, by moving an object around, or changing the lighting of the scene, the precomputed shading values of the scene must be updated. We will explore higher dimensional data structures that permit the rapid recomputation of appearance in dynamic scenes. Personnel of this proposal (Kavita Bala) has done preliminary research in 4D shading representations [BDT99b] and 5D representations to support dynamic scenes [BDT99a] in prior research.

Once these global illumination solutions are computed, they too will be added to the storage repository. The storage required for storing multidimensional representations depends on a variety of factors. The characteristics of the material properties of object's in a scene affect storage; glossier/shinier objects require more storage to represent shading because their appearance can vary significantly over the range of viewing directions (because of specular highlights). The range of viewing positions also affects storage; if a viewer is permitted to request extremely detailed views of a scene, the amount of storage required increases dramatically. Overall, for reasonable viewing parameters and object properties, the storage requirements can range up to several tens of Terabytes per scene. The computational resources required to precompute shading including global illumination effects range up to several thousands of processor-hours on current commodity CPUs.

6.1.4 Infrastructure Requirements and Usage

Capturing the measurements from the Spherical Gantry leads to large storage requirements. For instance, consider a measurement using a 1000x1000 resolution CCD camera that can capture images at around 30–60 Mbytes/s (for instance, a scientific camera or an HDTV camera), and a projector operating at 25x25 resolution. If we were to capture images at 300 positions each for the camera and projector (about 7 degree spacing over half the sphere), this will generate about 50 terabytes of data in a span of about 11 days *for a single object*. Storing, processing, and archiving this data requires tightly coupled storage and computational infrastructure that goes well beyond what we currently have available.

If this proposal is funded, we expect to acquire the desired resources and measure the scattering function on a scale much larger than has ever been attempted before. At capture time we will stream the data from the camera directly to disk, where it can be examined and verified (so that bad data, for instance due to the lamp in the source burning out, can be identified and re-captured) before it is compressed and then offloaded to tape storage. Previous work with related types of data has been forced to use lossy compression to store the data, which runs the risk of obscuring the very subtleties we are seeking to record. We would thus limit ourselves to lossless compression, saving perhaps a factor of 2 to 4 in archival storage space. However, for efficiency, we will keep the data that is to be streamed to the client uncompressed.

The large processing capabilities of modern processors also permit the precomputation of extremely large datasets of object appearance including global illumination effects. Besides the storage of the final precomputed objects, this also requires temporary storage during the process of generating the final data results. We expect tens of terabytes of storage to be required and thousands of GHz of processing to compute full global illumination solutions of complex scenes. Fast networking to stream this data to the computational engines would be required as well.

The data in the repository, including the measured datasets from the Spherical Gantry and the rendered datasets including global illumination, will be explored using a visual data exploration tool in a client-server architecture. The client component will run on a user's desktop and display a relatively small subset of the data, while the server component will be responsible for efficiently retrieving the appropriate subset and streaming it off disk and across the network to the client. The client code will also be made publicly available to the Graphics community as a whole so that researchers outside of Cornell can have access to the stored data.

6.2 The Structure and the Evolution of the Web

Researchers: William Arms, Daniel Huttenlocher, Jon Kleinberg, Jayavel Shanmugasundaram.

Cornell is a major center of research about the content and structure of the information on the Web. This research includes algorithms for understanding the structure and evolution of the Web (Jon Kleinberg, Daniel Huttenlocher), studies of the Deep Web (Jayavel Shanmugasundaram), and the related areas of scientific publishing and digital libraries (William Arms). A theme that runs through much of this research is how to measure and model the evolution of the Web, and in particular how to relate observations made on the Web itself against information obtained from usage data.

As a source of Web data, we have a fruitful partnership with the Internet Archive (Brewster Kahle) and with their Web crawling partners at the University of California Santa Cruz (Raymie Stata). The Internet Archive has snapshots of the Web dating back to 1996 as well as access to multiple crawlers for different crawls. The entire collection is about 500TB, is growing at about 10TB per month, and contains about 13 billion pages. Currently our access to this data is limited either to small snapshots (a few TB) or to use of computing resources at the Internet Archive. Those computing facilities were not designed for large-scale analyses of the content.

In addition, we have existing research ties with several large high traffic Web information resources including CiteSeer at the Pennsylvania State University [GBL98], and three prominent sites hosted at Cornell: the Legal Information Institute (search for legal information on Google) [Mar00], the physics ePrint arXiv (search for ePrint on Google) [Gin01], and the Laboratory of Ornithology (search for birds on Google). Each of these sites provides a trove of usage data, and great domain expertise. A common pattern of research is to use domain knowledge to develop hypotheses that can be tested on these data sets, before being extended to Web-scale experiments.

We propose to develop a large data store with gigabit speed connections to substantial computational resources so as to test and refine our algorithms on Web scale data. In working closely with the providers of production Web services such as the Internet Archive, the ePrint arXiv and Legal Information Institute, we have seen repeatedly that the infrastructure necessary to provide a reliable service to millions of users differs substantially from that needed to run large-scale experiments on terabytes of data. To the extent that we have run research code on machines co-located with the production environment, it has been necessary to be careful not to perform large compute or I/O intensive operations that interfere with the service (and at times despite our care our jobs have been killed by the system operators because of such interference). Thus a large data store tightly coupled with high capacity computing specifically for research is of fundamental importance to advancing our research.

6.2.1 Background and Prior Research

One of the central themes of Web research at Cornell is that of "Web dynamics" which encompasses both change in the Web itself and information about usage of the Web, gained through server logs at high traffic Web sites. Most current techniques for analyzing Web structure and content are based on static snapshots of the content of pages retrieved by a Web crawler (see for example [KL01, BB98, LG98, LG99, ABJ99, BKM$^+$00, BAJ99, FFF99, HA99, KRRT99]). Yet the Web itself is increasingly dynamic, in its content, its structure and its usage. For instance, Weblogs and news sites are major sources of highly dynamic content and usage patterns. While there are both commercial tools and research results that apply to such dynamic content, we have barely begun to be able to analyze the dynamics and evolution of the Web.

We are working on techniques to identify and analyze rapidly changing content on the Web, e.g., detecting hot topics and determining the evolution of topics over time. Such techniques are currently at a very primitive stage, yet their potential is enormous. In addition to identifying emerging topics, such methods can highlight portions of the Web that are undergoing rapid change at any point in time, to archive and summarize the Web content surrounding a fast-breaking news story, and to provide a means of structuring the content of emerging media like Weblogs. In contrast tools such as Google News are based on frequent crawls of a manually selected set of news sites rather than automatically identifying trends. A primitive form of trend spotting is provided on the Weblog DayPop, which draws on Kleinberg's research on burst detection [Kle02].

In order to illustrate other aspects of the Web that can be illuminated through the study of change over time, we briefly summarize some results from our recent work in this area [Kle02]. In this work, we evaluate the popularity over time of items available for download in the Internet Archive Media Collection. One particular popularity measure that we consider is the batting average, which is the fraction of visits to a page describing the item that in turn lead to the item being downloaded. For instance a batting average of .500 indicates that half the people who visited

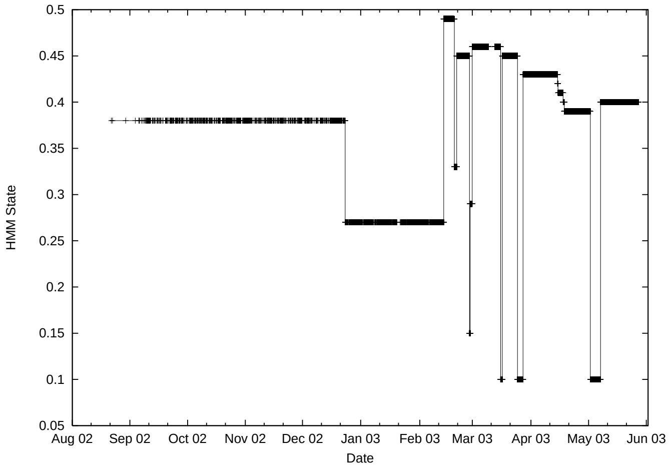

Figure 3: Tracking the batting average of *What You Should Know About Biological Warfare* as a function of time, using state transitions in a hidden Markov model.

the item description chose to download the item. We have developed a stochastic model of the batting average that estimates the most likely instantaneous batting average for an item, given a sequence of visits from the Web server log. One of these plots is shown in Figure 3, for the Internet Archive's online copy of the 1952 civil defense film *What You Should Know About Biological Warfare* and usage between November 2002 and April 2003.

Roughly, we see that the batting average begins at a high level (around .38), drops to a lower level (around .26), and returns to a higher level again (between .40 and .50), with the final higher period interrupted by five brief, sharp drops. Annotating these transitions in terms of events both on and off the site, a clear picture of the item's history emerges. The initial drop to a lower level in December 2002 occurred when the item was added to the Pick List, an unannotated list of (recommended) titles on a top-level page at the Internet Archive. The subsequent return to a higher level in February 2003 occurred when the item was moved (a week after Colin Powell's testimony on biological weapons to the UN Security Council) from the Pick List to the Collection Spotlight, a more extensive means of highlighting in which the title is accompanied by a brief description. Visitors arriving at the film's description from the Collection Spotlight were more likely to download it than visitors arriving from the less informative Pick List. Each of the five subsequent sharp drops can also be closely associated in time with an event involving the item. The first coincided with a referring link from the discussion site forums.somethingawful.com and the second with a referring link from the extremely active Weblog http://www.reason.com/hitandrun; in each case, the fraction of visitors arriving over these links who actually downloaded the film was very low. After the traffic from each of these referrers subsided, the batting average jumped back up. The final three drops correspond to technical failures on the site of varying lengths, which made it impossible to download the file.

This example illustrates several things. First and foremost, significant changes in the batting average for an active item are often correlated with "real-world events", both on and off the site, in which this item is featured. Second, the batting average undergoes sudden "discontinuous" changes in response to such external events. For instance the addition of a link occurs at some discrete point in time, and if it comes from a high traffic site, it causes a sudden and precipitous change in the instantaneous measure of item popularity. Third, significant events that were not necessarily apparent can be discovered after the fact through the effect that they have on items' batting averages.

6.2.2 Proposed Research: Evolution of the Web

Our goal is to develop techniques for simultaneously studying the evolution of the full Web and of specific, highly visible Web sites. As mentioned above, our previous work on the Internet Archive enabled us to detect things that are not available from static snapshots of the Web alone. Specifically, we were able to find periods where there is rapid change in usage patterns and to align those changes with external events. While we believe such analyses of usage data hold substantial promise, they do not even begin to explain more fundamental questions such as how the Media Collection at the Internet Archive grew from its initial introduction in September 2002 to receiving millions of visits a month over a matter of just a few months; this kind of visibility on the Web is not easy to achieve. We would like to be analyze what people were saying about the site, how its visibility increased, and examine the kinds of paths that lead to it. These are all questions that would be addressed by a combination of analyzing the Web graph, as it changes over time, together with usage data for the site.

Cornell is extremely well positioned to examine these kinds of questions on large Web datasets, but is currently limited by the availability of large disk storage coupled with the fast computing power necessary to run our algorithms on large data sets. We need access to a series of full snapshots of the Web, together with usage data from high-traffic Web sites (for a total of several hundreds of terabytes). With solely a series of Web snapshots, we are inferring people's behavior from just static links and text, which is questionable. With just usage data, we are only seeing the "last link" in many trails that lead to pages on a site such as the Internet Archive Media Collection. With both together, we can ask questions such as when did the Internet Archive receive news coverage, vs. when did traffic to the site change? What discussions were taking place in reference to the Internet Archive on the Web as a whole over time?

More generally, our past work has shown that information at large sites and its usage go through periods of discontinuous change, with more gradual behavior in between. Is there a meaningful sense in which this is true for the Web as a whole? Does the whole Web have seasonal, tidal variations? Currently these kinds of questions cannot be answered, and they cannot be answered without the ability to work with many large Web snapshots. In fact, much analysis of Web structure views hyperlinks as a kind of proxy for usage. If there is a link then it is implicitly assumed that people travel that link. This is a long-standing and untested assumption, and it's untested precisely because it's hard to look at both the link structure and usage simultaneously without access to multiple Web snapshots and usage data simultaneously.

6.2.3 Proposed Research: Web Measurement and Models

In the past few years, the Web research community (including researchers at Cornell) has produced measurements of some of the fundamental parameters of the Web, including its size, diameter and component structure, and distribution of node degrees. While many of these measurements have offered the first glimpses into basic structural features of the Web, there is need for a more solid analytical foundation to understand the generality and applicability of the results. Open questions range from the sampling methodology, in the case of size and diameter estimation, to inferring accurate functional forms, in the case of degree distributions. In this latter case, while it is easy to infer a heavy tail it is a much more subtle matter to distinguish power laws from closely related distributions. More generally, we need a method for separating the measured properties of a network from artifacts of the measuring process.

To illustrate the difficulties, we mention the “bow-tie” studies of Web structure, showing that the Web topology has a giant, strongly connected component, with roughly comparable numbers of nodes upstream and downstream from this component [BKM$^+$00]. Due to the enormous computational effort involved in this study, to our knowledge, it has been reproduced only once. Since both studies were done using an AltaVista crawl of the Web, it is crucial to understand if we can establish, in some precise manner, that the results of the study are an intrinsic property of the Web and not a consequence of heuristics used by the crawling software. We know that the crawl was smaller than the full Web; presumably a large fraction of the missing pages reside in the region upstream of the giant strongly connected component, since these are by far the hardest pages for Web indexing tools to gather. More subtle effects are possible too, as in a striking recent example of a graph traversal process on random graphs that can lead to incorrect inference of the degree distribution [LBCX03].

Our approach is to develop a framework for formal reasoning about the process of collecting Web measurements. We will use this framework, including precise models of the measurement process, to distinguish formally properties of networks that are relatively measurement-independent from those that are influenced by the method of measurement. Such a framework allows reasoning about network models at a more principled level. The development of this theoretical approach will proceed in conjunction with experiments on many large Web snapshots. By understanding the extent to which multiple views of the Web differ, we can better capture the inherent variability required of accurate Web graph models.

6.2.4 Infrastructure Requirements and Usage

The large storage requirements for this project stem from the fact that we propose to study the evolution of the Web. Thus, we need to store not just a single snapshot, but many different snapshots of the Web. A typical crawl of one billion pages contains over 25 billion hyperlinks, which take 1TB when stored naively and 100GB when stored with some sophistication. Even at 100GB, with snapshots over reasonably long periods of time such as 2-3 years, this requires several tens to hundreds of TB of data for multiple crawls of the Web. And this is not counting the space requirements for usage logs, which will require tens of TB of additional storage. We propose to develop this storage infrastructure using resources from this grant.

Web research also requires computational power that is tightly coupled with the storage infrastructure. Currently, almost all Web crawlers and search engines, and most Web researchers use collections of small Intel computers. These are a cost-effective way to assemble huge amounts of data and provide services to large numbers of users, but restricted in their ability to analyze the huge data sets. Moreover, the most powerful tools for coping with large volumes of information are accessible only to those with computing sophistication. For instance, to extract data from the Internet Archive, with its 400 parallel computers, is challenging, even for computer scientists.

Feature extraction is another little-appreciated aspect of Web research. Research is not done directly on the unparsed objects that come from the Web, but on features such as term sequences, term vectors, and hyperlinks [LPSW01, SBM00]. Given the vagaries of the Web, writing high-quality, reliable code to extract features is a non-trivial programming activity; extracting features from large collections requires massively parallel execution. Further, it requires large amounts (tens to hundreds of TB) of temporary space during computation.

Finally, feature extraction and analysis applications require high-speed network connections that will enable efficient computation at servers. High-speed network connections are also crucial for shipping relevant data sets to clients for further processing.

6.3 Large-scale Astronomical Surveys Using the Arecibo Telescope

Researchers: Alan Demers, Jim Cordes, Johannes Gehrke, Jayavel Shanmugasundaram.

The Arecibo Observatory (http://www.naic.edu) is operated and managed by Cornell University via the National Astronomy and Ionosphere Center under a cooperative agreement with the NSF. The Arecibo telescope is the world's largest radio telescope in terms of collecting area and thus can conduct the most sensitive surveys for pointlike objects. A multibeam feed system ALFA (Arecibo L-band Feed Array) is now under construction and will be installed by mid-2004. The feed array consists of seven dual-polarization horn feeds that provide 14 simultaneous data streams. The seven feeds point in different, but nearly contiguous directions, allowing a 7-fold speedup of the surveying rate. More details about ALFA can be found at http://alfa.naic.edu.

Cornell faculty are heavily involved in the ALFA project and serve as chairs of three scientific consortia that have formed, with international participation, for the main science areas: Galactic science, Extragalactic science, and Pulsar science. Jim Cordes is the chair of the Pulsar-ALFA consortium, which is planning a massive survey for pulsars, the P-ALFA surveys. Pulsars are neutron stars, one of the endpoints of stellar evolution, which are about 10 km in radius and have 1.4 times the mass of the sun. They are, in many ways, giant nuclei.

The pulsar surveys will be the deepest (reaching to the greatest distances) ever undertaken and are expected to yield not only a large number of new pulsars ($\sim$ 1000) but also exotic objects, including millisecond pulsars spinning near the break-up speed of a neutron star; neutron stars in compact binaries with orbital periods of a few hours or less, and companion stars that are other neutron stars or black holes; and neutron stars moving rapidly owing to their birth in off-center supernova explosions. These discoveries are expected to provide numerous opportunities for followup research on the equation of state of nuclear matter, gravitation physics, and gravitational waves.

The data storage requirements for the P-ALFA surveys are very large ("astronomical"). The raw data from the P-ALFA surveys will amount to about 800 Tetabytes, and it will take three to five years to acquire this data.

6.3.1 Background and Prior Research

Pulsar surveys not only discover new pulsars, but in particular aim to find exotic objects of interest, including millisecond pulsars with spin periods as fast as 1.6ms and potentially even shorter periods, and compact binaries with orbital periods of a few hours or less. For instance, compact binaries have sufficiently strong gravity that non-Newtonian gravitational theories (such as Einstein's General Theory of Relativity) are needed to account for the orbits, which are precision monitored through continued "timing" observations of a discovered pulsar.

Thus, pulsar research provides unique and important opportunities for studying extreme states of nuclear matter, gravitation physics, and gravitational waves. As an example of the physics payoff, a binary pulsar discovered in 1974 at Arecibo by Hulse and Taylor was used to infer the existence of gravitational waves in accordance with Einstein's General Theory of Relativity [HT75]. The Hulse-Taylor binary consists of two neutron stars in an 8-hour orbit with orbital separation smaller than the Sun. For this work, Hulse and Taylor were awarded the Nobel Prize in physics in 1993.

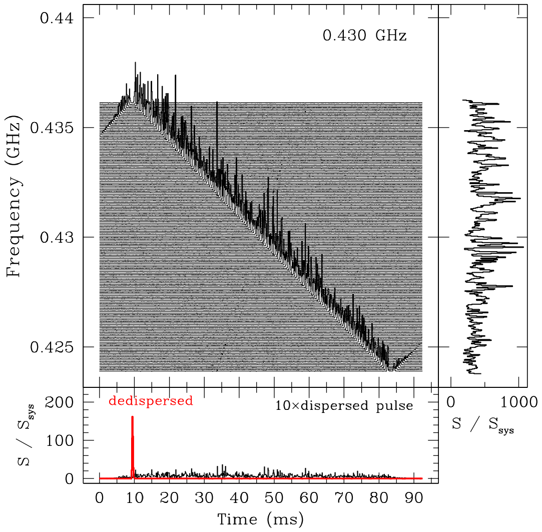

Figure 4: A single pulse from the Crab pulsar in the Crab Nebula.

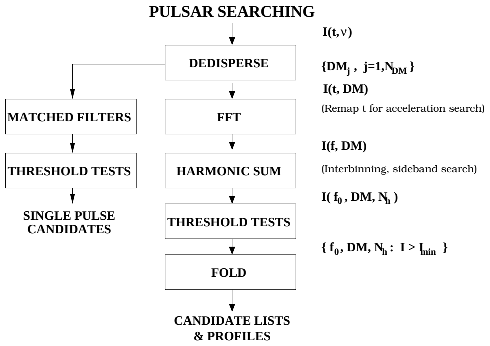

Figure 5: Flow diagram for a pulsar search algorithm

Pulsar search procedures have evolved over the past 30 years according to our growing understanding of astrophysics and populations of compact objects such as neutron stars and black holes, but also following the growth in computational and data storage capacity. To process the raw data, we apply a filter with three parameters: (a) the temporal period; (b) the dispersion measure, which quantifies how pulses arrive differentially at different frequencies; and (c) the pulse duty cycle, which is the pulse width divided by pulse period. Figure 4 shows a single pulse from a strong pulsar in the Crab Nebula and the differential arrival times owing to dispersion. A trace is shown for each of 256 frequency channels. The slope of the locus describing the arrival time of the pulse in each channel is proportional to the reciprocal of the "dispersion measure" (DM), a quantity that is unique to each line of sight (the endpoints of the pulse show positive slopes owing to aliasing in the sampling system). The bottom panel in Figure 4 shows the sum over frequency channels both with (”dedispersed”) and without (”dispersed”) compensation for the dispersion effect.

The main problem in pulsar surveys is that we do not know the parameters of the filter in advance. Thus, existing algorithms *search* a large space of possible parameters to find plausible settings that could mark the data as a candidate for a celestial object.

New pulsars are found largely through periodicity searches of pulse trains that are far weaker than the pulse shown. Single pulses are usually buried in noise, so a coherent analysis must be applied to long time series to bring out any periodic signal. While most data sets produce multiple candidate signals, the great majority of these are spurious signals resulting from locally generated radio frequency interference (RFI), one of the main problems of detecting real celestial signals. In order to recognize and remove RFI, we need to perform an analysis of the data products from the entire sky, taking into account that RFI signals repeat (but are episodic), while a single celestial signal should appear from only one direction on the sky. Due to the small size of existing pulsar surveys, current searches for filters could proceed brute-force without any sophisticated search strategies and high-performance algorithm implementations.

As preparation for full ALFA surveys, we have done a pilot survey of one square degree of sky using a single-beam system at Arecibo. The project has made use of pilot-Arecibo data to construct a data base and develop analysis tools for investigating candidate signals from pulsars that will warrant reobservation. Reobservations are needed to robustly confirm the reality of pulsars versus artificial radio frequency interference, which is a growing problem that requires new filtering algorithms.

6.3.2 Proposed Research: The P-ALFA Surveys

The proposed surveys for pulsars include (a) mapping of the entire Galactic plane of the Milky Way visible with Arecibo (within ± 5 deg of the midplane) where the greatest concentration of pulsars is expected; this represents a swath of sky equal to about 500 square deg, and (b) searching further out of the Galactic plane (e.g. ± 15 deg) in a shallower survey to find millisecond pulsars and binary pulsars that are expected to have a thicker population distribution than the bulk of pulsars.

Why search this much sky? The survey payoff comes in two forms. First, the number of detected rare, exotic objects scales with the expected total number (N) of new pulsar discoveries. Based on previous surveys, having N = 1000 implies that we will find a handful of exotic binaries, etc. that will provide the greatest payoff in basic physics. Second, the total sample can be used to map the ionized gas and magnetic field of the Galaxy. Pulsar signals carry information about intervening ionized and magnetized gas because the signals are distorted by the consequent index of refraction of the gas. Techniques are well known for measuring these distortions and thus quantifying the integrated (along the path) electron density and product of electron density and line-of-sight component of magnetic field. The basic effects are known as dispersion arrival times and Faraday rotation. Generating tomography-based maps of gas properties is much more challenging than making the basic measurements for each line of sight. Examples to date may be found in [TC93], [CL02] and [WCK$^+$03].

The raw data will be subjected to a novel multistep data analysis for identifying real celestial signals with expected (and hopefully new) signatures. The multiplicity of signals from the P-ALFA surveys drives data rates upwards and provides the opportunity for applying more intricate filtering algorithms that distinguish celestial signals from locally generated RFI. To realize the great potential of the ALFA system, new data mining algorithms are needed for tackling RFI and for detecting celestial objects of greatest interest.

The 14 signals from the "front-end" feed system (ALFA) will be signal processed in real time using a fast digital spectrometer that provides the basic "dynamic spectrum" $I(t,\nu)$, intensity vs. time and radio frequency, for each of the seven feed antennas. (The two polarization channels will be summed in real time, yielding seven distinct dynamic spectra.) Prior to the search analysis, the raw data will be filtered to remove radio frequency interference, which is expected to be nearly identical in all seven beams, whereas celestial signals will appear in at most 3 beams. As shown in Figure 5, each of the seven dynamic spectra is subjected to a detailed Fourier and statistical analysis whose results are winnowed to provide candidate periodic pulsar signals and candidate single pulse events. The dynamic spectra for each telescope pointing comprise a size equal to 7 beams $\times$ 1024 frequency channels $\times$ 10^7 time samples (approximately, since survey dwell times are now being optimized). The time duration for each pointing will be $\sim$ 300s. Each dynamic spectrum will be searched for about 10^3 separate values of dispersion measure, yielding $7{\times}10^3 \times 10^7 \times 32$ separate statistical tests, where the 32 factor is the number of separate harmonic sums that are calculated, or $\sim 2 \times 10^{12}$ separate tests. Thresholds are set so that, per beam, we will have only 10 to 10^2 events above threshold that require further investigation. *Subsequent to* the search analysis performed on each beam, candidate signals found in the 7 beams will be cross compared to remove low-level RFI that was missed previously, thus improving the quality of candidate lists. Additional complexity of the analysis involves optimization against potential motion of a pulsar in an orbit. To do so requires remapping the time series according to trial *accelerations* using typically about 200 trial values. Single-pulse candidate searches represent about 10% of the computations of the periodicity searches but add significantly to data products that need to be investigated in the meta analysis of all telescope pointings. Altogether, about 25,000 telescope pointings will be made to cover the Galactic plane with a similar number to cover portions of sky outside the plane. With the 300 s dwell time per pointing, it will take $\sim$ 2100 hr and $\sim$ 800 TB to acquire and store the raw data for the Galactic plane search.

From the above discussion, one can see that *efficient* analysis and processing of raw data at this scale is a very challenging *data mining* problem. With recent NSF funding, the PIs have developed some of the fastest known data mining algorithms for association rule mining and classification tree construction [GRG98, GGRL99, BCG01, AGYF02, BGKW02, DG02, KGBW03]. This software is available at http://himalaya-tools.sourceforge.net. We believe that data mining algorithms can reduce the filter parameter search time for two reasons. First, initial investigations of the search space have shown that it has structural properties that allow us to re-use techniques from existing efficient search algorithms, such as search algorithms for association rules. Second, we can formulate the prediction of filter parameters as a learning problem, and we can use data mining techniques to a-priori significantly reduce the search space.

Expected Results: One Example. Recall the discovery of the Hulse-Taylor binary, which consists of two neutron stars in an 8-hour orbit. In the last 30 years, only four additional binaries have been found because they are scarce and require a great deal of telescope and computer time. The orbits of double neutron star binaries (and any neutron-star/black-hole binaries) are shrinking because of energy losses associated with the gravitational waves. Consequently, we expect the Galaxy to contain binaries with orbital periods *much smaller than 8 hours*. Short-period binaries are difficult to detect because traditional search algorithms rely on the Fourier analysis of equally-spaced (or nearly so) pulses (see [Cor02] for an overview); Doppler shifts from orbital motion reduce the sensitivity of such searches unless compensating, computationally intensive algorithms are used [RCE03]. Processor speeds now are just beginning to make comprehensive searches for binaries feasible. We expect that important new binary systems will be found in ALFA surveys that will be far richer than the Hulse-Taylor binary in providing opportunities for testing gravitational physics.

6.3.3 Infrastructure Requirements and Usage

We currently do not have the capability for providing access to the raw data with the requisite disk and network capacities. This is a new venture for pulsar astronomy and cannot be funded with resources of the Arecibo Observatory. Accomplishing the overall program depends upon our ability at Cornell to gather the needed resources. Without the infrastructure, the project will have to be scaled back, with the consequence that the survey yield would be much smaller and the likelihood of finding rare compact binary pulsars, including pulsars with a black-hole companion, submillisecond pulsars, and high-velocity pulsars would be small.

We need a petabyte of total storage, a combination of disk (for the storage and interactive analysis of the derived data) and tape (for the storage of the raw data — with the capability to stage raw data from tape to disk to perform a new analysis). Local processing of the raw and derived data requires a powerful server that is tightly integrated with the storage resources.

Cornell is an excellent location for processing the raw data and consolidating the data products. The data needs to be accessible to Cornell faculty in the Department of Astronomy who manage the pulsar survey, and to faculty in the Department of Computer Science who will develop the data management and analysis infrastructure, and will perform research on new data mining algorithms. Cornell is also far more accessible from a networking point of view than the Arecibo Observatory in Puerto Rico, and this allows Cornell researchers and their international collaborators to interactively access the raw data and derived data.

Since the raw and processed data collected using the P-ALFA survey is expected to be of considerable value to astronomers and astrophysicists world-wide, we propose to provide a service-oriented interface to the data that allows users from all over the world to access the data, and to develop tools that will permit online meta-analysis of the results of the 25,000 telescope pointings for identifying candidate pulsar signals and potentially new types of signals. The Web-based set of tools we anticipate will allow investigators to interactively query the multidimensional search space and to allow interactive and efficient exploration of the data set.

Although the P-ALFA database project is developed for a particular application, the ideas should extend to similar projects. For example, the other ALFA surveys also involve large, multivariate data sets that require large-scale analysis.

6.4 Conclusion

In summary, the development of a large storage infrastructure and associated high-bandwidth network will enable us to pursue scientific research at a level that has not been attempted before due to demanding infrastructure requirements. Specifically, it will enable (a) highly accurate measurement and rendering in Computer Graphics, which is expected to lead to new insights on light reflection and rendering of complex structures and materials, (b) a large-scale and principled study of the evolution of the Web, and the development of new models of the Web, and (c) one of the largest pulsar surveys ever undertaken, which is expected to lead to new discoveries about the fundamental properties of matter. All the data sets generated as part of these projects will be made available to the general scientific community. Furthermore, the developed storage and computation architecture will allow the possibility of adding other data-intensive research programs in the future.