The Birth of the Universe and the Fate of the Earth:
$10^{12}$ UV Photons Meet Stan

1. Introduction

Astrophysics benefits from community-driven large sky surveys that generate vast homogeneously-observed data sets that serve a diverse range of scientific goals. Important, unanticipated discoveries often arise well after a survey is completed. Big survey data sets are themselves made more powerful over time through improvements in calibration, data mining and statistical analysis techniques, along with the creative ingenuity of new generations of scientists. Here we confront one large data set—every photon observed and recorded by the Galaxy Evolution Explorer (GALEX) ultraviolet space telescope (Martin et al. 2005)—with two new scientific questions that have risen to urgency only over the past couple of years and new statistical tools that are only now available to the community.

GALEX was launched in 2003 and operated for 10 years, collecting several trillion photons as part of a nested series of sky surveys. Its main mission was to address forefront questions in galaxy evolution: When and how are stars and galaxies formed in the Universe? Unlike conventional imagers, the detectors aboard GALEX record information for every incident photon (time, position and meta-data) which are then used in the construction of sky images. This data set has been used for time-domain studies of transient and periodic phenomena, including surprising last flares from stars falling into supermassive black holes.

In this era of exoplanet discovery, it has become evident that the time-domain data set of UV photometry from GALEX may itself contain signatures of planets, particularly around compact UV-luminous objects such white dwarfs. Stars such as the Sun will eventually reach this stage: What will happen to the Earth and other planets when the Sun becomes a white dwarf? With the GALEX photon data set, we can measure the distribution function of planets around white dwarfs and begin to answer this question. Even more recently, astronomers have discovered a faint polarization signal in the Cosmic Microwave Background, a possible signature of the inflationary phase of the early universe. Is this signal a true measurement of inflation or a poorly understood signature of Galactic dust? The GALEX all-sky photon data set may also hold clues as to the origin of this signal. Such recent advances motivate a novel, time-critical investigation into the GALEX large survey data set.

In the next two sections we discuss how we will address our scientific questions using a novel probabilistic modeling tool, Stan (Stan Development Team 2014). Stan is the only available tool designed to perform efficient hierarchical (multi-level) Bayesian inference (Hoffman & Gelman 2011) using a Hamiltonian Monte-Carlo algorithm, in a flexible framework that is well matched to the scientific questions posed here. Our data set contains more than $10^{12}$ photons observed over $10^8$ seconds, each of which could be individually modeled, an
intractable problem for any existing Bayesian inference tool. The process by which astrophysical data sets are typically studied, involving successive steps of “calibration”, “reduction” and “analysis” treat this problem by making implicit—though often piecemeal—assumptions about the correlation between system model parameters and the sources of variance. In hierarchical Bayesian modeling we can explicitly treat these assumptions in order to reduce the dimensionality of the problem (allowing us to e.g., “pool” our data) while also benefitting from the improved inference that can result. It is hard to overstate the importance of tools that allow efficient, iterative exploration of weakly- and strongly-correlated regions of parameter space.

Our proposed investigation is ideally suited to this IDSE call for several crucial reasons: 1) multi-level modeling applied to large, self-calibrating, observational data sets is a challenging endeavor, with broad applicability across the natural and data sciences; 2) this program will develop a set of statistically powerful astronomical analysis tools using Stan that will provide us with a strong competitive advantage in future astronomical observing and funding proposals; and 3) the short fuse on funding means that we can carry out our investigation in a time-frame that can have an impact on a rapidly developing field. While the scientific questions are of high priority in astrophysics, traditional funding opportunities are unable to provide rapid turnaround. Furthermore, funding agencies are reticent about supporting exploratory programs aimed at improved observational modeling and calibration. This initiative comes at an ideal time, both for fully exploiting the final GALEX data set and for leveraging and building on the development of Stan.

2. Stan considers the fate of the Earth

White dwarfs (WD) are compact, Earth-sized, stellar remnants that represent the end state in the evolutionary history of stars with mass comparable to our Sun. Stars arrive in this state after burning most of the hydrogen (and helium) in their core and after passing through one or more “giant” phases where the star expands to a radius comparable to the Earth’s orbital radius, losing its outer envelope in a wind and potentially engulfing or disturbing any planetary system that has formed around the star.

How many planets are lost in the process? Only in the past year have observations reached the point where the stellar companion mass-period distribution, $\phi(M_c, P_c, M_\ast)$ around typical “main-sequence” stars is constrained by observations (e.g., Marcy et al. 2014). These measurements combined with the white dwarf companion distribution, $\phi_{WD}(M_{WD})$, would yield the first-ever “planetary survival function,”, one that could ultimately provide the odds of survival of an Earth during the Sun’s late-stage evolution (Nordhaus & Spiegel 2013).

As white dwarfs are planet-sized, transiting planetary companions can produce an easily detectable, deep eclipse. WDs are UV-bright and every GALEX observation typically
Fig. 1.— Top left: Theoretical constraints on the distribution function of WD companion mass and period (Nordhaus & Spiegel 2013). Existing data on short-period planets around normal stars shown as points. Top right: Observed distribution of WD targets in UV/optical and optical flux-ratio space, with eclipse detections shown. Bottom left: Simple eclipse light curve model with 5 parameters (WD mass, radius, companion radius, impact parameter and orbit diameter/period). Bottom right: MCMC model fit to detected UV photon light curve from GALEX observations of a white dwarf.

includes at least one well-detected WD in its one square degree field of view. With GALEX photon data we have already identified several larger eclipsing companions to WDs (see Figure 1). To extend this analysis towards a measurement of the WD companion function, $\phi_{WD}$, we must accurately model the window function, or the fraction of time we could have detected an eclipse caused by a range of planetary companions. When transits are detected, we can also model light curves to further constrain the posterior likelihood of companion parameters (Fig. 1). Photon-level instrumental effects, including time variable background and response, must also be included in this model. This problem is characteristically multilevel, where prior information guides the form of the top-level distribution function, the object-level light curve model, and the instrumental photon-level effects.

Because of the computational challenges, it is only in the past few years that hierarchical
Bayesian modeling has been successfully used to analyze astrophysical data sets (e.g., Mandel et al. 2011). Recent work uses Stan to model the light curves of supernovae (Sanders et al. 2014), with similarities to the WD study we propose here. In our case, the top-level inference (the fate of planets), and 100x larger data set, adds an extra degree of difficulty. More specifically, the supernovae model considers missing data, but only for individual detected light curves. Our analysis will deduce the form of the top-level WD companion distribution function by additionally incorporating information about WD eclipse non-detection and non-observation. Approaches that incorporate non-detections when modeling and inferring a population’s distribution function have broad utility, both in astronomy and beyond.

3. Stan on the birth of the Universe OR a new Galactic dust component

This past March, the astrophysics community was stunned by revelations from the deepest-ever image obtained of polarization from the Cosmic Microwave Background (CMB), based on observations from the BICEP2 telescope located in near the South Pole (BICEP2 Collaboration et al. 2014). Their analysis supported the tentative conclusion that the polarization signal in the CMB, observed as a power spectrum (using spherical harmonics), was caused by the imprint of gravitational waves that themselves were created when the Universe was only $10^{-32}$ seconds old, during a period of dramatic inflation. This remarkable result, if true, would have lasting consequences on our ability to test the inflationary Big Bang model and to probe the earliest moments in the history of the Universe.

The data quality from BICEP2 and subsequent measurements are excellent. However a lingering issue is whether the inference drawn from the data, that the detected signal is a signature of the Big Bang, is correct. While the BICEP2 team has worked hard to rule out other astrophysical sources and foregrounds, their conclusions have recently been questioned because several underlying assumptions have been shown to be optimistic. One leading alternative theory is that dust grains in our own Galaxy may become aligned with Galactic magnetic fields, and produce a polarized emission signature at the same wavelengths as the detected polarization signal. While several groups have sought to better constrain the expected polarization contribution from this Galactic dust component, their work relies on a detailed model of the distribution of Galactic dust and its polarized dust emission. Both of these are poorly constrained, particularly at the level required to interpret the CMB data.

As it turns out, most of the diffuse UV background signal detected by GALEX is also strongly correlated with the intensity of infrared emission from Galactic dust, as shown in Figure 2. Dust grains are excellent scatterers of UV photons, and observations have revealed that at high Galactic latitudes, the intensity of the UV background can be used to estimate the dust column density (Hamden et al. 2013, recently completed Ph.D. w/ Schiminovich). This correlation while strong, is not perfect, and there are regions of significant excess UV emission, generally localized in longitude, that are as yet unexplained (Figure 2, top left).
Fig. 2.— Top left: Ratio of diffuse UV background intensity to dust far infrared emission over the complete sky, in Galactic coordinates (Hamden et al. 2013). White regions are unobserved by GALEX (too bright). Regions of UV excess are green/yellow/red. Red contour traces outline of BICEP2 survey field. Top right: UV and dust intensity are correlated over most of the sky, with saturation at low latitudes. Middle right: Zoom of UV background from BICEP2 field showing systematic offsets on scales comparable to the GALEX field of view (tile size 1 sq. degree) and structure in BICEP2 map. Bottom left: BICEP2 Polarization data with structure on 0.5-1 degree scales. Bottom right: Current model of Galactic dust polarization fraction based on data from the Planck mission (see e.g., Flauger et al. 2014). Some regions of high polarization fraction coincide with high UV excess. BICEP2 assumed a 5% dust polarization fraction in their field, while the latest estimate suggests that it could be higher than 10%.

One possibility is that the UV excess is caused by a heretofore undetected component of Galactic dust. Intriguingly, the observations made by BICEP2 are in a 300 square degree field at the edge of one of the two UV excess regions identified by GALEX (red outline in Figure 2). Can this combined data set be used to infer a new component of Galactic dust? While such a result would be a blow to the inferences made earlier this year, the conclusions would be particularly important, both for the design of future CMB polarization experiments and for our own understanding of the formation and destruction of dust grains and the physical origin of magnetic fields in our Galaxy.
How does Stan fit in? While the GALEX observations are strongly suggestive of a large-scale excess dust component, there are several reasons why it is hard to make a direct comparison with the BICEP2 results. BICEP2 detects a polarization signature on $\sim 1$ degree scales, which coincides with the diameter of the GALEX field of view. The GALEX background map contains systematic instrumental calibration effects on these scales that must be properly modeled before making precise probabilistic inference. A multi-level Bayesian model is exactly what we require for this task because the calibration residuals affecting many independent photon measurements are highly correlated in a way that can be modeled using a modest, but non-trivial, number of parameters. The net result will be a new, improved calibration of the entire GALEX data set, itself a highly valuable result.

Our scientific analysis will also be significantly expedited because we will also be able to use Stan to perform probabilistic inference on different classes of Galactic dust models, thereby determining whether the amplitude of a new Galactic dust component is consistent with the level required to produce the polarization signature in BICEP2. Posterior likelihoods on model parameters will not only be used for the BICEP2 field, but can be calculated for the entire Galactic high-latitude sky, producing a data product of lasting value to current and planned CMB measurements being carried out from both hemispheres.

4. Proposed Work Effort and Data Management

The scientific aims of our two investigations are quite distinct, yet the commonality in their use of the Stan hierarchical modeling tool simplifies and focuses the effort. Since both projects rely on the same GALEX photon data set, access to these data will share the same API. The CMB/Galactic dust polarization investigation has a higher priority, but the WD investigation may be more straightforward to implement because we can start with already developed models. Our proposed timeline is:

- **Jul-Dec 2014**: Natural science graduate student (NSGR) develops initial hierarchical model for CMB and WD investigations (e.g. using graphical models, etc.) and assembles data sets to be used for both projects. Data science graduate student (DSGR) will code models using Stan, performing initial investigations into model stability and convergence. Scientific emphasis will be devoted to the CMB investigation, with goal of producing a revised degree-scale calibration of the GALEX diffuse UV background and probabilistic inference on the amplitude of a new component of Galactic dust. Model will be published and released to community in a Healpix all-sky data format.

- **Jan-Jun 2015**: NGSR will focus on incorporation of top-level distribution function into WD model and detailed modeling of photon-level calibration. On-going model testing will use posterior predictive checks. DSGR will code new models; as model complexity/dimensionality increases will explore convergence bottlenecks. It is anticipated that
the DSGR will consider new approaches for achieving faster convergence within Stan (e.g., see Betancourt 2013), and implement in code as needed.

- Jul-Oct 2015: NSGR will publish final scientific results on WD-companion discovery and distribution function. Papers will include application of methods to future large surveys. DSGR will publish methods and algorithms developed using Stan, addressing scalability and future challenges. Team will also release new calibration tables and/or photon data to community and report on published results at a conference.

While proposed work effort involves a large 10-40 TB data set, everything required to carry out this project is accessible to Schiminovich and students. In addition to publication, any new data products that result from this investigation will be openly distributed to the community, either via Columbia-based web portal or through the GALEX high level science products website maintained at the Space Telescope Science Institute in Baltimore, MD. This project also builds on a collaborative effort between Columbia and NYU natural and data scientists, also providing a connection that extends to other institutions supported by the Moore/Sloan data sciences initiative.

REFERENCES

Betancourt, M. J. 2013, Generalizing the No-U-Turn Sampler to Riemannian Manifolds, arXiv, 1304


7